Multi-Scale Gaussian Mixtures for Cross-species Study

ZiHua Yang, Abdullatif Alwatban, Richard Everson and Zheng Rong Yang

Abstract— **Cross-species studies using microarray gene expressions help discover genetic diversity among different species, the results of which are fundamental to comparative genomics. Various approaches have been used for cross-species studies, such as homogeneity test and cluster analysis. A homogeneity test provides a homogeneity significance ranking for each gene pair whilst cluster analysis looks at the discovery of co-expressed meta-genes. We propose a unified method to extract both homogeneous and heterogeneous expression patterns across species. The basic idea is to model the sum and difference of expressions across species using multi-scale Gaussians which reveal information about homogeneous and heterogeneous expression patterns respectively. We show using both simulated and real data that the proposed method is suited to identifying both homogeneous and heterogeneous gene expression patterns.**

I. Introduction

Cross-species study compares multiple data sets from biological/medical experiments for revealing how genes are conserved among distantly related species [1]. For instance, it is assumed that the primary structure of an orthologous gene will be conserved when species evolve even if their evolutionary distance is large [2]. Biological investigation in one species on conserved genes can therefore be tested in another species if two species are related. For instance, clinic trials for disease intervention are conducted on mice prior to being tested on human and there have been many human-mice species comparison studies [3-7]. However, gene expression levels vary significantly across species due to sample variation, technique resolution, posing a challenging issue in cross-species studies [1, 2, 8-10].

Two types of quantitative analytical approaches are typically used for cross-species analysis. The first approach tests homogeneity gene expressions across species. For example, the homogeneity test based on correlation was used to determine the genetic components of alcohol consumption between human and rats [11]. Those genes with large correlation coefficients across the two species were classified as homogeneous genes. In extension to correlation analysis which captures only second order and linear information, some studies looked at higher order statistics such as mutual information to detect early stress responses in rodent models of lung injury across species [12]. Evolutionary conservation is a complex implementation of correlation analysis for cross-species analysis [13-15]. Both correlation analysis and mutual information rely on large sample sizes. When examining homogeneity across more than two species, the Fisher combined probability test [16] is used when the species number is sufficiently large. It combines p values derived from significance analysis carried out separately in each species to deliver a combined p value of the dependence across species [17-19]. The Fisher combined probability test is unreliable when some p values are extremely small [20].

The second approach uses multivariate analysis such as unsupervised and supervised learning algorithms for cross-species studies. Supervised learning algorithms are used to examine whether a pattern reserved in one species can be a predictive factor for the other species. The algorithms used for supervised cross-species analysis include artificial neural network [21], linear discriminant analysis [22] and k-nearest neighbors [22]. For instance, artificial neural network algorithm was used for identifying conserved and divergent transcriptional modules across species [23], linear discriminant analysis was used for lung injury biomarker detection across multiple species [24], k-nearest neighbor algorithm was used for probe sequence identification [25]. Because phenotypic data are not always available or difficult to acquire, unsupervised learning algorithms are often used in real applications. Nonnegative matrix factorization [26, 27] is such an unsupervised learning algorithm and has been used to analyze common meta-genes across two species [28-30]. Using nonnegative matrix factorization (NMF), a gene expression matrix is decomposed in to a meta-gene expression matrix and a coefficient matrix of individual gene contributions to the meta-genes. Mathematically the expression level of a gene is a linear combination of the expression levels of all meta-genes. Based on the magnitude of coefficients, one can quantify the relationship of each gene to a meta-gene, giving a partition or cluster of the data. Common meta-genes can then be identified across species by inspecting separate NMF models. Cluster analysis as another set of important unsupervised learning algorithms for partitioning data. Each cluster corresponds to a meta-gene. Clustering algorithms including k-means algorithm, mixture models and self-organizing map [31] have been used for cross-species study [32-35]. When clustering gene expression data for cross-species analysis, data can be separately or jointly analyzed. However in cross-species studies it models gene expression matrices of two species separately and

ZiHua Ynag is with University of Queen Mary, UK (e-mail: z.h.yang@qmul.ac.uk). Abdullatif Alwatban is with University of Exeter (e-mail: asa220@ex. colostate.edu). Richard Everson is with University of Exeter (e-mail: reverson@ex.ac.uk). Zheng Rong Yang is with University of Exeter (e-mail: z.r.yang@ex.ac.uk).

requires a post analysis to compare two species. In addition, the determination of the number of meta-genes is based on *a priori* knowledge. In case such knowledge is unavailable, it is then difficult to accurately identify meta-genes.

Cluster analysis partitions data explicitly. During cluster analysis, each data point is assigned to a cluster and a cluster represents a meta-gene. When using cluster analysis for cross-species, one can model a gene expression matrix for each species separately [36-38] or model a combined gene expression matrix from multiple species [2, 32, 39]. Separate clustering suits any data size because each species is modelled individually to generate a cluster model through an unsupervised learning process. After individual cluster models have been generated, follow-up-manual work extracts co-expressed patterns across species. Combining expression matrices across species into one matrix, on the other hand, can avoid the difficulty of manual work [39]. Using combined expression matrix for cross-species analysis, gene expression matrices are normalized separately and then are merged into one matrix for using cluster analysis. Using this approach, we do not need a follow-up-manual comparison. Clusters of genes from multiple species directly show the co-expression of genes across species. However this approach has a stringent requirement of data pairing, i.e. all the expression matrices must have the same number of samples. Moreover, when the number of species becomes large, the number of samples is likely to be larger than the number of genes, i.e. the number of variables is larger than the number of instances in the context of machine learning. In this case, cluster analysis model can be unreliable.

A major distinction between the two quantitative approaches is the focus on co-expressed genes versus the focus on co-expressed meta-genes. A homogeneity test can detect exactly which subset of genes is co-expressed across species and measure the significance of co-expressions for gene ranking. Cluster analysis does not produce statistics of significance but summarises information based on meta-genes. One critical issue for cluster analysis is that cluster structures are often badly estimated when data noise is large. For instance, two overlapped clusters can easily be merged into one. When a cluster structure (or multiple cluster structures using separate clustering approach) is wrongly estimated, true homogeneous co-expressed genes can hardly be accurately predicted.

We introduce a new method using the multi-scale Gaussian (MSG) model for probabilistic cross-species analysis. We use MSG to model the sum and difference of DEs across two species. We use simulated and real data to illustrate the successful detection of both homogeneous and heterogeneous differential expression patterns (DEPs) using the proposed method.

II. Methods and Materials

A. *Multi-scale Gaussian*

The multi-scale Gaussian model is a mixture model [40-42], composed of two components which have similar or identical centres but different variances. We denote a vector by $Z=\{z_n\}_{n=1}^{N}\in\Re$. We assume that Z follows a mixture of two Gaussians, one with a small variance which is referred to as the null density while the other with a large variance referred to as the alternative density. A standard multi-scale Gaussian density function $f(z\mid\theta)$ given model parameters $\theta=(\mu_0,\mu_1,\sigma_0^2,\sigma_1^2,w_0,w_1)$ has the form:

$$f(z\mid\theta)=w_0\mathcal{G}(z\mid\mu_0,\sigma_0^2)+w_1\mathcal{G}(z\mid\mu_1,\sigma_1^2)$$

where

$$\mathcal{G}(z\mid\mu_m,\sigma_m^2)=\frac{1}{\sqrt{2\pi\sigma_m^2}}\exp\left(-\tfrac{1}{2}\sigma_m^{-2}(z-\mu_m)^2\right)$$

which is centred at μ_m with standard deviation σ_m (m = 0, 1); w_m is the mixing coefficient or weight satisfying $w_0+w_1=1$; σ_0 is the standard deviation of the null density, $\sigma_1\gg\sigma_0$ is the standard deviation of the alternative density. This model is also called a two-component hierarchical normal mixture model. In this model, $\mathcal{G}(z\mid\mu_0,\sigma_0^2)$ models the uninformative part of the data (null data -little variation from zero) and $\mathcal{G}(z\mid\mu_1,\sigma_1^2)$ models the informative part (alternative data - large deviation from zero). A small σ_0 naturally leads to a sharp peak centred at μ_0. However the density modelled by $\mathcal{G}(z\mid\mu_1,\sigma_1^2)$ can extend infinitely to two tails when $\sigma_1\gg\sigma_0$, $\mathcal{G}(z\mid\mu_1,\sigma_1^2)$ is approximates a uniform distribution and the identification of null data is straightforward. The likelihood function is defined as

$$P(Z\mid\theta)=\prod_{n=1}^{N}f(z_n\mid\theta)$$

We place (conjugate) inverse gamma distributed priors on the variances:

$$IG(\sigma_m^2\mid a_m,b_m)=\frac{b_m^{a_m}\sigma_m^{-2(a_m+1)}}{\Gamma(a_m)}\exp(-b_m\sigma_m^{-2})$$

The mixing coefficients are modelled using non-informative priors and we assign each of μ_0 and μ_1 a Gaussian prior with zero mean and small deviations τ_m: $\mu_m\sim\mathcal{G}(0,\tau_m^2)$. Moreover $|z|\to 0$ indicates null data. Such a mixture can learn $\mathcal{G}(z\mid\mu_0,\sigma_0^2)$ for data with a property of $|z|\to 0$. Therefore $|\mu_0|\to 0$ should be true if sufficient data have been acquired. On the other hand, estimating μ_1 largely depends on data. Because of large σ_1, slight difference between positive alternative data and negative alternative data will affect $|\mu_1|$ accordingly. We set the prior means to be zero based on our observation that both centers are close to zero and set both τ_m to be 0.5. The posterior is:

$$P(\theta \mid Z,\ \alpha) \propto P(Z \mid \theta)P(\theta \mid\ \alpha)$$
$$= \prod_{n=1}^{N}\{w_0\mathcal{G}(z_n \mid \mu_0, \beta_0^{-1}) + w_1\mathcal{G}(z_n \mid \mu_1, \beta_1^{-1})\}$$
$$\prod_{m=0}^{1}\frac{b_m^{a_m}\beta_m^{a_m+1}}{\Gamma(a_m)}\exp(-b_m\beta_m)\prod_{m=0}^{1}\mathcal{G}(0, \upsilon_m)$$

where $\alpha = (a_0, a_1, b_0, b_1, \tau_0, \tau_1)$ is the hyper-parameter set. The log-posterior can be written as:

$$\log P(\theta \mid Z,\ \alpha) \propto \sum_{n=1}^{N}\log f(\mathbf{z}_n \mid \theta) + \sum_{m=0}^{1}\log B_m$$

Here

$$B_m = (a_m + 1)\log\beta_m - b_m\beta_m - 0.5\upsilon_m\mu_m^2$$

and $\upsilon_m = \tau_m^{-2}$. The model parameters are estimated by maximising the posterior, giving the iterative variance updates

$$\sigma_m^2 = \frac{\sum\vartheta_{n,m}z_n^2 + 2\,b_m}{\sum\vartheta_{n,m} + 2a_m + 2}$$

where $\vartheta_{n,m}$ is defined as:

$$\vartheta_{n,m} = \frac{w_m\mathcal{G}(z_n \mid \mu_m, \sigma_m^2)}{f(z_n)}$$

The iterative update rule for the mixing coefficients can be written as: $w_m = (1/N)\sum_{n=1}^{N}\vartheta_{n,m}$ and the iterative update rule for each center is:

$$\mu_m = \frac{\beta_m\sum\vartheta_{n,m}z_n + \upsilon_m\mu_{m0}}{\beta_m\sum\vartheta_{n,m} + \upsilon_m}$$

where $\mu_{m0} = 0$. The learning process is initiated by assigning 0.5 to w_m , 1 to σ_m , and 0 to μ_m We set $\mu_{m0} = 0$ to reflect our observation that the null data are distributed around zero. Setting $w_m = 0.5$ assumes that there are roughly an equal number of alternative and null data initially when there is no *a priori* knowledge. We also set $\sigma_m = 1$ and $\mu_m = 0$ initially. For both Gaussians, the variance hyper-parameters a_0 and a_1 were set to one. b_0 was set to be the standard deviation of 90% of the smallest absolute data as we assume that the majority of data correspond to those of null genes. b_1 was the 90th percentile of all data. Bayesian learning has desirable convergence properties [43, 44]. The convergence of multi-scale Gaussians is usually determined by the maximum learning cycle and parameter stability. After convergence is achieved, the null density and the alternative density are estimated for each gene. The Bayes rule is then used to determine whether a data is alternative.

B. Using multi-scale Gaussian mixture

To use multi-scale Gaussian to model differential expressions (DEs) for a microarray expression data, we denote a DE matrix by $X = \{\mathbf{x}_n\}_{n=1}^{N} \in \Re^d$. It has N rows of genes or probe sets. We use the biological significance [45, 46] of the expression data and denote it by $Z = \mathrm{E}(X) \in \Re$.

C. Cross-species differential expression pattern discovery

We aimed to discover two kinds of cross-species differential expression patterns (DEPs). The first refers to the subset of genes with similar DE direction as well as magnitude across species - homogeneous DEP. The second refers to a subset of genes with opposite DE directions across species - heterogeneous DEP. Homogeneous DEPs inform how two species demonstrate similar response to stress whilst heterogeneous DEPs infer species diversity.

We denoted two DE matrices by $X = \{\mathbf{x}_n\}_{n=1}^{N} \in \Re^{d_x}$ and $Y = \{\mathbf{y}_n\}_{n=1}^{N} \in \Re^{d_y}$. Both matrices have N rows of genes or probe sets. From X and Y, we denoted a co-differential expression vector by $Z = \mathrm{E}(X) \otimes \mathrm{E}(Y)$, where $\otimes = (+, -)$ has two operations for the two kinds of DEPs. Here the sum function + was used for revealing homogeneous DEP (DEP0) across species and the difference function − was used to explore heterogeneous DEP (DEP1) across species.

D. Experimental design

You Simulated data were designed to evaluate the method in identifying DEP0s and DEP1s. We simulated 1000 genes across two species, each with ten control and ten test samples. The control samples and the test samples of non-DEGs were random samples of a normal distribution of mean ten and a varying standard deviation. The standard deviation values were 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1. The test samples of DEGs were random samples of a normal distribution with mean ten, plus (minus) one unit for up (down) regulated genes. Table 1 shows how these 1,000 genes were distributed in two data sets.

Table 1. Experimental design of simulated data with combinations of non-DEGs (Null), up-regulated DEG (UP) and down-regulated DEG (Down).

	Data set 1	Data set 2
800	Null	Null
25	Null	Up
25	Null	Down
25	Up	Null
25	Down	Null
25	Up	Up
25	Down	Down
25	Up	Down
25	Down	Up

We used two measurements to evaluate algorithms for the simulated data, i.e. sensitivity and false discovery proportion. The sensitivity was derived as follows for different algorithms. The Fisher test and correlation analysis return vectors of p values for the genes. Suppose the design vector is denoted as t. The sensitivity was calculated using $\#(t = 1 \,\&\, p < p_0)/\#(t = 1)$, where p_0 is the critical p value and $\#(x)$ means of the number of x. The false discovery proportion was calculated using $\#(t = 0 \,\&\, p < p_0)/\#(p < p_0)$. The largest sensitivity approaches one meaning that the p values for all the designed dependent/independent (homogeneous/heterogeneous) genes approach zero. The smallest sensitivity will approach zero. The smallest false discovery proportion is zero if none of the designed non-dependent/independent genes has its p value less than a critical p value. MSG returns a vector of the posterior probabilities (P) for the designed dependent/independent genes. Therefore the sensitivity was calculated using $\#(t=1 \& P>P_0)/\#(t=1)$, where P_0 is the critical posterior probability. The false discovery proportion was calculated using $\#(t=0 \& P>P_0)/\#(P>P_0)$.

A clustering approach will return two vectors of cluster memberships for genes. We therefore used different measurement techniques for calculating the sensitivity and false discovery proportion. Moreover, the separate clustering approach returns two separate models, we therefore used correlation to pair clusters from two models. We scanned the clusters of one model for one data set one by one and found one cluster from the other model with maximised correlation which is larger than a pre-defined threshold. We denoted the membership vectors as f_A and f_B. For one cluster (say k), $\vartheta_A^k=\{f_A=k\}$ and $\vartheta_B^k=\{f_B=k\}$ denote genes from two data sets A and B. For the sum function, a commonality proportion for the cluster was calculated as $\pi_k=\#\vartheta_A^k \bigcap \vartheta_B^k/\#(\vartheta_A^k \bigcup \vartheta_B^k)$. For the difference function, a uniqueness proportion for the cluster was calculated as $\pi_k=\#(\vartheta_A^k \bigcup \vartheta_B^k)\backslash(\vartheta_A^k \bigcap \vartheta_B^k)/\#(\vartheta_A^k \bigcup \vartheta_B^k)$. π_k was then assigned to all genes belonging to the cluster. We denoted each gene's sensitivity as s_i. The overall sensitivity was calculated as

$$\frac{1}{\#(t_i=1)}\sum_{t_i=1} s_i$$

and the false discovery proportion was calculated as

$$\frac{1}{\#(s_i>0)}\sum_{t_i=0} s_i$$

We set the critical p value to be 0.01 and the critical posterior probability to be 0.95. We downloaded a data set from the Gene Expression Omnibus (GEO, accession number GSE44337). The data set originated from a study on conserved gene differentiation in aggressive B lymphomas across human species (human diffuse large B cell) and mouse species (B6 iMyc).The human species expression data was generated using the GPL570 platform with 54,675 probe sets. It is composed of three samples of the wide type and nine tumour samples. The mouse species was generated using the GPL1261 platform with 45,101 probe sets. It is composed of three wide type samples and seven tumour samples. As both data sets have unmatched sample number between wide type and tumour, we used a one-to-one sample pairing approach, i.e., each wide type sample is paired with a tumour sample to generate a DE. Because different data sets in a cross-species study may use different platforms, we match probe sets to gene symbols. For gene symbols with more than one probe set, we selected the probe set with the maximum variance.

III. Results

Simulation with DEP0. *Fig. 1 shows the sensitivity evaluation of three approaches, separate clustering, joint clustering and MSG for identifying ten simulated DEP0 with varying noise levels. The Fisher test and correlation failed to provide meaningful results. It can be seen that the joint clustering approach and MSG were similar in performance while the separate clustering approach was very unreliable.* Fig. 2 illustrates the false discovery proportions for these simulations. We can see that MSG achieved the lowest false discovery proportion among three approaches.

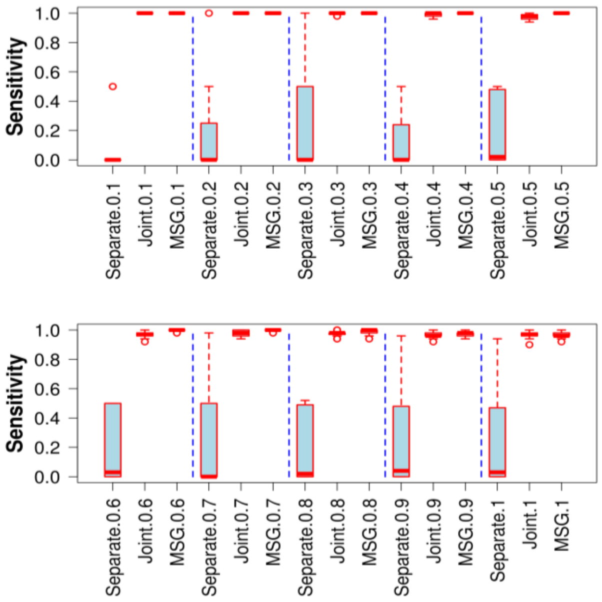

Fig. 1. The sensitivity measures for detecting DEP0.

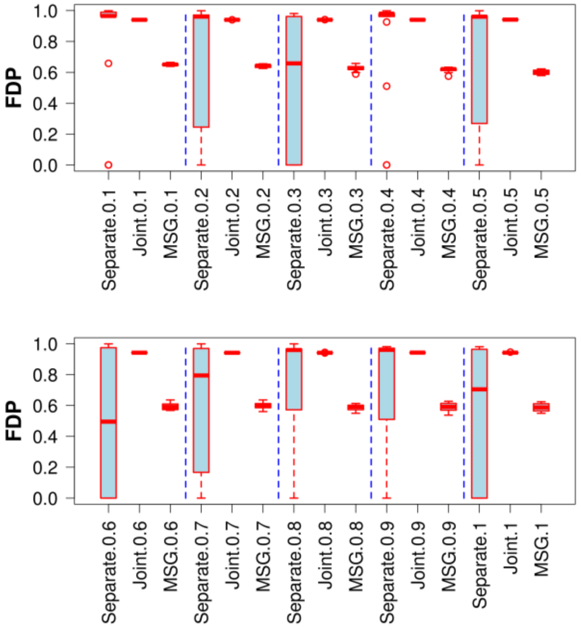

Fig. 2. The false discovery proportion measures for detecting DEP0.

Simulation with DEP0. Fig. 3 shows the sensitivity measures for detecting DEP1. It can be seen that only MSG worked well. Both the separate clustering approach and the joint clustering approach failed to predict independent DEGs. Fig. 4 shows the false discovery proportion. MSG's false discovery proportion is at 60% similar to that shown in detecting DEP0. However the false discovery proportion measures for the separate clustering approach and the joint clustering approach were nearly one meaning all the predicted independent DEGs were false ones.

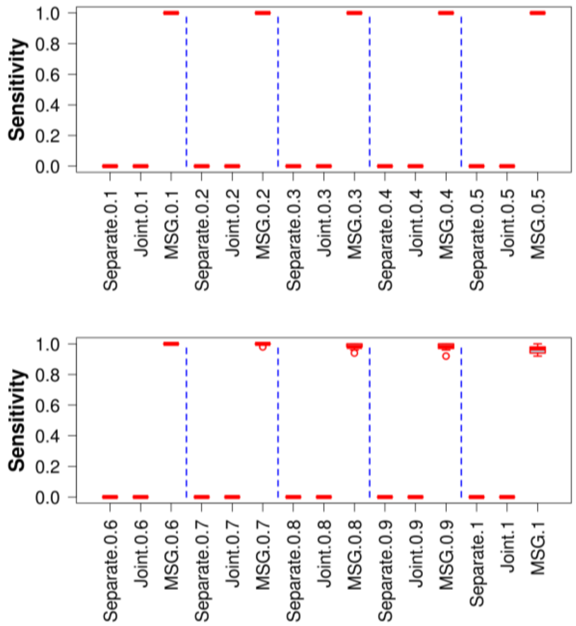

Fig. 3. The sensitivity measures for detecting DEP1.

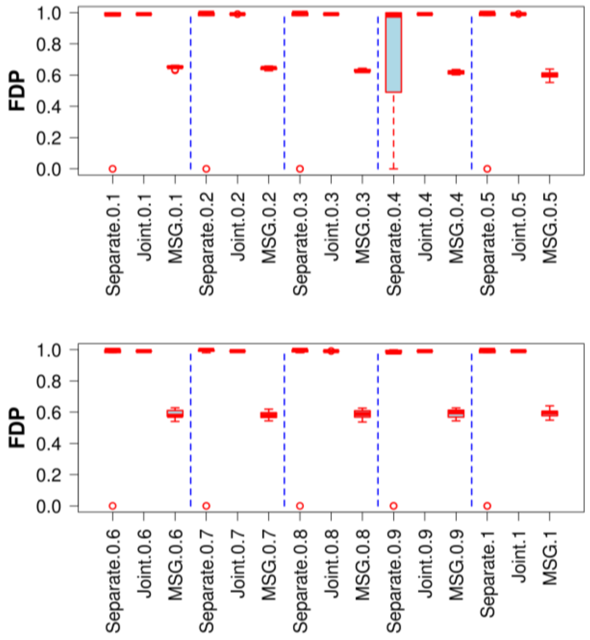

Fig. 4. The false discovery proportion measures for detecting DEP1.

Real data. We extracted 27 DEs for each gene in the human data set and 21 DEs for the mouse data set. After probe set - gene symbol mapping, we have selected 13,992 probe sets for the cross-species study. Using a 0.95 as the critical posterior probability, we identified 546 homogeneous DEGs and 309 heterogeneous DEGs. Fig. 5 shows the top ten homogeneous DEGs, where six genes show homogeneous up-regulation in both species and the rest show homogeneous down-regulation in both species. We then mapped these top ten genes to Gene Ontology Biological Processes and have found that 74 are identical among 87 biological processes for the human species and 74 mice biological processes for the mice species. This shows that the biological processes are well conserved for these homogeneous DEGs.

Fig. 6 shows the top ten heterogeneous DEGs. When mapping these top ten genes to Gene ontology biological processes, we have found that 68 are identical among 73 biological processes for the human species and 77 mice biological processes for the mice species.

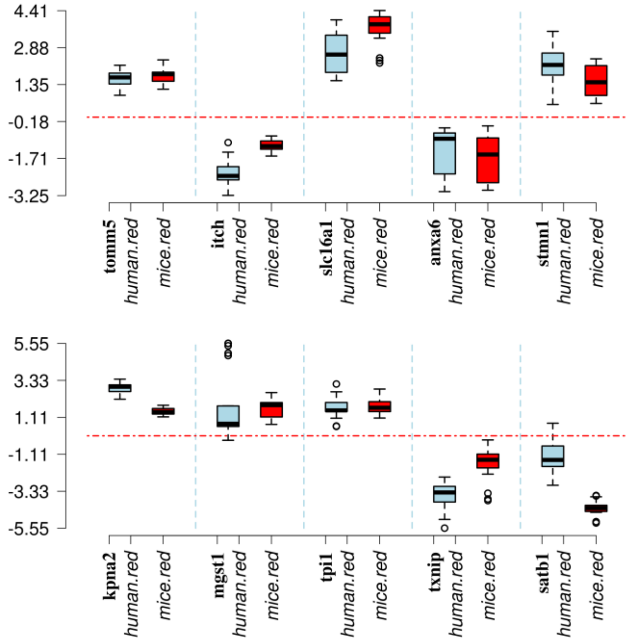

Fig. 5. Expressions of the top ten homogeneous genes across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The horizontal lines in red represent the cutoffs for gene significance.

Table 2 summarizes the top ten homogeneous and heterogeneous DEGs across the human and mice species. Among them, several have been both tested in human and mice. For instance, TOMM5 has been tested in human and mice tissue for examining the histopathology value and used mice as the baseline assay for high-throughput phenotyping [47]. CALU has been examined in human tissue for using IRF5 as a tumour suppressor in splenic marginal-zone lymphoma [48]. BCAR3 has been studied in relation with B lymphoma in both mice and human tissue [49]. ANXA6 was studied in human tissue for lymphoma [50]. STMN1 has been studied in human tissue for its anti-cancer activity [51].

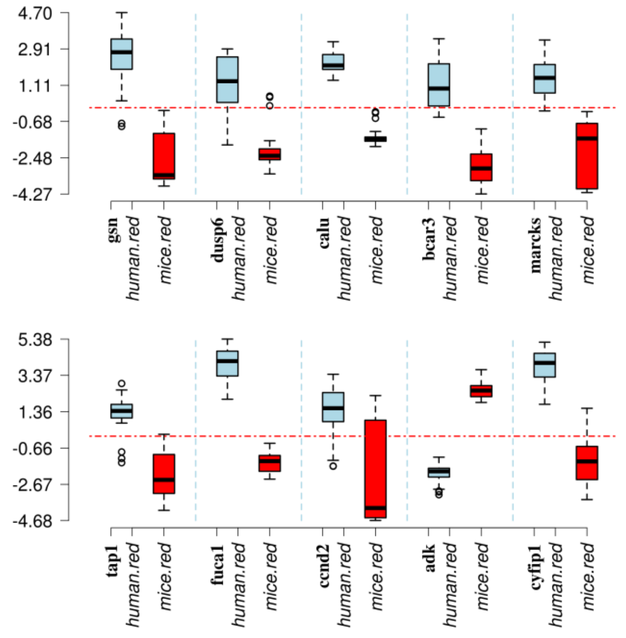

Fig. 6. Expressions of the top ten heterogeneous genes across mice (blue) and human (red). The boxes in light blue (red) represent the 27 DEs for each gene in the human species. The boxes in red represent 21 DEs for each gene in the mouse species. The horizontal lines in red represent the cutoffs for gene significance.

Table 2. Top ten homogeneous and heterogeneous differentially co-expressed genes. FC means Fold Change.

Homogeneous			Heterogeneous		
Symbol	FC_{human}	FC_{mice}	Symbol	FC_{human}	FC_{mice}
TOMM5	1.6112	1.721	SN	2.4890	-2.443
ITCH	-2.352	-1.169	DUSP6	1.2692	-2.009
SLC16A1	2.6025	3.6919	CALU	2.2258	-1.414
ANXA6	-1.401	-1.718	BCAR3	1.1410	-2.929
STMN1	2.1274	1.5015	MARCKS	1.4876	-1.941
KPNA2	2.8837	1.4777	TAP1	1.2625	-2.224
MGST1	1.6952	1.6459	FUCA1	3.9882	-1.406
TPI1	1.7194	1.7664	CCND2	1.4121	-2.458
TXNIP	-3.590	-1.649	ADK	-2.095	2.5622
SATB1	-1.252	-4.376	CYFIP1	3.8051	-1.391

IV. Conclusion

We have presented a new method for exploring homogeneous and heterogeneous DEPs across species using multi-scale Gaussian mixtures. The algorithm is motivated by the limitations in using existing homogeneity tests and cluster analysis techniques for cross-species analysis. The former suffers from low sensitivity and the latter is often severely affected by data noise. Density function is able to model two types of data with little difficulty. They are non-differentiable (hence informative modelled by a Gaussian density with a small variance) and differentiable (hence uninformative modelled by a Gaussian with a large variance) between two species. We introduced the sum/difference functions of DEs for homogeneous and heterogeneous DEP discovery. With these functions, co-expressed genes across species can be detected using the sum function while heterogeneous DEGs can be detected using the difference function. The method is simple to implement but can be extended to deal with more general problem settings in future work. We use simulated data and real data to illustrate the efficacy of this method in revealing both homogeneous and heterogeneous DEPs.

References

[1] E. Kristiansson, Osterlund, T, Gunnarsson, L, Arne, G, Joakim, Larsson, DG, Nerman, O, "A novel method for cross-species gene expression analysis," *BMC Bioinformatics,* vol. 14, p. 70, 2013.

[2] X. Zheng-Bradley, Rung, J, Parkinson, H, Brazma, A, "Large scale comparison of global gene expression patterns in human and mouse," *Genome Biol,* vol. 11, p. r124, 2010.

[3] U. Ala, Piro, RM, Grassi, E, Damasco, C, Silengo, L, Oti, M, Provero, P, Di Cunto, F, "Prediction of human disease genes by human-mouse conserved coexpression analysis," *PLoS Comput Biol,* vol. 4, p. e1000043, 2009.

[4] A. Sweet-Cordero, Mukherjee, S, You, ASH, Roix, JJ, Ladd-Acosta, C, Mesirov, J, Golub, TR, Jacks, T, "An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis," *Nat Genet,* vol. 37, pp. 48-55, 2005.

[5] J. Miller, Horvath, S, Geschwind, DH, "Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways," *PNAS,* vol. 107, pp. 220-229, 2010.

[6] A. Rasche, Al-Hasani, H, Herwig, R, "Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 Diabetes mellitus," *BMC Genomics,* vol. 9, p. 310, 2008.

[7] E. Segal, Friedman. N, Kaminski, N, Regev, A, Koller, D, "From signatures to models: understanding cancer using microarrays," *Nat Genet,* vol. 37, pp. S38-45, 2005.

[8] D. Allison, Cui, X, Page, GP, Sabripour, M, "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet,* vol. 14, pp. 55-6, 2006.

[9] M. Consortium, "The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nat Biotechnol,* vol. 14, pp. 1151-61, 2006.

[10] W. Kuo, Liu, F, Trimarchi, J, Punzo, C, Lombardi, M, Sarang, J, Whipple, ME, Maysuria, M, Serikawa, K, Lee, SY, McCrann, D, Kang, J, Shearstone, JR, Burke, J, Park, DJ, Wang, X, Rector, TL, Ricciardi-Castagnoli, P, Perrin, S, Choi, S, Bumgarner, R, Kim, JH, III, GFS, Freeman, MW, Seed, B, Jensen, R, Church, GM, Hovig, E, Cepko, CL, Park, P, Ohno-Machado, L, Jenssen, TK, "A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies," *Nat Biotechnol,* vol. 14, pp. 832-40, 2006.

[11] B. Tabakoff, Saba, L, Printz, M, Flodman, P, Hodgkinson, C, Goldman, D, Koob, G, Richardson, HN, Kechris, K, Bell, RL, Hübner, N, Heinig, M, Pravenec, M, Mangion, J, Legault, L, Dongier, M, Conigrave, KM, Whitfield, JB, Saunders, J, Grant, B, Hoffman, PL, "Genetical genomic determinants of alcohol consumption in rats and humans," *BMC Biol,* vol. 7, p. 70, 2009.

[12] S. Ma, Grigoryev, DN, Taylor, AD, Nonas, S, Sammani, S, Ye, SQ, Garcia, JG, "Bioinformatic identification of novel early stress response genes in rodent models of lung injury," *Am J Physiol Lung Cell Mol Physiol,* vol. 289, pp. L468-77, 2005.

[13] B. Liao, Zhang, JZ, "Evolutionary conservation of expression profiles between human and mouse orthologous genes," *Mol Biol Evol,* vol. 23, pp. 530-40, 2006.

[14] B. Dutilh, Huynen, MA, Snel, B, "A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation," *BMC Genomics,* vol. 7, p. 10, 2006.

[15] K. Essien, Hannenhalli, S, Stoeckert, CJ, "Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to Plasmodium phenotypic diversity," *PLoS One,* vol. 3, p. e3122, 2008.

[16] R. Fisher, "Answer to question 14 on combining independent tests of significance," *Amer Statistician,* vol. 2, p. 30, 1948.

[17] P. Hu, Greenwood, CMT, Beyene, J, "Statistical methods for meta-analysis of microarray data: a comparative study," *Inf Syst Front,* vol. 8, pp. 9-20, 2006.

[18] A. Campain, Yang, YH, "Comparison study of microarray

meta-analysis methods," *BMC Bioinformatics,* vol. 3, p. 408, 2010.

[19] G. Tseng, Ghosh, D, Feingold, E, "Comprehensive literature review and statistical considerations for microarray meta-analysis," *Nucleic Acids Res,* vol. 40, pp. 3785-99, 2012.

[20] W. Rice, "A consensus combined P-value test and the family-wide signi_cance of component tests," *Biometrics,* vol. 46 pp. 303-8, 1990.

[21] D. E. Rumelhart, McClelland, J.L, *Parallel Dsitributed Processing*. Cambridge, MA, USA: MIT press, 1986.

[22] R. O. Duda, Hart, P.E., Stork, D.G., *Pattern Classification*, 2nd ed.: Wiley-Interscience, 2000.

[23] H. Li, Zhan, M, "Identifying Conserved and Divergent Transcriptional Modules by Cross-species Matrix Decomposition on Microarray Data," *J Proteomics Bioinform,* vol. 2, p. 117, 2009.

[24] P. Hu, Wang, X, Haitsma, JJ, Furmli, S, Masoom, H, Liu, M, Imai, Y, Slutsky, AS, Beyene, J, Greenwood, CM, dos Santos, C, "Microarray meta-analysis identifies acute lung injury biomarkers in donor lungs that predict development of primary graft failure in recipients," *PLoS One,* vol. 7, p. e45506, 2012.

[25] T. Royce, Rozowsky, JS, Gerstein, MB, "Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification," *Nucleic Acids Res,* vol. 35, p. e99, 2007.

[26] D. D. Lee, Seung, H.S., "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, pp. 788-91, 1999.

[27] D. D. Lee, Seung, H.S., "Algorithms for non-negative matrix factorization," *Adv Neural Info Proc Syst,* vol. 13, pp. 556–62, 2001.

[28] P. Tamayo, Scanfeld, D., Ebert, B.L., Gillette, M.A., Roberts, C.W., Mesirov, J.P., "Metagene projection for cross-platform, cross-species characterization of global transcriptional states," *PNAS,* vol. 104, pp. 5959-64, 2007.

[29] J. M. e. a. Stuart, "A gene-coexpression network for global discovery of conserved genetic modules," *Science,* vol. 302, pp. 249-255, 2003.

[30] J. P. Brunet, Tamayo, P., Golub, T.R., Mesirov, J.P., "Metagenes and molecular pattern discovery using matrix factorization," *PNAS,* vol. 101, pp. 4164-9, 2004.

[31] T. Kohonen, *Self-Organizing Maps*. Berlin: Springer, 2001.

[32] J. Cai, Xie, D, Fan, Z, Chipperfield, H, Marden, J, Wong, WH, Zhong, S, "Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells," *PLoS Comput Biol,* vol. 6, p. e1000707, 2010.

[33] Z. Mi, Shen, K., Song, N., Cheng, C., Song, C., Kaminski, N., Tseng, G.C., "Module-based prediction approach for robust inter-study predictions in microarray data," *Bioinformatics,* vol. 26, pp. 2586-93, 2010.

[34] Y. Lu, He, X, Zhong, S, "Cross-species microarray analysis with the OSCAR system suggests an INSR->Pax6->NQO1 neuro-protective pathway in aging and Alzheimer's disease," *Nucleic Acids Res,* vol. 35, pp. W105-14, 2007.

[35] Y. Lu, Huggins, P., Bar-Joseph, Z., "Cross species analysis of microarray expression data," *Bioinformatics,* vol. 25, pp. 1476-83, 2009.

[36] M. Schlicht, Matysiak, B, Brodzeller, T, Wen, X, Liu, H, Zhou, G, Dhir, R, Hessner, MJ, Tonellato, P, Suckow, M, Pollard, M, Datta, MW, "Cross-species global and subset gene expression profiling identifies genes involved in prostate cancer response to selenium," *BMC Genomics,* vol. 5, p. 58, 2004.

[37] Y. Lu, He, X., Zhong, S., "Cross-species microarray analysis with the OSCAR system suggests an INSR->Pax6->NQO1 neuro-protective pathway in aging and Alzheimer's disease," *NAR,* vol. 35, pp. W105–14, 2007.

[38] Y. Lu, Yi, Y, Liu, P, Wen, W, James, M, Wang, D, You, M, "Common human cancer genes discovered by integrated gene-expression analysis," *PLoS One,* vol. 2, p. e1149, 2007.

[39] W. Huang, Cao, X, Zhong, S, "Network-based comparison of temporal gene expression patterns," *Bioinformatics,* vol. 26, pp. 2944-51, 2010.

[40] P. Kramer, Kurbanmuradov, O, Sabelfeld, K, "Comparative analysis of multiscale Gaussian random field simulation algorithms," *Journal of Computational Physics,* vol. 226, pp. 897-24, 2007.

[41] E. Bertschinger, "Multiscale Gaussian random fields and their application to cosmological simulations," *The Astrophysical Journal,* vol. 137, pp. 1-20, 2001.

[42] S. Billings, Wei, H, Balikhin, MA, "Generalised multiscale radial basis function networks," *Neural Networks,* vol. 20, pp. 1081-94, 2007.

[43] C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer, 2006.

[44] A. Webb, *Statistical Pattern Recognition*. Chichester: John Wiley & Sons Ltd, 2002.

[45] Y. Xiao, Hsiao, T.H., Suresh, U., Chen, H.I., Wu, X., Wolf, S.E., Chen, Y., "A Novel Significance Score for Gene Selection and Ranking," *Bioinformatics,* vol. in press, 2012.

[46] D. McCarthy, Smyth, GK, "Testing significance relative to a fold-change threshold is a TREAT," *Bioinformatics,* vol. 25, pp. 765–71, 2009.

[47] P. Vogel, Read, RW, Rehg, JE, Hansen, GM, "Cryptogenic organizing pneumonia in Tomm5(-/-) mice," *Vet Pathol,* vol. 50, pp. 65-75, 2013.

[48] V. Fresquet, Robles, EF, Parker, A, et al, "High-throughput sequencing analysis of the chromosome 7q32 deletion reveals IRF5 as a potential tumour suppressor in splenic marginal-zone lymphoma," *Br J Haematol,* vol. 158, pp. 712-26, 2012.

[49] Y. Nakayama, Iwamoto, Y, Maher, SE, Tanaka, Y, Bothwell, AL, "Altered gene expression upon BCR cross-linking in Burkitt's lymphoma B cell line," *Biochem Biophys Res Commun,* vol. 277, pp. 124-7, 2000.

[50] A. Turtoi, Sharan, RN, Srivastava, A, Schneeweiss, FH, "Proteomic and genomic modulations induced by &#x03B3;-irradiation of human blood lymphocytes," *Int J Radiat Biol,* vol. 86, pp. 888-904, 2010.

[51] S. Rana, Maples, PB, Senzer, N, Nemunaitis, J, "Stathmin 1: a novel therapeutic target for anticancer activity," *Expert Rev Anticancer Ther,* vol. 8, pp. 1461-70, 2008.