Evolutionary Transcriptome Analysis Based on Differentially Expressed (DE) Genes

Xun Gu^{1*}

¹ Department of Genetics, Development and Cell Biology, Iowa State University, Ames,
IA 50011, USA.

*Corresponding author. E-mail: xgu@iastate.edu

Abstract

To address how gene regulation plays a key role in phenotypic innovations through high throughput transcriptomes, it is desirable to develop statistically-sound methods that enable researchers to study the pattern of transcriptome evolution. Most methods currently available are based on the Ornstein-Uhlenbeck (OU) model that considers the stabilizing selection as the baseline model of transcriptome evolution. In this paper, we developed a new evolutionary approach, based on the genome-wide p -value profile arising from statistical testing of differentially expressed (DE) genes between species. Our current approach is focused on the estimation of transcriptome distance between species. We first establish the relationship between the evolutionary model (the Markov-chain or Poisson model) and the proportion of null hypothesis (u_0), which can be used to estimate the transcriptome distance. Further, we calculate the posterior probability of a gene being DE when a p -value is given, denoted by $Q=P(DE/p)$, and develop a simple algorithm to estimate the transcriptome distance for any number of genes in the genome. Our compute simulations showed the statistical performance of these new methods are generally satisfactory.

Introduction

With enormous amounts of transcriptome (RNA-seq) data from multiple tissues with diverse species available, it is the exciting time for evolutionary biologists to address how gene regulation plays a key role in phenotypic innovations (King and Wilson 1975; Lehner 2013) (e.g., Wang, et al. 2009; Brawand, et al. 2011; McCarthy, et al. 2012; Necsulea and Kaessmann 2014; Xu, et al. 2018; Cardoso-Moreira, et al. 2019). To help achieve this goal, development of statistically-sound methods that enable researchers to study the pattern of transcriptome evolution is desirable (Gu and Su 2007; Pereira, et al. 2009; Gu, et al. 2013; Gu 2016; Ruan, et al. 2016; Gu, et al. 2017; Sudmant, et al. 2015; Yang et al. 2018; Gu et al. 2019). We (Yang et al. 2019) recently developed an R-package, *TreeExp2* (*Tree*-dependent *Expression* analysis for short), which provides a suite of up-to-date statistical methods for phylogenetic analysis of transcriptomes. It should be noticed that most methods implemented in *TreeExp2* are based on the Ornstein-Uhlenbeck (OU) model that considers the stabilizing selection as the baseline model of transcriptome evolution (Hansen and Martins 1996; Lemos, et al. 2005), that is, stabilizing selection, which maintains the optima under the background of random mutations, dominates the transcriptome evolution (Bedford and Hartl 2009).

In a high-throughput transcriptome study, statistical detection of differentially expressed (DE) genes between two sets of samples is probably the most common practice in the transcriptome data analysis (Cheng et al. 2004; Efron 2004; Efron et al. 2001). Each statistical hypothesis testing (usually for each gene) results in a p -value summarizing the level of statistical significance of the observed differential gene expression. Xu et al. (2018) conducted an analysis to predict differentially expressed (DE) genes between the human and chimpanzee in several brain regions. Their study indicated that detection of DE genes between species is a valuable approach to understanding the pattern of transcriptome evolution.

In the past decades a wealth of statistical literatures were reported to deal with the multiple DE-test problem, i.e., many null hypotheses being tested simultaneously in a high throughput experimentation. Our goal is to formulate an evolutionary model based on the statistical framework primarily developed for the detection of DE genes

(Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001). In this paper we focus on the estimation of transcriptome distance between species, a basic parameter for understanding the regulatory evolution. After invoking the widely-used BUM (beta-uniform mixture) model for the genome-wide p -value distribution (Pounds and Morris 2003; Smyth 2004; Storey and Tibshirani 2003; Zhang and Grant 2004; Zhang 2011), we first study the relationship between the evolutionary model (the Markov-chain or Poisson model) and the proportion of null hypothesis (u_0), which can be used to estimate the transcriptome distance. Further, we calculate the posterior probability of a gene being DE denoted by $Q=P(DE/p)$ based on the BUM model, and develop a simple algorithm to estimate the transcriptome distance for any number of genes with special interests. While compute simulation is used to evaluate the statistical performance of these new methods, case studies demonstrate their potential applications. The differences between these DE-based methods and OU-based methods are discussed.

Results

Statistically modeling of p -values for differentially expressed genes

While p -values arising from the null hypothesis are distributed uniformly on the interval $(0, 1)$, those arising from the alternative hypothesis follows a distribution that is generally modeled by a beta distribution whose pdf is given by

$$\phi(p) = \frac{p^{\beta-1}(1-p)^{\beta_0-1}}{B(\beta, \beta_0)} \quad (1)$$

It follows that the distribution of a set of p -values can be written as a beta-uniform mixture model (BUM) consisting of a uniform $(0, 1)$ component for the null hypothesis and a beta component for the alternative hypothesis, with the pdf given by

$$\Phi(p) = u_0 + (1 - u_0)\phi(p) \quad (2)$$

where u_0 is the proportion of null hypothesis. The BUM model can be intuitively interpreted as follows. Under the null hypothesis, the p -values will have a uniform density corresponding to a flat horizontal line. Under the alternative hypothesis, the p -values will have a distribution that has high density for small p -values and the density will decrease as the p -values increase. The overall BUM distribution will be a mixture of

p -values arising from the two hypotheses. Note that Eq.(1) involves two parameters, β and β_0 . With the restrictions of $0 < \beta < 1$ and $\beta_0 \geq 1$, Eq.(2) provides a reasonable model for the distribution of p -values arising from high throughput genomics (Storey and Tibshirani 2003): it asymptotes at $p=0$ and monotonically decreases to its minimum of u_0 at $p=1$.

u_0 -Evolutionary distance of transcriptome

When the differentially expressed (DE) gene detection is applied to two species samples of the same tissue, one may speculate that the number of DE genes is small between closely-related species, whereas it is large between distantly-related species. An intriguing question is how it may relate to the distance of transcriptome evolution. As shown below, we will address this issue under two models: the Markov chain model, and the Poisson model.

The Markov-chain model

Suppose any gene in a given tissue has two states: unexpressed or expressed. Let π_1 be the proportion of expressed genes, and $\pi_0=1-\pi_1$ be that of unexpressed genes. Let λ be the rate of transcriptome evolution. Under a two-state Markov-chain model, the gain rate and the loss rate of an expression status are given by $\pi_1\lambda$ and $\pi_0\lambda$, respectively. It follows that the transition probabilities from state i to state j after the t time units, $P_{ij}(t)$, $i, j=0$ (unexpressed), or 1 (expressed), are given by $P_{11}(t)=\pi_1+\pi_0e^{-\lambda t}$, and $P_{10}(t)=\pi_0(1-e^{-\lambda t})$, respectively (t is the speciation time), and similarly, we have $P_{00}(t)=\pi_0+\pi_1e^{-\lambda t}$, and $P_{01}(t)=\pi_1(1-e^{-\lambda t})$. For any orthologous gene pair, the probability for any observed expression pattern $P(r_A, r_B)$ can be calculated as follows

$$\begin{aligned} P(1, 1) &= \pi_1^2 + \pi_1\pi_0e^{-2\lambda t} \\ P(1, 0) &= \pi_1\pi_0(1 - e^{-2\lambda t}) \\ P(0, 1) &= \pi_1\pi_0(1 - e^{-2\lambda t}) \\ P(0, 0) &= \pi_0^2 + \pi_1\pi_0e^{-2\lambda t} \end{aligned} \quad (3)$$

For instance, $P(1, 1)$ is the probability that both genes are expressed; $P(1, 0)$ is the probability that gene-1 is expressed but gene-2 is unexpressed; and so forth.

The Markov-chain model depicted by Eq.(3) provides an evolutionary

interpretation for the BUM model of DE genes: the probability of a gene being no differentially expressed (null hypothesis) is given by $P(1, 1)+P(0, 0)$, as predicted by u_0 , the proportion of null hypotheses. The relationship $u_0=P(1, 1)+P(0, 0)$ implies that

$$u_0 = 1 - 2\pi_0\pi_1 (1 - e^{-2\lambda t}) \quad (4)$$

Apparently, a large proportion of non-DE genes is expected for closely-related species ($u_0 \rightarrow 1$ as $t \rightarrow 0$), while it decreases for distantly-related species, ultimately toward a saturation level of $(1-2\pi_1\pi_0)$ as $t \rightarrow \infty$. Next, we define the distance of transcriptome evolution by $G=(2\pi_1\pi_0)2\lambda t$, the expected number of expression changes per gene per time unit. According to Eq.(4), one can estimate G by the following formula

$$G = -2\pi_0\pi_1 \left(1 - \frac{1 - u_0}{2\pi_0\pi_1}\right) \quad (5)$$

provided that u_0 , the proportion of null hypotheses, is estimated.

The Poisson model

Under the Poisson model, the probability of a gene being no differentially expressed (non-DE) between two species is given by $e^{-2\lambda t}$. Hence, the u_0 -transcriptome distance defined by $G=2\lambda t$ can be simply estimated by the relationship $e^{-2\lambda t}=u_0$, i.e.,

$$G = -\ln(u_0) \quad (6)$$

Usually, the Poisson distance given by Eq.(6) tends to underestimate the transcriptome distance. Yet, it is useful in practice in the case when π_1 (or π_0) is difficult to estimate, or when Eq.(5) is inapplicable, which may occur between two distantly-related species. At any rate, both Eq.(5) and Eq.(6) converge to $1-u_0$ for closely-related species.

Estimation procedure of u_0 -transcriptome distance and case study

Non-parametric estimation of u_0

Suppose we have multiple hypothesis tests with simultaneously M p -values. With some minor modifications, we implemented the algorithm developed by Zhang and Grant (2004) to estimate u_0 , the proportion of null hypotheses. First, the M p -values are first sorted in ascending order so that $p_1 \leq p_2 \leq p_3 \leq \dots p_{M-1} \leq p_M$. An empirical cumulative

distribution of p -values is obtained by plotting i/M versus p_i . Second, we connect each point $(i/M, p_i)$ on the empirical cumulative plot to the end point $(1.0, 1.0)$ by a straight line, and calculates the slope of the line as

$$c_i = \frac{1 - i/M}{1 - p_i} \quad (7)$$

Then the median value of c_i for a given range of p_i , $p_l \leq p_i \leq p_u$, is taken as an estimate of u_0 ,

$$\hat{u}_0 = \text{medium}(c_i | p_l \leq p_i \leq p_u) \quad (8)$$

with $p_l=0.40$, $p_u=0.95$ as the default values of the method parameters.

Data availability

We downloaded published RNA-seq datasets of four primates, human, chimpanzee, gorilla and gibbon (Xu et al. 2108). Each species consists of eight brain areas, including cerebellum (CB), hippocampus (HIP), striatum (STR), anterior cingulate cortex (ACC), primary visual cortex (V1C), premotor cortex (PMC), ventrolateral prefrontal cortex (VPFC), DPFC (dorsolateral prefrontal cortex (DPFC)). All primate species possess 2-6 biological replicates except that gibbon only has one for all samples. Among the cerebral region, the neocortical area includes ACC, V1C, PMC, VPFC and DPFC, whereas the subcortical area includes HIP and STR. For each RNA-seq sample, the median ratio method was applied to normalize the raw reads count data to make these expression values comparable cross species and brain areas. Following the original authors, a total of 27,991 expressed genes are used in the current study.

The u_0 -transcriptome distances between human and chimpanzee brain regions

For each of eight brain regions, we carried out the differentially expressed (DE) gene analyses between the human and chimpanzee, using three software packages, *DESeq2* (Love et al. 2014), *edgeR* (Robison et al. 2010) and *limma* (Ritchie et al. 2015). The genome-wide p -value profile for each method in each brain region was obtained. We then estimated u_0 , the proportion of null hypothesis, in each case by Eq.(7) and Eq.(8). As shown in Table 1, the estimate of u_0 varies among methods as well as brain regions. The average u_0 (over brain regions) is 0.88, 0.91.and 0.95 for *DESeq2*, *edgeR* and *limma*,

respectively. The range of u_0 by *DESeq2*, for instance, is from 0.825 (CB) to 0.914 (PMC).

To estimate u_0 -transcriptome distance (G) under the Markov-chain model, we have to estimate π_1 and π_0 . From ENSEMBL, the total number of human genes is roughly 39K, one can estimate $\pi_1=27991/39000=0.717$ and $\pi_0=0.283$. For the comparison, we estimated the transcriptome distance under the stationary OU model (Yang et al. 2019). Interestingly, for each brain region, the estimate of SOU transcriptome distance is somewhere among three u_0 -transcriptome distances (Fig.1). In other words, differences among statistical methods for DE detections may accounts for most estimation differences. Yet the average u_0 -distance is actually impressively consistent with the SOU-distance. As shown below, the new evolutionary framework may provide some insights on the long-term research topic on how transcriptome evolved between the human and chimpanzee (Enard et al. 2002; Gu and Gu 2003; Gilad et al. 2006; Nowick et al. 2009; Brawand et al. 2011; Xun et al. 2018).

A Bayesian approach to estimating transcriptome distance

The posterior probability for a gene to be differentially expressed, given a p-value

Instead of the most common p -value assigned for each DE test, we anticipate that the posterior probability of a gene being differentially expressed, denoted by $P(DE|p)$, may be useful for the analysis of transcriptome evolution (Pounds and Morris 2003; Smyth 2004; Storey and Tibshirani 2003). According to the BUM model of Eq.(2), it is straightforward to calculate $Q=P(DE|p)$ by the Bayes rule, that is

$$P(DE|p) = (1 - u_0) \frac{\phi(p)}{\Phi(p)} = \frac{(1 - u_0)\phi(p)}{u_0 + (1 - u_0)\phi(p)} \quad (9)$$

In the following analysis we use a special case of BUM with $\beta_0=1$ (Storey and Tibshirani 2003), in which $P(DE|p)$ can be simplified as follows

$$P(DE|p) = \frac{a}{p^{1-\beta} + a} \quad (10)$$

where $a=(1-u_0)\beta/u_0$. We implemented the following empirical Bayesian procedure: (i) estimate u_0 by the nonparametric method of Eq.(8); and (ii) after treating the estimate of

u_0 as known, the maximum likelihood (ML) estimates of β in Eq.(2) can be obtained by fitting the beta-uniform mixture (BUM) distribution with $\beta_0=1$ to the observed p -values (Pounds and Morris 2003; Storey and Tibshirani 2003).

Q-evolutionary distance of transcriptome

Suppose we have M hypothesis tests (p -values) at the genome level. Let $Q_k=P(\text{DE}|p_k)$ be the posterior probability of the k -th gene being differentially expressed, $k=1, \dots, M$. Intuitively, one may use the average Q given by

$$\bar{Q}_M = \sum_{k=1}^M Q_k / M \quad (11)$$

for empirically estimate of $q=P(1, 0)+P(0, 1)$, the proportion of differentially expressed genes between species. Hence, under the Markov-chain model, the Q -transcriptome distance is given by

$$G = -2\pi_0\pi_1 \ln \left(1 - \frac{\bar{Q}_M}{2\pi_0\pi_1} \right) \quad (12)$$

and that under the Poisson model given by

$$G_m = -\ln (1 - \bar{Q}_m) \quad (13)$$

Equivalence between the u_0 -based and Q -based transcriptome distances

An intriguing question is to what extent the u_0 -based and Q -based transcriptome distances may differ. We address this issue as follows. Since the posterior probability $P(\text{DE}|p)$ conditional of p can be mathematically viewed as a function of p -value, one can calculate the conditional expectation of $P(\text{DE}|p)$ with respect to $\Phi(p)$ under the BUM model, denoted by $E[P(\text{DE}|p)]$. From Eq.(9) one can show

$$\begin{aligned} E[DE|p] &= \int_0^1 P(DE|p)\Phi(p)dp \\ &= (1 - u_0) \int_0^1 \phi(p)dp = 1 - u_0 \end{aligned} \quad (14)$$

When the number of genes (M) is sufficiently large, Eq.(14) implies that the average of Q 's given by Eq.(11) is expected to be

$$\bar{Q}_M = \sum_{k=1}^M Q_k / M \approx 1 - u_0 \quad (15)$$

Therefore, the u_0 -based transcriptome distance and the Q -based transcriptome distances are equivalent by the means of empirical Bayesian: to calculate the posterior probability $Q_k = P(\text{DE}|p_k)$, one may choose the prior u_0 as the empirical estimate from the data.

Q-based transcriptome distance for a small set of genes

Some important issues in evolutionary genomics require transcriptome distance analysis of a specific set (m) of genes; for instance, to evaluate a lineage-specific fast-evolving pattern. However, when the number of genes (m) is small, estimation of u_0 under the BUM model is notoriously difficult (Pounds and Morris 2003; Smyth 2004; Storey and Tibshirani 2003; Zhang and Grant 2004; Zhang 2011): our simulation study (see below) showed that estimation of u_0 could be subject to a large sampling variance as well as a large bias when m is around hundreds or less. We developed an algorithm that provides a practically feasible solution, which includes the following steps.

- (i) Estimation of u_0 and β by the methods we developed above, based on genome-wide (M genes) p -values.
- (ii) In the case of Markov-chain model, π_1 is estimated by the proportion of genes expressed in the transcriptome sample, and $\pi_0 = 1 - \pi_1$.
- (iii) Suppose we have m genes under study ($m < M$). Calculate the posterior probability of each gene being DE when the p -value is given, denoted by $Q_k = P(\text{DE}|p_k)$, $k = 1, \dots, m$. And the average of posterior probabilities of being DE over m genes is given by

$$\bar{Q}_m = \sum_{k=1}^m Q_k / m \quad (16)$$

- (iv) Let λ_m be the evolutionary rate of transcriptome for those m genes. In analogy to Eq.(12), under the Markov-chain model, the transcriptome distance defined by $G_m = (2\pi_1\pi_0)\lambda_m t$ can be estimated by

$$G_m = -2\pi_0\pi_1 \ln \left(1 - \frac{\bar{Q}_m}{2\pi_0\pi_1} \right) \quad (17)$$

while under the Poisson model the transcriptome distance defined by $G_m=2\lambda_mt$ is given by

$$G_m = -\ln(1 - \bar{Q}_m) \quad (18)$$

Sampling variance of the transcriptome distance

While it is straightforward to estimate the transcriptome distance for any set of genes by Eq.(17) or Eq.(18), the question is how we can evaluate the statistical property of the distance estimation. There are two resources of estimation errors: the sample size, i.e., the number of genes under study (m), and the statistical uncertainty of high throughput experiments. Each gene k may have some contributions to this high throughput uncertainty, approximately measured by $Q_k(1 - Q_k)$ (Efron 2001); indeed, the contribution is the lowest (zero) if $Q_k=0$ or 1 (no uncertainty), or the highest (0.25) if $Q_k=0.5$ (maximum uncertainty).

We then developed an approximate method to estimate the sampling variance of transcriptome distance. By the delta method, we showed that the sampling variance of Q_m is approximately given by

$$Var(\bar{Q}_m) \approx \frac{\bar{Q}_m(1 - \bar{Q}_m)}{m} + \frac{\sum_{k=1}^m Q_k(1 - Q_k)}{m} \quad (19)$$

The first term on the right hand of Eq.(19) represents the sample size effect, which approaches to zero when the number of genes $m \rightarrow \infty$. The second term represents the variance due to high throughput uncertainty, which generally cannot vanish except for all Q_k is zero or one. It follows that the variance of G_m under the Markov-chain model is given by

$$Var(\hat{G}_m) = \frac{Var(\bar{Q}_m)}{[1 - \bar{Q}_m/(2\pi_0\pi_1)]^2} \quad (20)$$

and that under the Poisson distance is given by

$$Var(\hat{G}_m) = \frac{Var(\bar{Q}_m)}{[1 - \bar{Q}_m]^2} \quad (21)$$

Simulation studies

We carried out simulation studies to evaluate the statistical performance of Q -transcriptome distances for any number of genes, which can be concisely described as follows. *Theme-A.* (i) Under a two-species scenario of transcriptome evolution, simulate the expression patterns of M genes according to a specified two-state Markov-chain model, and then calculate u_0 . (ii) Simulate the BUM distribution under this u_0 and a specified β for a p -value profile of M hypotheses. (iii) Estimate u_0 and β based on those simulated p -values. *Theme-B.* Repeat the first two steps in Theme-A for pre-specified m genes, under a given Markov-chain model and BUM model. After those m p -values are simulated, calculate the posterior probability Q_k for each gene, using the estimates of u_0 and β by step-(iii) in Theme-A. In both themes A and B, the simulation will be repeated for 1000 rounds and then evaluate the statistical performance in each simulation case.

Our main results are summarized in Table 2. (i) Estimations of both u_0 and β are statistically reliable and asymptotically unbiased as long as $M > 500$. (ii) Estimation of the Q -transcriptome distance is, in general, asymptotically unbiased. That is, the estimation bias tends to negligible as long as the number of genes (m) is sufficiently large. (iii) The estimation variance decreases with the increasing of the number (m) of genes, but cannot vanish ultimately, due the high throughput uncertainty. (iv) In the case when m is sufficiently large, say, $m > 1000$, the value of β is the key-factor to determine the accuracy of distance estimation. And (v) the variance of distance estimation given by Eq.(20) or Eq.(21) can be used as a good proxy in practice to the variance of distance estimation.

Discussion

Independent assumption in the BUM model

The BUM model implicitly assumes that the multiple test statistics are independent (Pounds and Morris 2003; Smyth 2004; Storey and Tibshirani 2003; Zhang and Grant 2004; Zhang 2011). Because this independent assumption is likely unrealistic in the context of transcriptome experiments, estimation of u_0 under the BUM would be biased. Indeed, estimating the proportion of null genes with nontrivial inter-gene dependence is beyond the scope of current study, which has been shown very difficult (Zhang 2011).

While this independent assumption makes our new methods of transcriptome distance practically feasible, one should be cautious that it might not be adequate in some extreme cases.

Underlying distribution of gene expressions

The new method for estimating the evolutionary distance of transcriptome are based on the BUM model of p -values. While the advantage is flexibility and wide applicability, the effect of the underlying gene-expression distribution on the accuracy of u_0 -estimation, as well as other parameters (β and β_0), remains largely unknown. It would be interesting to see how the underlying distributions affect the accuracies of these u_0 -estimation methods, especially how robust these p -value based methods are when the actual distribution of gene expression departs from the assumed theoretical distribution. This would form the basis of a future investigation (Pounds and Morris 2003; Smyth 2004; Storey and Tibshirani 2003; Zhang and Grant 2004; Zhang 2011).

Different methods for u_0 -estimation

The method here to estimate u_0 is nonparametric, which does not depend on the specific form of statistical tests being used, as long as the p -values pertaining to the tests are obtained. Since there are many different methods proposed to estimate u_0 , it remains an open question to what extent the estimation of the evolutionary distance of transcriptome is sensitive to different methods used. In the future, we will carry extensive computer simulations to examine the relative performance of our transcriptome distance estimation to the different methods of u_0 -estimation (Pounds and Morris 2003; Smyth 2004; Storey and Tibshirani 2003; Zhang and Grant 2004; Zhang 2011).

Effects of biological replicates

Both theoretical analysis and computer simulation has established an inverse relationship between the number of biological replicates (R) and the BUM parameter (β) (Zhang and Grant 2004; Zhang 2011). Our simulation study showed that the sampling variance of the estimated transcriptome distance may inflate considerably when $\beta=0.08$ or more. According to previous study, it implies that the new method tends to be

statistically unreliable when the biological replicates is small, say, $R=3$ or less. Hence, it is a challenge for an evolutionary researcher facing the prospect of biological replicates in order to achieve a reasonably accuracy of distance estimation.

Implementation

Under the *R*-environment, we implement a computational procedure to analyze the evolutionary pattern of transcriptome evolution based on p -values. At the current stage, the software is available upon request.

Table 1 Transcriptome distances between the human and chimpanzee brain regions

Methods	u_0 -transcriptome distance ^b						SOU distance ^d
	<i>DEseq2</i>		edgeR		limma		
Brain regions ^a	u_0	G	u_0	G	u_0	G	D
CB	0.825	0.2300	0.854	0.1813	0.881	0.1411	0.0899
HIP	0.850	0.1874	0.883	0.1381	0.961	0.0414	0.0974
STR	0.875	0.1502	0.917	0.0924	0.931	0.0760	0.0816
ACC	0.883	0.1380	0.893	0.1243	0.965	0.0367	0.0712
V1C	0.901	0.1136	0.936	0.0692	0.966	0.0353	0.0850
PMC	0.914	0.0961	0.962	0.0397	0.983	0.0177	0.0808
VPFC	0.905	0.1083	0.943	0.0616	0.968	0.0331	0.0653
DPFC	0.885	0.1347	0.934	0.0724	0.965	0.0363	0.0708
Mean ±s.e.	0.880 ±0.010	0.1448 ±0.0147	0.915 ±0.012	0.0974 ±0.0156	0.952 ±0.011	0.0522 ±0.0131	0.0803 ±0.0035

Notes. ^a brain regions: cerebellum (CB), hippocampus (HIP), striatum (STR), anterior cingulate cortex (ACC), primary visual cortex (V1C), premotor cortex (PMC), ventrolateral prefrontal cortex (VPFC), and dorsolateral prefrontal cortex (DPFC). ^b u_0 -transcriptome distance (G) based on three DE detection software packages (DESeq2, edgeR and Limma). ^d The transcriptome distance based on the stationary OU model.

Table 2. Simulation results of Q -transcriptome distances for given the number of genes (m).

Distance ^a	β ^b	$m=100$	$m=500$	$m=1000$	$m=\infty$
0.1	0.08	0.094±0.088 ^c	0.104±0.084	0.098±0.081	0.100±0.080
	0.05	0.097±0.050	0.103±0.042	0.099±0.041	0.100±0.040
	0.01	0.097±0.035	0.102±0.023	0.100±0.021	0.100±0.019
0.5	0.08	0.473±0.167	0.487±0.123	0.494±0.116	0.500±0.109
	0.05	0.479±0.134	0.488±0.071	0.505±0.059	0.500±0.044
	0.01	0.490±0.127	0.498±0.058	0.503±0.043	0.500±0.015
1.0	0.08	1.173±0.567	0.977±0.460	1.094±0.445	1.000±0.429
	0.05	1.104±0.453	0.958±0.309	0.965±0.286	1.000±0.261
	0.01	0.920±0.399	0.958±0.223	1.034±0.190	1.000±0.150

Notes. ^aThe true transcriptome distance is simulated according to a Markov-chain model between two species, where $\pi_1=0.70$ and $\pi_0=0.30$. ^b β is the parameter of BUM distribution; roughly speaking, a small value of β indicates a large biological replicates or *vice versa*. ^c The mean of Q -distance estimation and the derivation are based on 1000 simulations.

Fig.1. u_0 -Transcriptome distance between the human and chimpanzee brain regions, as well as the transcriptome distance based on stationary OU model (SOU for short)

References

- Bedford T, Hartl DL 2009. Optimization of gene expression by natural selection. *Proceedings of the National Academy of Sciences of the United States of America* 106: 1133-1138
- Brawand D, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478: 343-348
- Benjamini, Y. and Y. Hochberg (1995): “Controlling the false discovery rate-a practical and powerful approach to multiple testing,” *J Roy Stat Soc B*, 57, 289–300.
- Benjamini, Y. and D. Yekutieli (2001): “The control of the false discovery rate in multiple testing under dependency,” *Ann Stat*, 29, 1165–88.
- Cardoso-Moreira M, et al. 2019. Gene expression across mammalian organ development. *Nature*
- Cheng C, Pounds SB, Boyett JM, Pei D, Kuo ML, et al. (2004) Statistical Significance Threshold Criteria For Analysis of Microarray Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology* 3: Article 36.
- Efron, B. (2004): Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Am Stat Assoc*, 99, 96–104.
- Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001): “Empirical bayes analysis of a microarray experiment,” *Journal of Computational Biology*, 96, 1151–60.
- Enard W, et al. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* 296: 340-343
- Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440: 242-245
- GU J and GU X. 2003 Induced gene expression in human brain after the split from chimpanzee. *Trends in Genetics* 19: 63-65
- Gu X 2016. Understanding tissue expression evolution: from expression phylogeny to phylogenetic network. *Brief Bioinform* 17: 249-254
- Gu X, Ruan H, Su ZX, Zou YY 2017. Brownian model of transcriptome evolution and phylogenetic network visualization between tissues. *Mol Phylogenet Evol* 114: 34-39
- Gu X, Ruan H, Yang J 2019. Estimating the Strength of Expression Conservation from High Throughput RNA-seq Data. *Bioinformatics*
- Gu X, Su ZX 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *P Natl Acad Sci USA* 104: 2779-2784
- Gu X, et al. 2013. Phylogenomic Distance Method for Analyzing Transcriptome Evolution Based on RNA-seq Data. *Genome Biol Evol* 5: 1746-1753

- Hansen TF, Martins EP 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution* 50: 1404-1417
- King MC, Wilson AC 1975. Evolution at 2 Levels in Humans and Chimpanzees. *Science* 188: 107-116
- Lehner B 2013. Genotype to phenotype: lessons from model organisms for human genetics. *Nat Rev Genet* 14: 168-178
- Lemos B, Meiklejohn CD, Caceres M, Hartl DL 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: Stabilizing selection and variability among functional categories. *Evolution* 59: 126-137
- LOVE MI, HUBER W, ANDERS S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 [J]. *Genome Biol*, 2014, 15(12):
- McCarthy DJ, Chen YS, Smyth GK 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 40: 4288-4297
- Necsulea A, Kaessmann H 2014. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet* 15: 734-748
- NOWICK K, GERNAT T, ALMAAS E, et al. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain [J]. *P Natl Acad Sci USA*, 2009, 106(52): 22358-63.
- Pereira V, Waxman D, Eyre-Walker A 2009. A Problem With the Correlation Coefficient as a Measure of Gene Expression Divergence. *Genetics* 183: 1597-1600
- Pounds, S. and S. W. Morris (2003): "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values," *Bioinformatics*, 19, 1236–1242.
- RITCHIE ME, PHIPSON B, WU D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies [J]. *Nucleic Acids Res*, 2015, 43(7):
- ROBINSON MD, MCCARTHY DJ, SMYTH GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010, 26(1): 139-40.
- Ruan H, Su ZX, Gu X 2016. TREEEXP1.0: R Package for Analyzing Expression Evolution Based on RNA-Seq Data. *J Exp Zool Part B* 326: 394-402
- Smyth, G. K. (2004): "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical Applications in Genetics and Molecular Biology*, 1, 3.
- Storey, J. D. and R. Tibshirani (2003): "Statistical significance for genome-wide studies," *Proc Natl Acad Sci USA*, 100, 9440–9445.
- Sudmant PH, Alexis MS, Burge CB 2015. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol* 16: 287
- Wang Z, Gerstein M, Snyder M 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57-63
- Xu C, et al. 2018. Human-specific features of spatial gene expression and regulation in eight brain regions. *Genome research* 28: 1097-1110

Yang J, Ruan H, Zou Y, Su Z, Gu X 2018. Ancestral transcriptome inference based on RNA-Seq and ChIP-seq data. *Methods*

YANG J W, RUAN H, XU W J, et al. TreeExp2: An Integrated Framework for Phylogenetic Transcriptome Analysis [J]. *Genome Biol Evol*, 2019, 11(11): 3276-82.

Zhang SD, Gant TW (2004) A statistical framework for the design of microarray experiments and effective detection of differential gene expression. *Bioinformatics* 20: 2821–2828.

Zhang S-D (2011) Towards Accurate Estimation of the Proportion of True Null Hypotheses in Multiple Testing. *PLoS ONE* 6(4): e18874. doi:10.1371/journal.pone.0018874