LOOK ACADEMIC PUBLISHERS
OPEN ACCESS

INTERNATIONAL JOURNAL OF ENVIRONMENTAL & SCIENCE EDUCATION
2016, VOL. 11, NO. 18, 13001-13022

Possible Positive Selection on the Regulation of TMEM159-ZP2-CRYM in Human Cerebellum

Ganjcai Xie[a], Xi Jiang[a], Ning Fu[b] and Philip Khaytovich[a, b, c]

[a]Shanghai Institute for Biological Sciences, CAS, Shanghai, CHINA;
[b]Skolkovo Institute for Science and Technology, Skolkovo, RUSSIA;
[c]The Immanuel Kant Baltic Federal University, Kaliningrad, RUSSIA.

ABSTRACT

At this post-genome age, we have already known that human has quite similar genome compared to its closest extant cousin, chimpanzee. It had been proposed that the changes at the regulation level rather than at the protein coding sequence level play more important roles during human evolution. In this study, we focused on the genes that are conserved among human, chimpanzee and rhesus macaques, and examined the ones that possibly went through positive selection on gene expression regulation in human. Interestingly, our study revealed one previously un-characterized gene cluster TMEM159-ZP2-CRYM that is specifically regulated in the human genome. The genes in this cluster show dramatic age-related changes in human cerebellum, specifically, they are co-up regulated at early human cerebellum post-natal developmental stage and keep their high expression levels to the whole later life span. To carry out this inter-species gene expression comparison, we had developed a new method named BITS, which is based on high-throughput transcriptome sequencing data and can estimate the gene expression level for different species in more conserved and balanced way. Based on BITS method, we observed significant divergence to diversity ratio difference between protein-coding genes and pseudogenes as well as more species-specific up-regulated genes in human brain areas than in non-brain tissues. This study could be valuable for further functional study of human specific features as well as inter-species gene expression comparison. Finally, we show that the down-regulation of TMEM159-ZP2-CRYM is correlated with several human diseases, which might indicate their important functions in human cerebellum.

KEYWORDS
Human brain; sequencing (RNA-seq); transcriptome.

ARTICLE HISTORY
Received 27 July 2016
Revised 14 November 2016
Accepted 7 December 2016

Introduction

The interests to the study of human gene regulation changes could be date back to decades ago, when were proposed the changes in gene regulation level rather than at the protein-coding sequence level might play an important role for human evolution. Humans, compared to our closest extant relative chimpanzees, exhibit quite dramatic features, such as much larger brain size, sophisticated language systems, much more complex cultures, much higher intelligence and walking in up-right manner (Pelletier & Sonenberg, 1985; Sprinzl et al., 1998). However, to answer the evolutionary reason for these special features is still a daunting task. In this decade, with the advent of microarray technologies and especially the later coming high-

CORRESPONDENCE Ganjcai Xie ✉ gangcai@picb.ac.cn

throughput sequencing technologies, researchers could study human specific gene expression changes in much larger scale (Chen et al., 2005; Oeder et al., 2007; Salmena et al., 2011; Prescott & Proudfoot, 2002). The state-of-art situations for these studies could be found in published reviews. Generally, hundreds and even thousands of genes have been found specifically expressed in human tissues, and the shift of gene expression during early post-natal development in human had also been observed by our group's studies. Recent study had applied RNA-seq technology to the profiling of gene expression in major mammalian lineages and birds, which revealed different rates of gene expression evolution for different species and tissues (Katayama et al., 2005; Zhu et al., 2009; Yelin et al., 2003; Wang et al., 2005; Storz et al., 2005). However, this study used very limited samples to cover individual variations.

Compared to gene expression studies in one species, inter-species study usually suffers from several drawbacks. The microarray-based methods largely depends on known gene annotation and requires exact match of the probe to orthologous transcripts in order to eliminate inter-species bias, which may largely reduce the probe coverage for each gene and the detection power for certain divergent genes (Dan et al., 2002; Chu & Dolnick, 2002). And the current methods for inter-species gene expression analysis based on RNA-seq require mapping the reads to the genome of corresponding species (Spielmann et al., 2012; Smalheiser, 2012). However, different species may have quite different qualities of genome assemblies, which might introduce bias for higher quality species (Markham & Zuker, 2005; Liu et al., 2005; Griffiths-Jones et al., 2005). Furthermore, these methods either need to self-construct gene homologous relationship or rely on public homologous gene databases. In this study, we proposed a more conserved and balanced method by mapping the RNA-seq reads onto the consensus genome of studied species, which allows the mapping and annotation for concerned species done on one genomic level and also allows annotation-free analysis for these species.

Neutral evolution had been proposed for human gene expression in the previous studies, however, they were based on very limited number of background genes (pseudogenes). And further studies by other groups raised some concerns about this neutral model, where they found strong purifying selection for human gene expression. Using BITs method mentioned above, we could capture more expressed pseudogenes in our samples, which provide a chance to re-evaluate this issue.

Furthermore, to our knowledge, currently there has been no single report about gene cluster that shows age-related specific expression change in human tissues. And the genes predicted by previous studies as human-specific also had been rarely experimentally validated.

Materials and Methods

RNA extraction and sequencing

90 individual samples were collected in this study, which include 5 tissues (cerebellar cortex: CBC, prefrontal cortex: PFC, visual cortex: VC, kidney: K and muscle:M) of 3 species (human, chimpanzee and rhesus monkey) and were dissected from frozen postmortem tissue of healthy adult individuals. The source of the sample can be found in Table 1. After dissection, total RNAs were extracted by using Trizol Kit (Invitrogen), and polyA(+) RNAs were sorted out by oligo-dT beads. RNA-seq library was prepared following TruSeq library preparation protocol and the sequencing was done on illumina HiSeq 2000.

Table 1. Down-regulation of ZP2 region genes is significantly enriched in human diseases.

Human cerebellum related ZP2 expression change in NextBio Disease Atlas			
Public ID	Compare type	Fold change	p value
GSE28192	Medullablastoma primary tumors vs normal adult cerebellums	-37.4	1.90E-09
GSE28192	Anaplastic medulloblastomas vs normal adult cerebellums	-38.8	4.60E-09
GSE28192	Classic medulloblastomas vs normal adult cerebellums	-37.3	4.40E-08
GSE3790	Cerebellum- Huntingtons disease grade 2 vs control_GPL96	-1.96	1.00E-07
GSE28192	Large cell medulloblastomas vs normal adult cerebellums	-38.5	8.90E-06
GSE28192	Nodular medulloblastomas vs normal adult cerebellums	-30.4	1.20E-05
GSE35978	Cerebellum of patients with bipolar disorder vs unaffected controls	-1.62	0.0005
GSE3790	Cerebellum- Huntingtons disease grade 1 vs control_GPL96	-1.55	0.0028
GSE35978	Cerebellum of patients with schizophrenia vs unaffected controls	-1.28	0.0072
GSE38322	Cerebellum of autistic patients vs normal controls	-1.67	0.0377
Human cerebellum related CRYM expression change in NextBio Disease Atlas			
Public ID	Compare type	Fold change	p value
GSE28192	Anaplastic medulloblastomas vs normal adult cerebellums	-21.6	1.60E-09
GSE28192	Medullablastoma primary tumors vs normal adult cerebellums	-20.9	3.80E-06
GSE28192	Classic medulloblastomas vs normal adult cerebellums	-20.2	4.40E-06
GSE21687	Ependymoma primary tumors from posterior fossa vs cerebellum	-3.24	4.40E-06
GSE3790	Cerebellum- Huntingtons disease grade 2 vs control_GPL96	-1.46	4.50E-06
GSE4058	Glioblastoma multiforme solid tumor vs normal cerebellum	-18.3	0.0002
GSE38322	Cerebellum of autistic patients vs normal controls	-2.43	0.0005
GSE28192	Nodular medulloblastomas vs normal adult cerebellums	-21	0.0005
GSE28192	Large cell medulloblastomas vs normal adult cerebellums	-21	0.0005
GSE4058	Oligoastrocytoma solid tumor vs normal cerebellum	-25.1	0.0112
GSE3790	Cerebellum- Huntingtons disease grade 1 vs control_GPL96	-1.22	0.0149
Human cerebellum related TMEM159 expression change in NextBio Disease Atlas			
Public ID	Compare type	Fold change	p value
GSE28192	Anaplastic medulloblastomas vs normal adult cerebellums	-3.73	9.00E-08
GSE28192	Medullablastoma primary tumors vs normal adult cerebellums	-3.73	2.50E-07
GSE28192	Large cell medulloblastomas vs normal adult cerebellums	-6.35	5.20E-05
GSE28192	Nodular medulloblastomas vs normal adult cerebellums	-3.4	0.0004
GSE4058	Oligoastrocytoma solid tumor vs normal cerebellum	-3.13	0.0181

BITS method and gene expression calculation

In the procedure of BITS for HCR study, the 3 species genome-wide alignment was constructed by multiz using blastz chained and netted pairwise alignment files downloaded from UCSC (hg19vsRheMac3 and hg19vsPanTro3). The mask was done on hg19 according to the multiple alignment result, and the sites that show discordances or in six bases flanking region of insertion/deletion sites were masked as N. All the reads passing illumina quality control were combined together with sample information added, and were then mapped onto the BITS constructed consensus genome by RNA-seq mapper STAR. Gencode V14 was used for human gene annotation (ClustalW2, 2016), and the non-redundant exons (NE) were computed based on this annotation. The number of reads located on each NE was

counted, and if the reads is partially overlapped with NEs, then the overlapping fraction was used for the counting. Only the genes with more than 10 reads on average in at least one tissue were considered in the following analyzing. Next, TMM method was used for counts normalization and the normalized counts was further normalized to reads per million (RPM) by total number of normalized counts in all the genes.

EdgeR was used for differential gene expression calculation for pairwise inter-species comparison, where TMM method was used for reads counts normalization, and the p values were further corrected by Benjamini-Hochberg method.

For one given tissue type, human specific change index is defined as:

$$RC_{median} = Median(C_{median}, R_{median})$$

$$hSindex = \frac{(H_{median} - RC_{median})}{Min(H_{median}, RC_{median}) + 0.5}$$

And chimpanzee specific change index is defined as:

$$RH_{median} = Median(H_{median}, R_{median})$$

$$cSindex = \frac{(C_{median} - RH_{median})}{Min(C_{median}, RH_{median}) + 0.5}$$

Where

H_{median}= median RPM of humans

C_{median}= median RPM of chimpanzees

R_{median}= median RPM of rhesus macaques

For each tissue, the sDEGs satisfy three criteria: first, between species differential expression p value (BH adjusted) should be less than 0.01, second, Sindex absolute value should be larger than 1 and third the RPMs of all the biological replicates in considered species should be consistently larger or smaller than the other species.

Three-way reciprocal LiftOver

Gencode V14 was used as human gene annotation, and NEs were constructed based on this annotation as described above. Human NEs genomic coordinates (hg19) were first converted into chimpanzee genomic coordinates (panTro3) (Q05996 (ZP2_HUMAN), 2016), and then chimpanzee NEs genomic coordinates were converted into rhesus macaque genomic coordinates (rheMac3). And the reverse conversion processes were carried out using the same method, and at each process only the NEs with 1-1 matches were kept and the final converted NEs in human should have the same coordinates as the original human NEs. The gene structures in each species were constructed by final NEs after 3-way reciprocal LiftOver.

Real time PCR and in-situ hybridization

The RNA extraction for five tissues of the three species (each with 3 adult individual replicates) was performed the same as in the sequencing part, further sample preparation was done according to Roche LightCycler 480 SYBR Green I Master protocols (500 reactions) and then real-time PCR was carried out in eppendorf epMotion 5070. GAPDH was used as control for expression normalization.

Following are the primers information for ZP2 and GAPDH:

>ZP2_Forward
CAACCTTATGGGGAAAACGA
>ZP2_Reverse
TGATCAGGATGGGTCACAGA
>GAPDH_Forward
GAAGATGGTGATGGGATTTC

>GAPDH_Reverse
GAAGGTGAAGGTCGGAGTC

The slices for one adult human cerebellum and adult rhesus macaque cerebellum were cut by machine name. The RNA in-situ hybridization (ISH) was performed according to standard ISH protocols, where pGEM-5Zf(+) was used to construct the vectors containing Zp2 probe sequences. The same primers were used as in real-time PCR. The probe sequences had been validated by Sanger sequencing, the information of which can be found in the supplementary supporting information.

The definition of real-time PCR relative expression level is:

$$RE = \frac{E_{query}}{Mean(E_{RCBC})}$$

Where,

$$E_{query} = 2^{(CP_{query} - Mean(CP_{GAPDH}))}$$

$$E_{RCBC} = 2^{(CP_{RhesusCBC} - Mean(CP_{GAPDH}))}$$

Variance and trio analysis

In the analysis of gene expression variance explained by each factor, anova (model ~ species+tissue+gender) was used for variance calculation, and the sum of square (SS) for each factor and total sum of squares (TSS) were extracted. The percentage of variance explained by each factor is defined as:

$$If \quad there \quad are \quad n \quad genes \quad and \quad m \quad factors:$$

$$variance \quad \exp lained \quad for \quad factor \quad k \quad = \frac{\sum_{i=1}^{n} SS_{ik} / n}{\sum_{i=1}^{n} TSS_i / n}$$

$$Where \quad for \quad gene \quad i: \; TSS_i = \sum_{k=1}^{m} SS_{ik} + SS_{i\_residue}$$

In the analysis of the relationship between BV and WV, for each gene in each tissue of each pairwise species (human-chimpanzee, human-rhesus or chimpanzee-rhesus) study, anova analysis (model ~ species) was applied. Then, BV is defined as the Mean of Square (MS) for species factor and WV is defined as MS of residuals.

In the trio-analysis of human specificity, for each gene in each tissue, the samples were grouped into either human or nonhuman, and anova analysis (model~group) was applied. Then, BV is defined as the MS for group factor and WV is defined as the MS of residues. The ratio of divergence to polymorphism (D/P ratio) is equal to the log2-transformed ratio of BV/WV. For each gene, the tissue with highest D/P ratio (called target tissue) was used for human enrichment and tissue specificity calculation. In human samples, we calculated the average RPM fold change of target tissue to each of other tissues, and the tissue specificity was defined as the smallest fold change with log2 transformation. For the target tissue samples, the average RPM fold changes between humans and chimpanzees or between humans and rhesus monkeys were calculated, and the smaller one of which was defined as human enrichment. The same procedure was applied to the trio-analysis of chimpanzee specificity, where the focus was on chimpanzee rather than on human.

Results and Discussion

BITS: a new method for inter-species gene expression comparison

We proposed a new method for inter-species gene expression comparison, which is named as Balanced Inter-species Transcriptome Study (BITS) method (figure 1). Basically, all the sequenced reads from different species are mapped to the same consensus genome, which is constructed based on multiple-species genome-wide alignment. In order to better use annotation information, one of the well-studied and annotated species can be chosen as reference genome for consensus genome construction, where inter-species discordant sites according to genome-wide alignment are masked as N. After mapping, similar strategies for one-species gene expression study could be applied to this inter-species comparison, including the estimation of gene expression level (bottom box in figure 1).

Currently, there are several ad hoc methods for such comparative study. One method is homologue-based, which needs to first construct the homologous genes of all the species considered. In this method, all the species studied should have their genome sequenced, and RNA-seq reads were mapped to the genome of the corresponding species. In order to making clear conclusion, these studies also only keep the homologs that have one to one relationship among species. After the calculation of gene expression separately in each species for each one-to-one homologue, the expression values were combined for inter-species comparative study. Although such homologous information can be downloaded from public available database, such as Ensembl (2016), not all the species with available genomes are included in these orthologous databases. Another method is based on inter-species genomic region matching, which usually uses Liftover (2016). Similar as the first method, only the genes with one to one relationship matching is kept for the inter-species gene expression comparison. The third method combines the homologs from the first method and does the further local alignment based on mapping tools, such as BLAT (2016) or BLAST (2016).

However, different species might have quite different qualities of genome assemblies, for example in HCR (human-chimpanzee-rhesus) comparison, the number of unplaced genomic fragments in current UCSC genome assemblies, for human (hg19), chimpanzee (panTro4) and rhesus (rheMac3), is 59, 24103 and 34081 respectively, which indicates human genome has much better quality than others. All of afore mentioned methods require mapping the reads onto their own genomes, which could provide bias due to different quality of genome assemblies. Another aspect is that, for high-throughput sequencing data analyzing, researchers commonly require visualizing the raw sequencing reads for their interested genes or genomic regions. Due to using different genomes, it's not easy to visualize the raw reads from different species directly. Furthermore, in certain situations, annotation-free analysis would be needed.

The liftover method had been widely used in previous inter-species studies, as a proof of concept, we applied this method and our method to 18 RNA-seq PFC samples, which are composed by equal number of adult HCR individuals. We first calculated the number of expressed genes that can be captured by each method. As shown by figure 2A, BITS can consistently capture more expressed genes compared to Liftover method at different expression cutoffs. However, BITS could sacrifice the accuracy of gene expression estimation for gaining more expression signals. In order to solve this problem, we then calculated inter-species gene expression correlation for both methods. The result (figure 2B) indicates that BITS can significantly improve the inter-species gene expression correlation. And interestingly, BITS shows more balanced inter-species gene expression level estimation than Liftover, as shown by

figure 2B, the difference between CR (chimpanzee-rhesus) gene expression correlation and HR (human-rhesus) gene expression correlation is much smaller based on BITS calculation than based on Liftover.

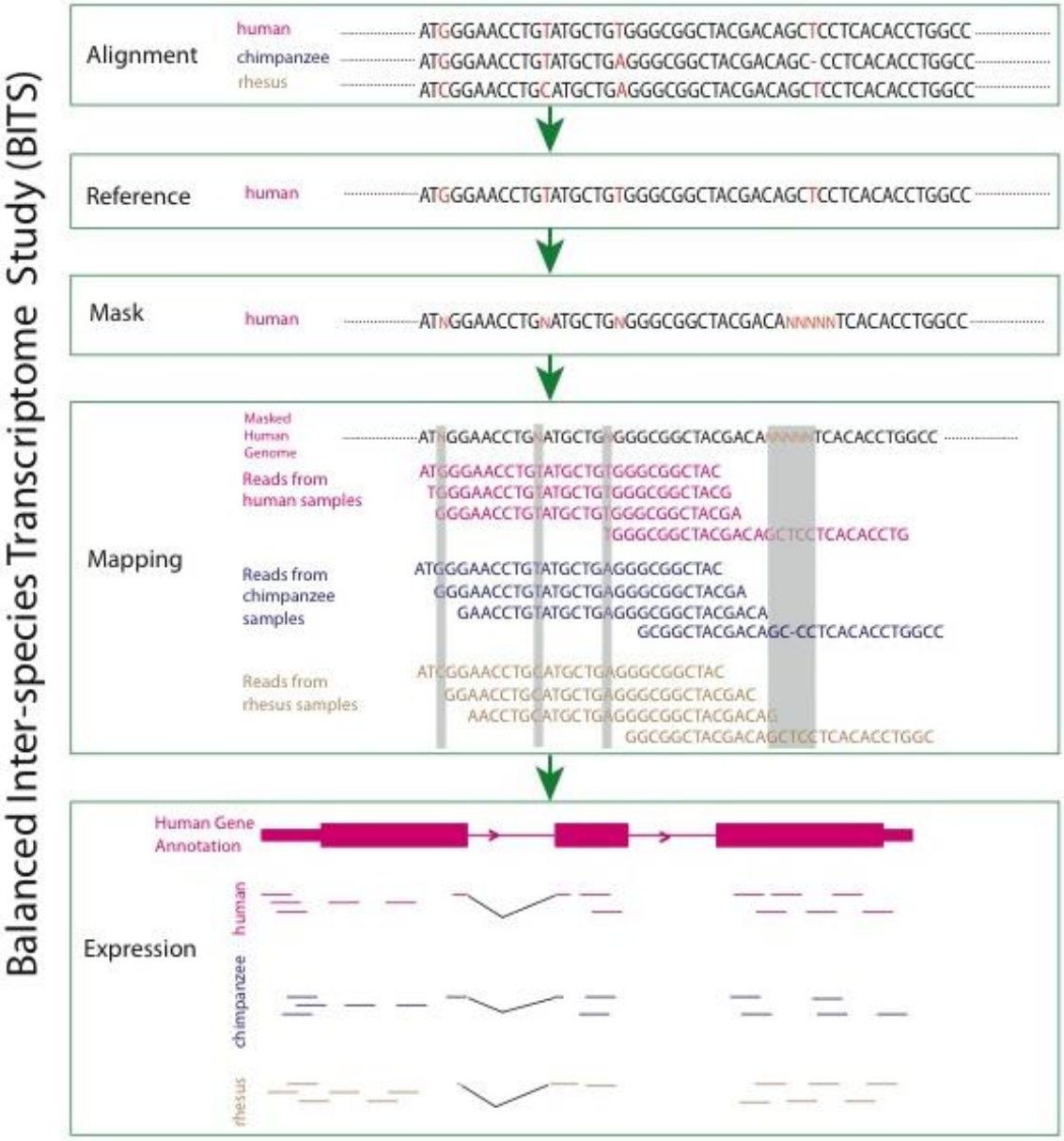

Figure 1. Brief introduction to BITS method. Multiz was used for multiple genome alignment, and the sites showing discordant alignments were masked on the reference genome. The reads from all the species were pooled together and mapped onto the masked genome (or consensus genome) by STAR mapper. The expression level of reference gene was calculated for each species at the final stage.

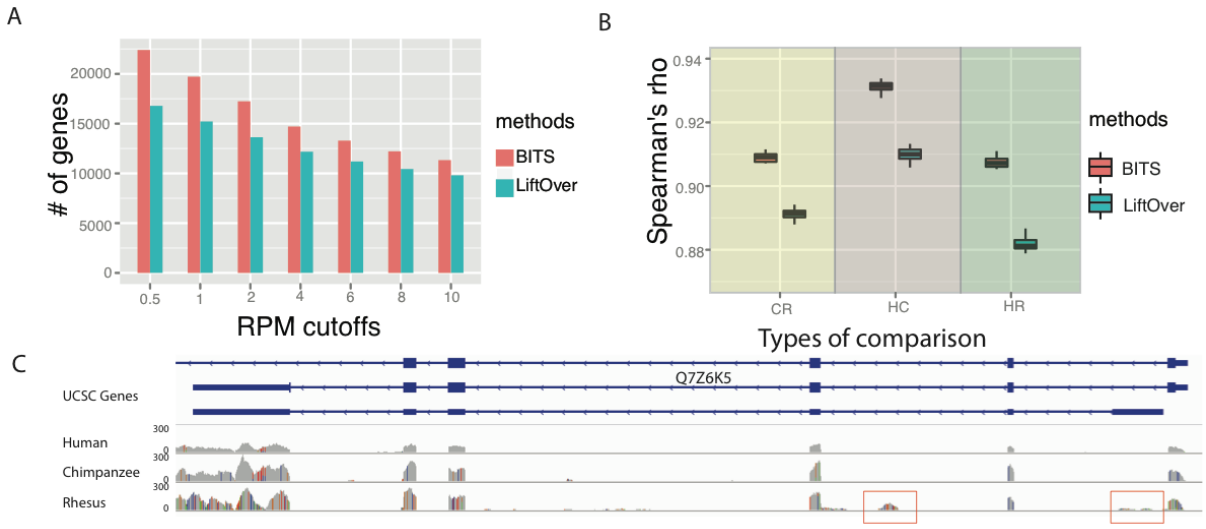

Figure 2. BITS method comparison and visualization. (A) Number of expressed genes comparison at different expression level cutoffs. RPM (Reads Per Million, see method section for detail). (B) Inter-species gene expression correlation comparison. CR: chimpanzee-rhesus comparison, HC: human-chimpanzee comparison, HR: human-rhesus comparison. (C) Example of visualizing inter-species raw sequencing reads based on BITS mapping result. The regions emphasized by red rectangle are species specific-spliced regions. The RNA-seq reads are stacked according to the mapping information and IGV tool was used to color the sites showing discordance with the human reference genome.

Furthermore, as shown by figure 2C, BITS makes it possible to visualize raw-sequencing information for different species, which is important for single gene study.

Possible positive selection on the regulation of brain protein-coding genes

Through previous studies, both of neutral evolution and purifying selection had been proposed as main forces for recent human gene expression evolution. Our result here supports neutral evolution of gene expression for majority human genes, but also revealed some human brain protein-coding genes could be possibly under selection pressure. We applied afore mentioned BITS method to RNA-seq data of 90 samples, which are composed by five tissues (PFC, VC, CBC, K and M) of three species (human, chimpanzee and rhesus monkey) with equal number of adult individuals (sample information is available in Table 1). On average, we got about 16 million 100bp-reads for each sample, among which about 70% can be mapped to the masked genome and 65% can be uniquely mapped (Table 1 and 2). For each tissue, more than 15 thousand genes are expressed (RPM >1), and no bias is observed for power of expressed gene detection for different species and high inter-species gene expression correlation is observed (Figures 3, 4, and Table 3). It is consistent with previous study that tissue explains the largest part of the total gene expression variation, and then species (see figure 5 for PCA analysis and figure 6A for variance analysis). Interestingly, we observed high correlation of between-species variation (BV) and within-species variation (WV) based on human-chimpanzee comparison (figure 6B shows HC PFC protein-coding gene comparison, for others tissues and non-protein coding genes please check supplementary figure 7 and figure 8), which support the neutral evolution as a major force for recent human gene expression evolution.

Table 2. Summary of sequencing and mapping

Sample type	Average Total Reads	Average reads mapped	Average Reads uniquely mapped	Average Mapping %	Average Unique mapping %
cC	17,269,442	13,020,638	13,020,638	75.50%	70.10%
cK	17,474,363	11,755,715	11,755,715	67.00%	60.20%
cM	20,210,198	14,981,782	14,981,782	74.00%	67.80%
cP	13,671,724	9,952,590	9,952,590	72.80%	67.40%
cV	15,923,475	11,374,334	11,374,334	71.40%	66.00%
hC	14,885,291	11,301,948	11,301,948	76.10%	69.90%
hK	15,090,163	10,236,693	10,236,693	68.20%	61.20%
hM	14,714,657	10,392,099	10,392,099	70.60%	65.10%
hP	15,730,647	11,582,775	11,582,775	73.40%	68.10%
hV	15,470,217	10,769,182	10,769,182	69.00%	63.40%
rC	15,316,850	10,864,546	10,864,546	71.00%	66.50%
rK	19,249,031	12,166,981	12,166,981	63.00%	57.40%
rM	19,260,540	12,970,151	12,970,151	68.00%	62.60%
rP	13,532,665	9,286,830	9,286,830	68.30%	63.70%
rV	14,560,426	10,203,255	10,203,255	70.20%	65.90%

The expression of majority pseudogenes had been proposed to be under neutral evolution, which could be used as a neutral background for the study of the selection signal in other genes. It had been shown in the previous study that there was no significant difference between the ratio of divergence to diversity ratio for protein coding genes and pseudogenes. However, in that study, only 23 expressed pseudogenes were considered. Here, we combined RNA-seq and BITS method, and captured much more expressed pseudogenes (around 1500, see figure 6C). As shown in Figure 6C, we observed strong difference between brain region tissues and non-brain tissues for the difference of protein coding genes and pseudogenes in BV to WV ratios. One side Wilcoxon rank-sum test (protein coding genes > pseudogenes) was applied for each tissue, and only the differences in brain regions are statistically significant ($p<0.05$).

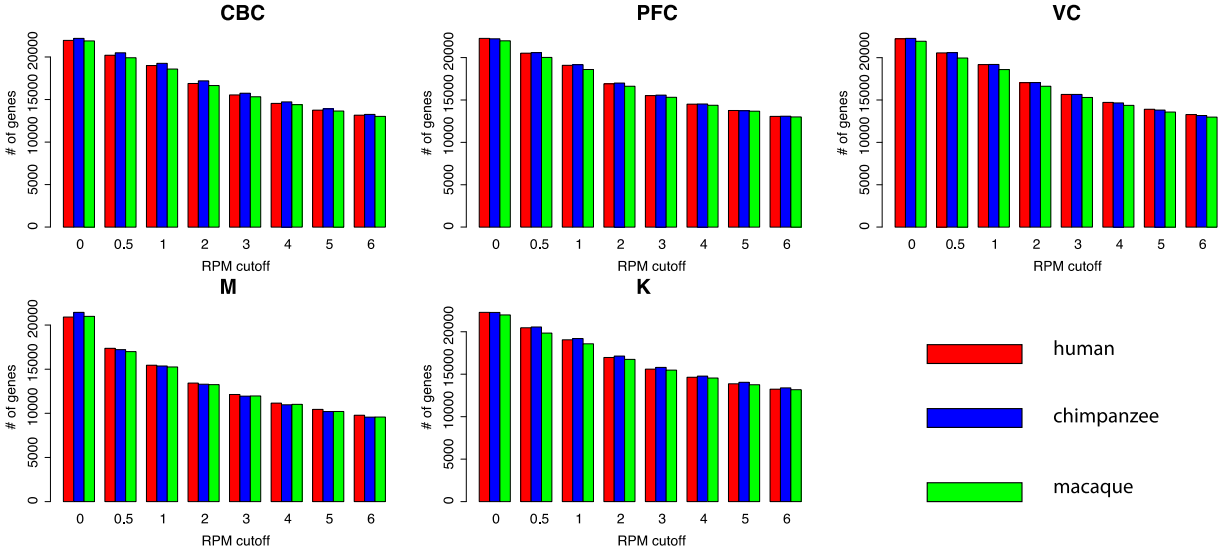

Figure 3. Number of expressed genes at given RPM cutoff. Each subfigure shows the gene number counting information for each tissue of the 3 species. At give RPM cutoff, the number of genes expressed in certain tissue of certain species is defined as the number of genes with median RPM (median of the 6 biological replicates) larger than this cutoff. X-axis and y-axis represent RPM cutoff value and the number of genes above this cutoff respectively, and different colors mean different species as shown in the legend bars.

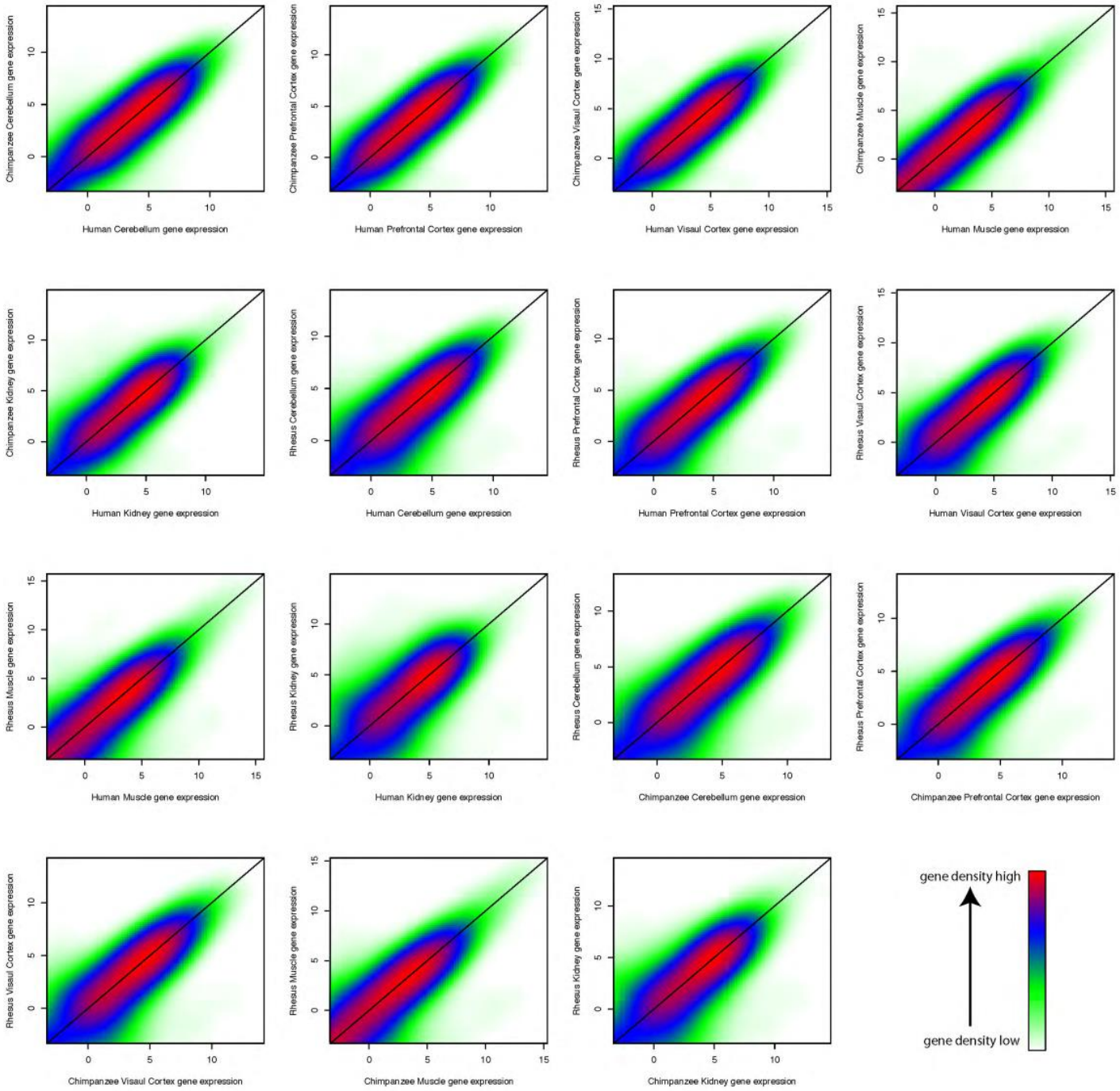

Figure 4. Scatterplot between species in each tissue. Only genes with average reads count in each tissue larger than 10 (or Expressed Genes) were used in the comparison, and expression level was based on the gene median RPM in each tissue of each species. Red color means high density of genes in given region and green color means low density of genes in given region. And the plotting area is square with y=x diagonal line drawn. X-axis and y-axisrepresentcorrespondent gene median RPM in given tissue of the two compared species.

Table 3. Spearman's rho between species in each issue. Only genes with average reads count in each tissue larger than 10 (or Expressed Genes) were used in the calculation of the correlation, and the calculation was based on the gene median RPM in each tissue of each species. (HC: Human-Chimpanzee correlation, HM: Human-Macaque correlation, CM: Chimpanzee-Macaque correlation).

	HC	HM	CM
CBC	0.954	0.901	0.912
PFC	0.965	0.923	0.926
VC	0.962	0.913	0.923
M	0.949	0.902	0.911
K	0.928	0.867	0.887

The study based on variation analysis doesn't tell about which species had gone through natural selection, could be either chimpanzees or humans, or even both. Using rhesus macaque as out-group, we examined human or chimpanzee specific expressed genes. Interestingly, we found there are more brain region species-specific higher expressed genes than non-brain tissue related ones for both of humans and chimpanzees (figure 6D), which indicates possible natural selection for the regulation of brain-related genes in both species.

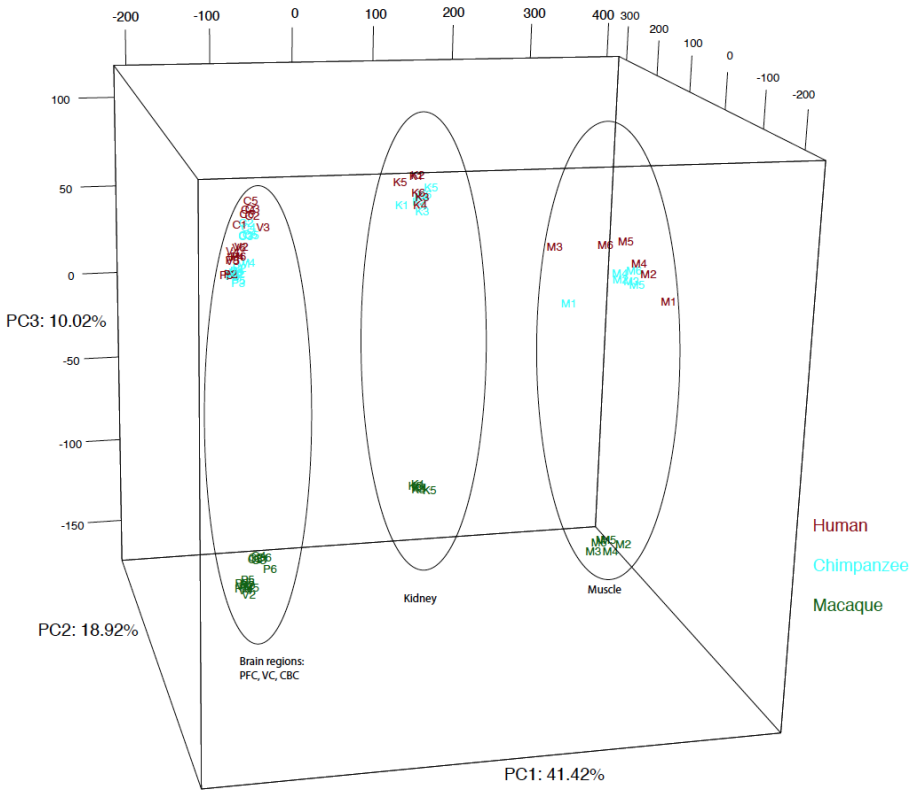

Figure 5. PCA based on Expressed Genes. The first 3 principal components were drawn on respective dimensions, with PC1 represents the first principal component, PC2 the second one and PC3 the third one. The 3 species were colored as shown, and the tissue names with sample number were drawn in correspondent locations. Three ellipse lines were drawn for better grouping different tissues separated in the PC1.

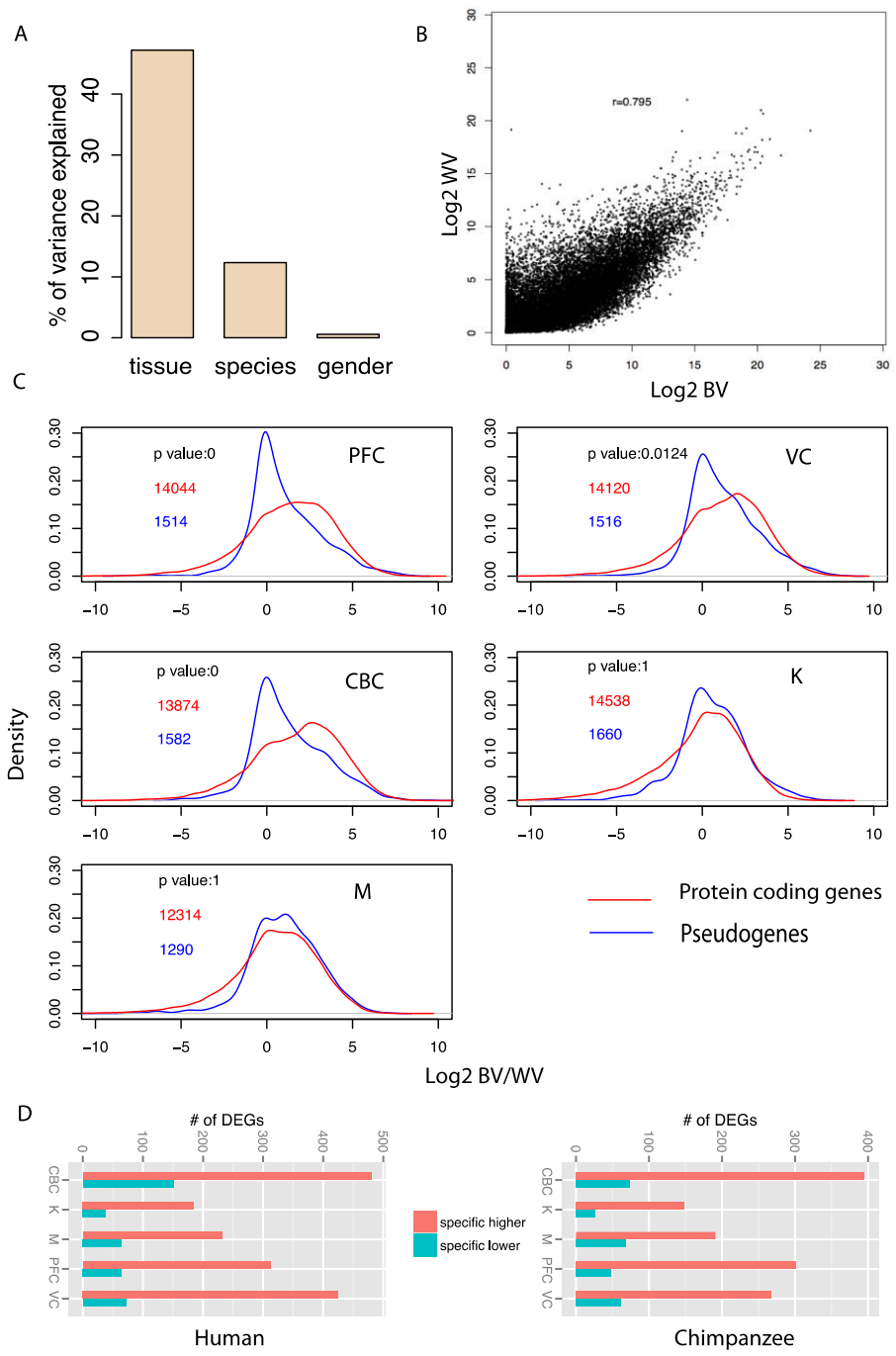

Figure 6. Human brain-expressed proteins were under possible positive selection. (A) Percentage of total variance explained by each factor. See detail calculation in method section. (B) Linear relationship between WV and BV. WV: within-species gene expression variation; BV: between-species gene expression variation. The analysis was done between human and chimpanzee PFCs, and both of WV and BV were log2 transformed. (Check Figure 7 and 8 for other tissues and non-protein coding genes) (C) Comparison between protein coding genes and pseudogenes for human-chimpanzee BV to WV ratio difference. Only the genes with average RPM larger than 1 were considered in the comparison. One side (red > blue) Wilcoxon rank-sum test was applied for BV/WV distribution of protein coding genes and pseudogenes. The colored number in each sub-figure represent the number of expressed protein coding genes (red) or pseudogenes (blue). (D) Species-specific expressed genes. Rhesus macaque was used as out-group for species-specific expression examination.

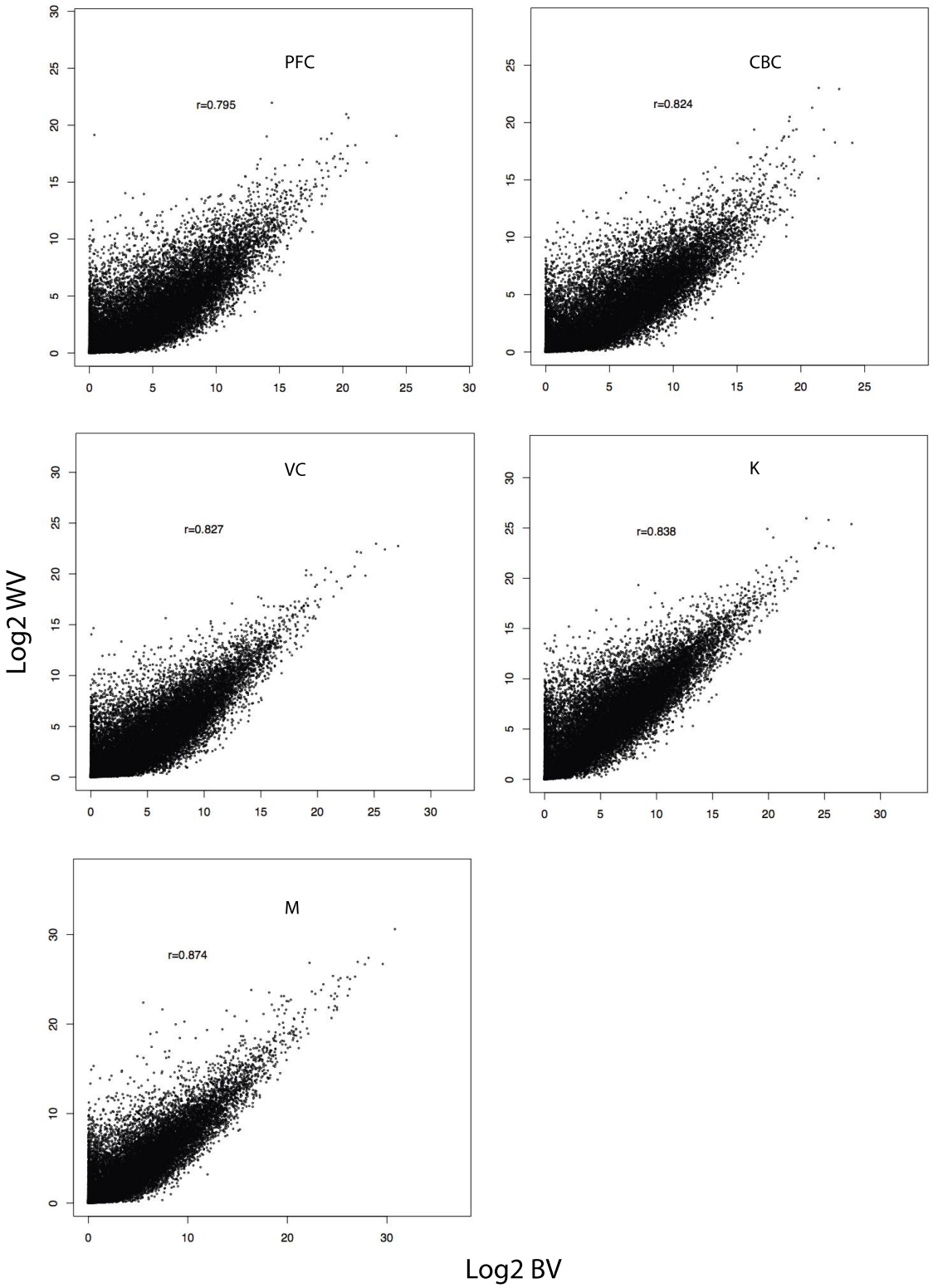

Figure 7. Linear correlation between BV and WV for human-chimpanzee protein coding genes.

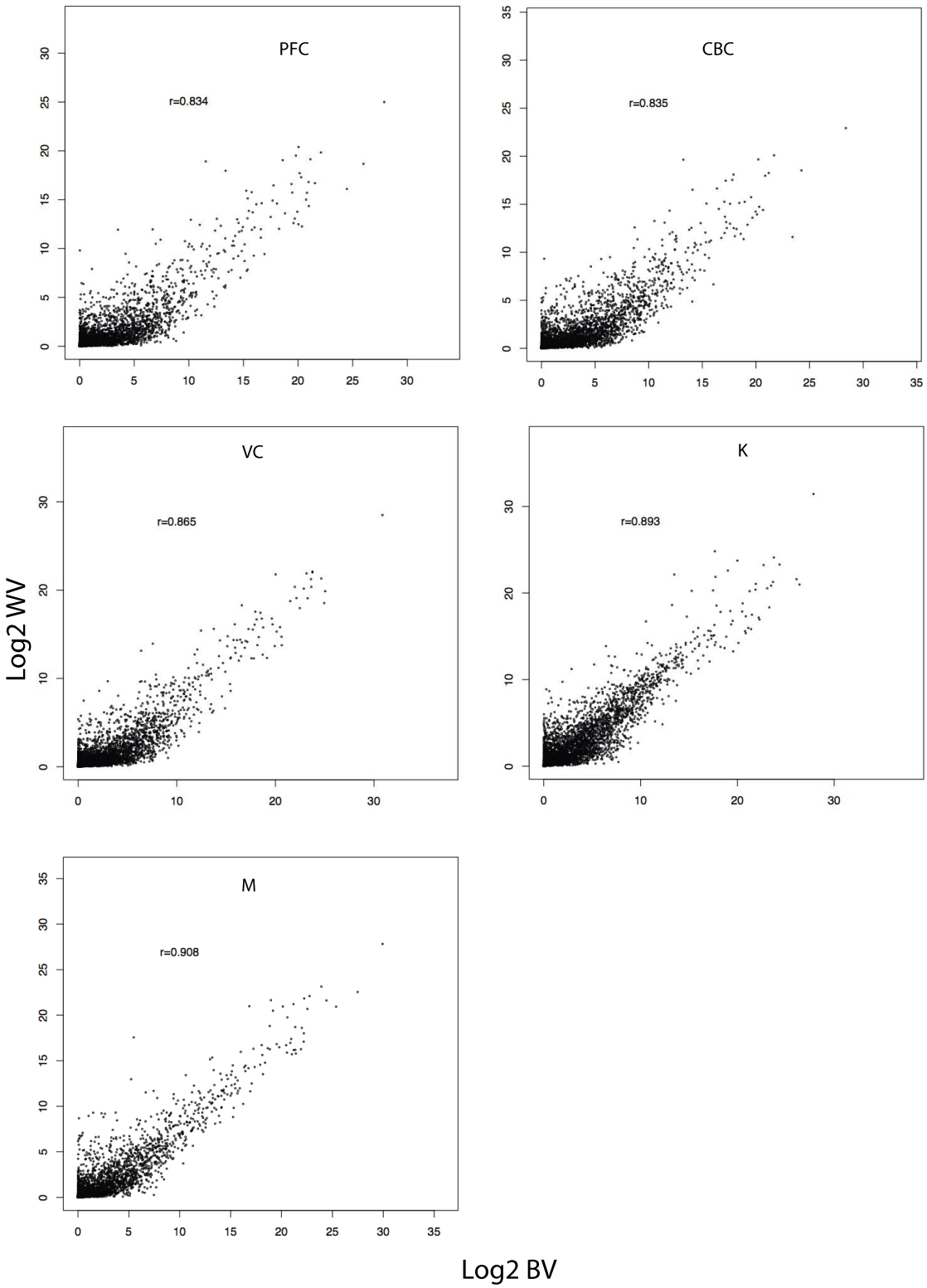

Figure 8. Linear correlation between BV and WV for human-chimpanzee non-protein coding genes.

There are two possible explanations for the higher level of divergence to diversity in brain-expressed genes. One is that part of brain-expressed genes underwent recent stabilizing selection after human-chimpanzee separation from theirmost recent common ancestors (MRCA), which made the WV smaller than expected by chance. Another explanation is that there was positive selection for part of the brain-expressed genes during the separation of humans and chimpanzees from their MRCAs, which could enlarge the divergence of these genes between the two species.

Dramatic change of ZP2 expression in human cerebellum

The genes that show the most dramatic expression change in humans are more likely under selection in human lineage. So, we examined all the genes that are expressed in our samples to search for the ones that show both of higher species specificity and tissue specificity. In order to do this, we constructed a trio-analysis of the gene expression change among species and tissues (method part for detail information). Basically, in this trio-analysis, we had considered three factors, which are divergence to diversity (or polymorphism) ratio change, tissue specificity and the abundance fold change. As illustrated by the three-dimensional plot for this trio-analysis (Figure 9A), Zp2 is the gene that is within the top quantile value (top 1%) for all three trio-analysis factors and especially it shows the highest abundance fold change. However, such kind of dramatic change in the three levels only happened in human brain region, and non-brain tissues or chimpanzee tissues don't contain such gene (Figure 9A). Further examination of ZP2 expression based on our RNA-seq shows that ZP2 is highly expressed in human cerebellum but barely expressed in other brain regions or non-brain tissues (figure 9B). Such kind of human cerebellum specificity was also confirmed by real-time PCR experiments (figure 9C and figure 10). We then ask if such kind of specificity still holds true when comparing to other primates or even mammalians. To answer this, we checked ZP2 expression in the RNA-seq data (Figure 11), and as expected, Zp2 is only highly expressed in human cerebellum. It is barely or not expressed in heart, kidney, liver and testis or other species, including non-mammalian chicken.

Human cerebellar cortex is a complex system, which can be basically divided into molecular layer, purkinje layer and granular layer from surface to the inner part. The granular layer majorly contains densely organized granule cells, purkinje layer contains a thin layer of purkinje cells above granular layer, and molecular layer is sparsely filled with the cells, such as stellate and basket cells. In order to study the expression of ZP2 in different layers, we carried out RNA in-situ hybridization in human cerebellar cortex. As shown in figure 12A, ZP2 is highly expressed in granular layer and purkinje cells, and further information shows that it is majorly expressed in purkinje cytoplasm areas rather than in nucleus areas (Figure 12B). The high expression level of ZP2 in Purkinje cells is also confirmed by the microarray data for this cell type (Figure 13).

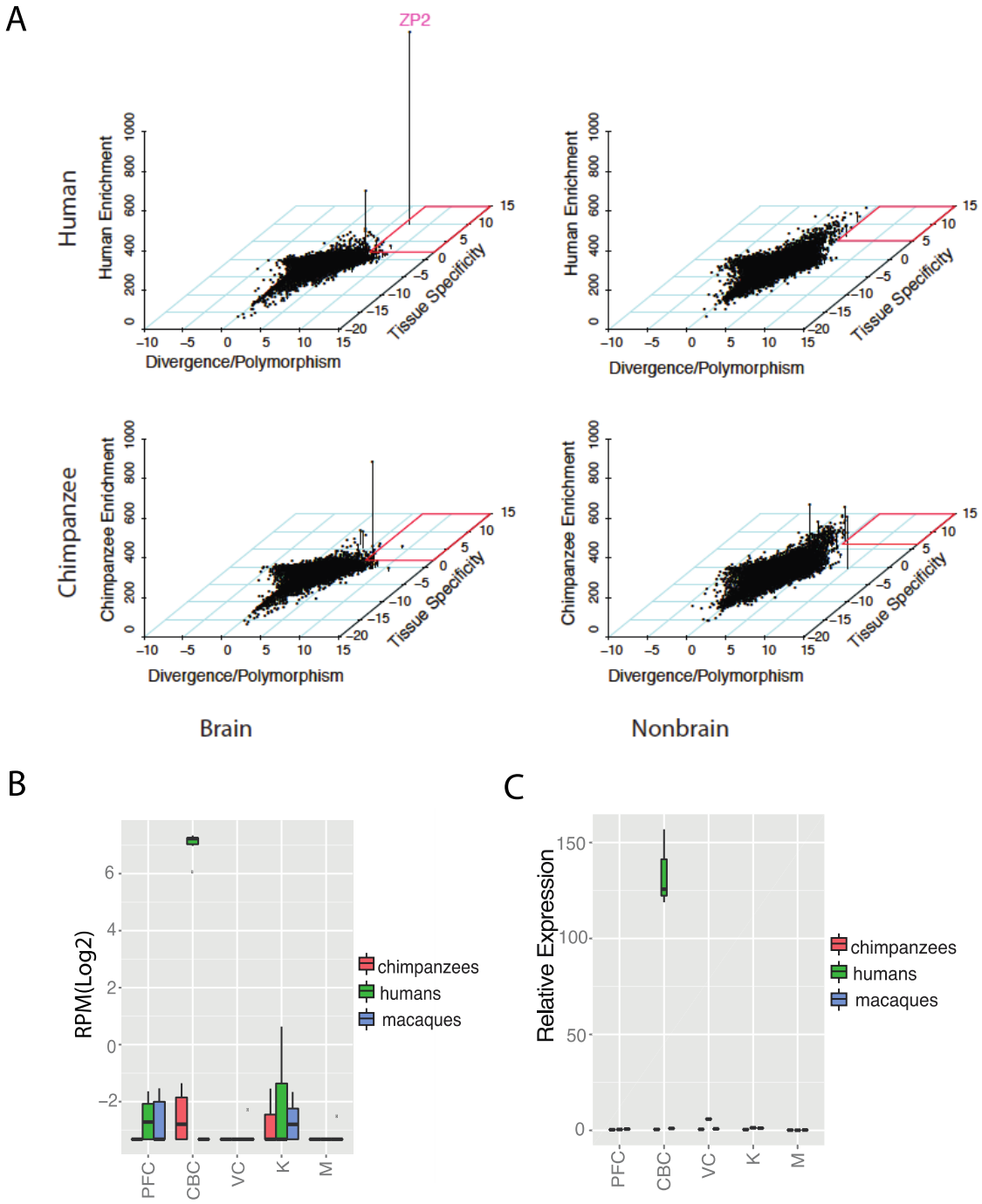

Figure 9. Human-specifically dramatic change of ZP2 expression. (A) Trio-analysis reveals high specificity of ZP2. Divergence to polymorphism ratio (D/P) was calculated for each gene in each tissue, and the tissue with largest D/P ratio (X axis) was selected for both of tissue specificity (Y axis) and species enrichment analysis (Z axis). Brain (left two subfigures) means the D/P ratio selection was done on PFC, VC and CBC and nonBrain (right two subfigures) means the selection was based on K and M. The dots with highest 1 percentile D/P ratio and tissue specificity are located at the red rectangle region. See more information at the method section. (B) Boxplot illustration of human ZP2 expression specificity based on RNA-seq. The Y-axis value is the logarithm of RPM value to base 2. (C) Real-time PCR validation of ZP2 expression specificity. The definition for relative expression level can be found in the method part, and the raw Cp values are available in Table 1.

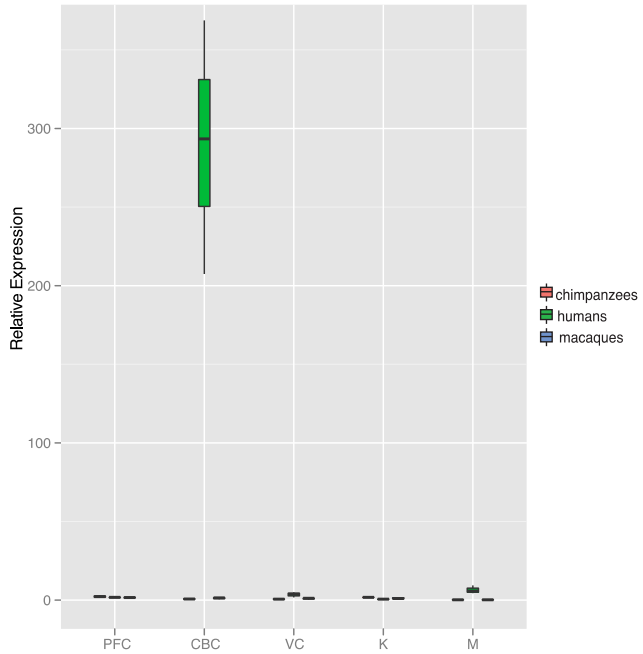

Figure 10. Biological replicates for qPCR validation of ZP2 expression specificity. Different individuals were used compared to the ones shown in the main text, and same analysis procedure was applied for this new batch of qPCR data. The Calculation of relative expression can be found in the method part, and the raw Cp values are available in the Table S1.

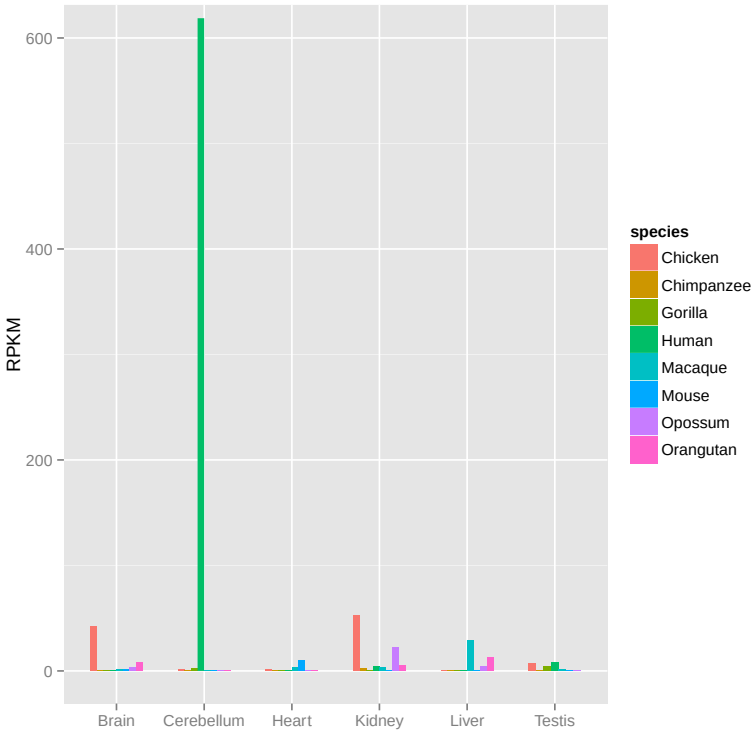

Figure 11. ZP2 expression in various tissues of multiple species. This barplot shows the ZP2 expression level in different tissues of various species. The Reads Per Million Per Kilobase (RPKM) is the normalized reads counts based on the public data provided by Brawand et al, where the raw reads per base reads numbers provided by this study were normalized by total number of raw reads and scaled to 1kb. For the samples with replicates, the mean RPKM was used.

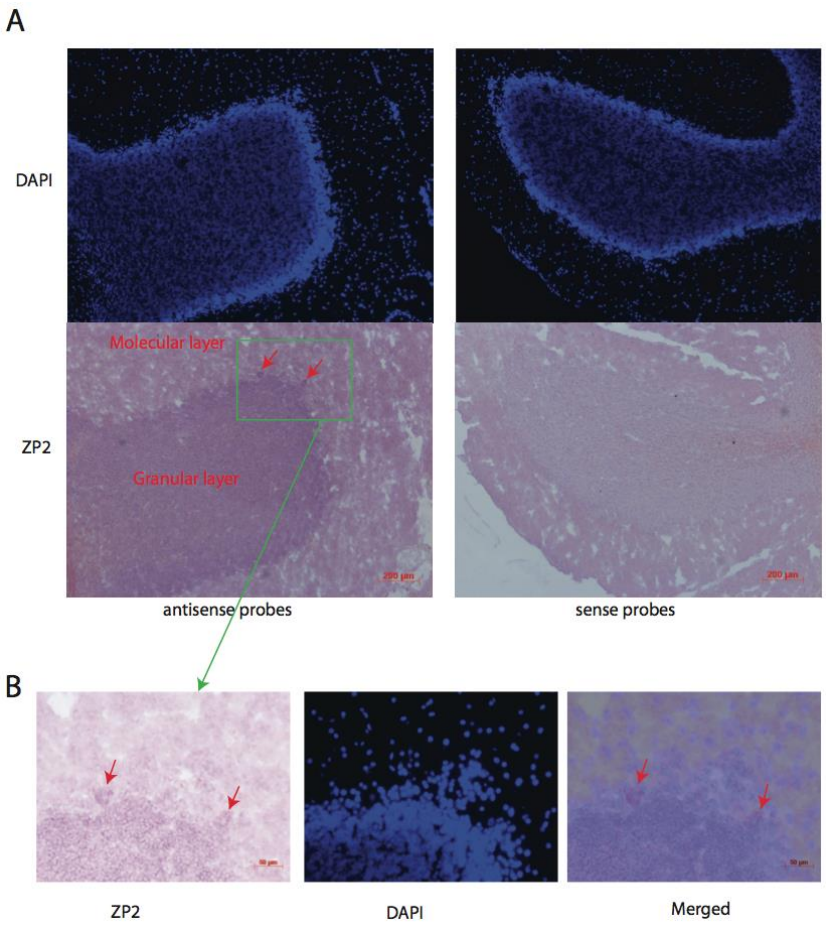

Figure 12. ZP2 RNA in situ hybridization (ISH) in human cerebellar cortex. (A) ISH comparison revealed expression of ZP2 in granular layer and purkinje cells. Sense and antisense probes were designed according to ZP2 messenger RNA as described in the method part. (B) ISH results show that ZP2 is expressed in the purkinje cytoplasm.

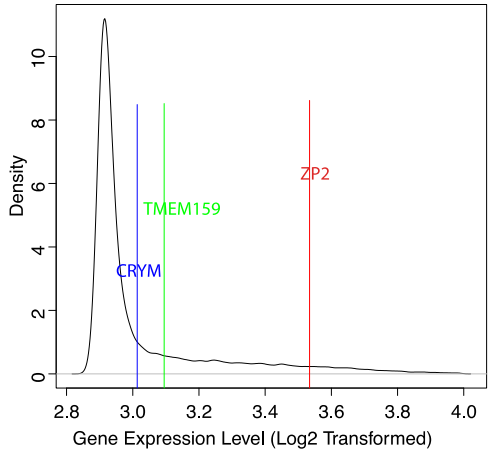

Figure 13. Zp2, CRYM and TMEM159 are highly expressed in human Purkinje cells. Expression data from GSE37205 was further quantile normalized, and the mean expression level from different species was used in this plot. The expression levels of the three genes are indicated by the vertical lines.

Human specific co-regulation of TMEM159-ZP2-CRYM

Gene expression usually exhibits special spatiotemporal patterns, so it's important to find out human specific time-related gene expression changes for the study of human specificity. As already shown, ZP2 is highly expressed in adult human cerebellum but not in the corresponding tissues of other species. Taking the advantage of age-series gene expression profiles done in our group for the humans, chimpanzees and rhesus macaques, we checked the expression of ZP2 during the whole post-natal life span in the three species. The result indicates that ZP2 is highly up regulated at the early human cerebellum post-natal developmental stage and keeps such high expression level at the later stages with slight drop during aging process (Figure 14A).

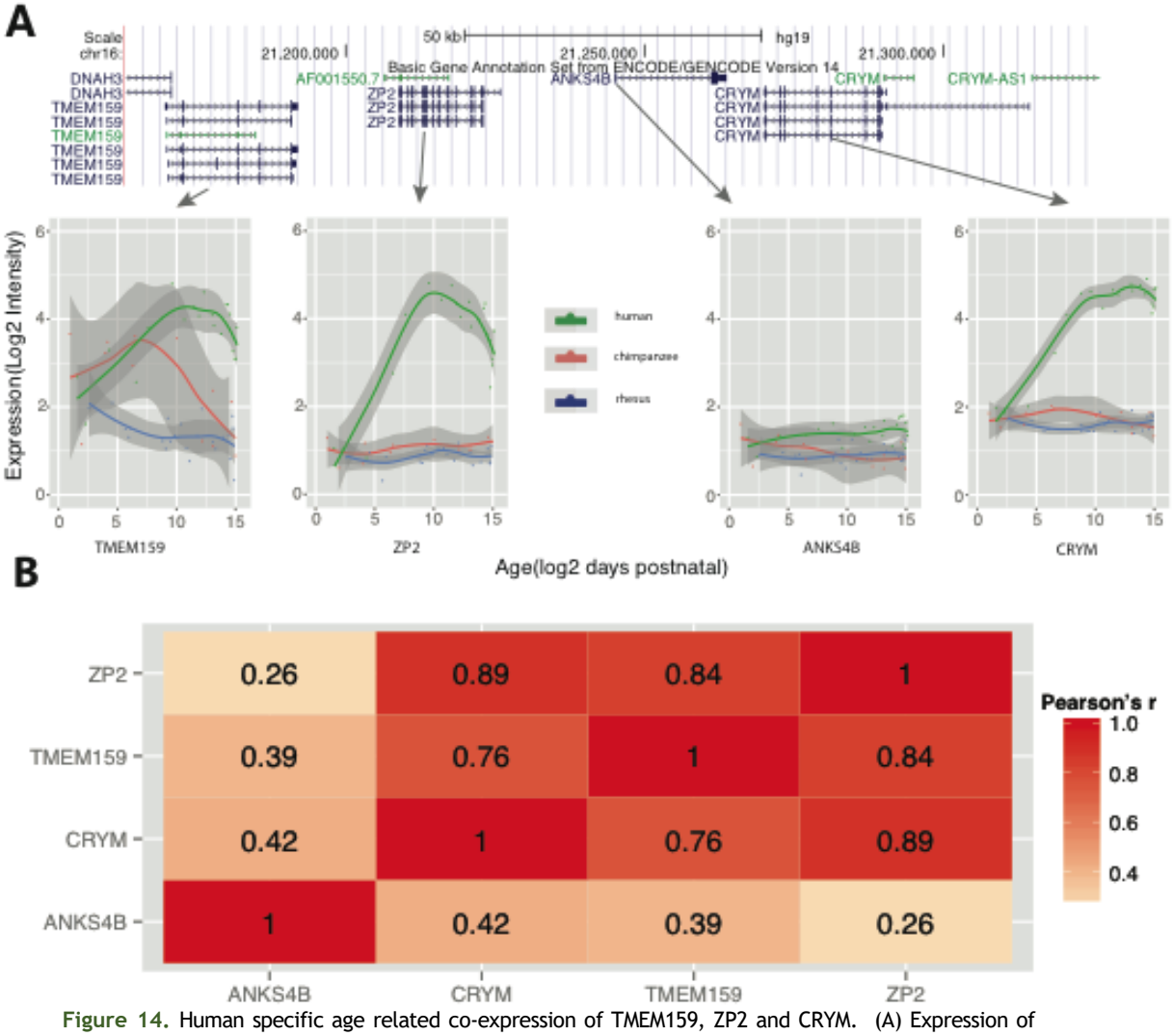

Figure 14. Human specific age related co-expression of TMEM159, ZP2 and CRYM. (A) Expression of ZP2 and its neighbor genes in human, chimpanzee and rhesus cerebellum during early post-natal development and aging process. Loess regression was used for smoothing and the shaded region surrounding the smoothed line represents 0.95 confidence interval of the expression level. The expression data is based on Liu et al's study. (B) Pairwise gene expression correlation of ZP2 and its neighbor genes. The same data was used here as the one used in A.

Interestingly, we also found that the neighbor genes of ZP2 (TMEM159 and CRYM) exhibits similar human specific age-related expression pattern (Figure 14A). According to our literature search, there is no known interaction or shared function of these three genes. Both of ZP2 and TMEM159 had been proposed contain transmembrane domains, but their function in human cerebellum is unknown. The age-related co-up regulation of these three genes are significantly correlated with each other, as shown in figure 14B. However, another ZP2 neighbor gene, which is located within TMEM159-ZP2-CRYM genomic area, doesn't show strong correlation with these three genes (Figure 14B) and is lowly expressed at all three species at the whole post-natal life span (Figure 14A).

The human specific gene expression change, especially the changes happen at large genomic area, could also be related with human disease. We found that the down-regulation of these three genes in human cerebellum is significantly related to medulloblastomas, and both of the down-regulations of ZP2 and CRYM are significantly associated with huntington's disease (Table 1). The correlation of the down-regulation of these three genes and human disease suggests they might play a possible functional role in human cerebellum, and further experimental evaluation of this hypothesis is needed.

Conclusion

Finding out human-specific genetic change is one important step to understand human evolution. The previous studies in searching of genes under positive selection based on genetic context had generated a list of human-specific changed genes at DNA level, which provided basis for further functional studies in model organisms. However, there is still lack of strong gene candidates that are possibly under positive selection based on human specific expression change. Here, we show such kind of strong candidate gene ZP2 that exhibits high level of human cerebellum specificity. The human specific expression of ZP2 led us to look at this genomic region in more detail, which revealed that the upstream and downstream of ZP2 genes are also human specifically highly expressed. The co-up regulation of ZP2 region could be simply due to human specific opening of the chromatin at this region or transcriptional read-through, however, we also found that ANKS4B that locates at this region is not specifically highly up regulated in human cerebellum.

ZP2 is conserved among mammals, shows about 60% identity between human and mouse in protein level (DATASTATISTICSBROWSER: HUMAN, 2016), and the promoter sequence of human ZP2 can drive the expression of reporter gene in mouse. During sperm-egg interaction in mouse, ZP2 is modified following fertilization, which contributes to the part of the changes in certain properties of zonae pellucida . To our knowledge, there is no single study that has illustrated the function of ZP2 in brain region of any organism. By checking ZP2 expression data in different human brain regions, we found that ZP2 is only highly expressed in cerebellum. Our study highlights the importance of study of ZP2 in the future in order to better understanding of human specific phenotypes, especially ZP2 upstream and downstream genes are also human specifically co-up regulated.

It has long been observed that human and close primates walk in different ways, we walk up right but our closest relative chimpanzees have knuckle walking. Such kind of gesture changes might require human specific movement coordination change, which may involve the change of human cerebellum. In our disease analysis, we noticed that the expression of ZP2 and CRYM is significantly down regulated in Huntington's disease, the patients of which have muscle coordination problems. However, more sophisticated experimental deign for supporting this hypothesis is needed.

Furthermore, besides this strong age-related regulation change in human cerebellum, we also observed the divergence to diversity ratio difference between protein coding genes and pseudogenes at human brain regions but not in non-brain tissues, and such observation is contradictory to the previous claims. However, in the previous claims, they only considered very limited number of pseudogenes in the comparison but in this study thousand of pseudogenes show expression signals.

Here, we also attached the expected ZP2 probe sequences and the ISH ZP2-fragments containing vector Sanger sequencing product.

Expected ZP2 probes sequences:

>ZP2 sense probe

CAACCTTATGGGGAAAACGAGTACCCTCTAGTGAGATTCCTCCGCCAACCAATTTACATG
GAAGTGAGAGTCCTAAACAGGGATGACCCCAACATCAAGCTGGTCTTAGATGACTGCTGGGCG
ACGTCCACCATGGATCCAGACTCTTTCCCCCAGTGGAACGTTGTCGTGGATGGCTGTGCATAT
GACCTGGACAACTACCAGACCACCTTCCATCCAGTCGGCTCCTCTGTGACCCATCCTGATCA

>ZP2 antisense probe

TGATCAGGATGGGTCACAGAGGAGCCGACTGGATGGAAGGTGGTCTGGTAGTTGTCCAG
GTCATATGCACAGCCATCCACGACAACGTTCCACTGGGGGAAAGAGTCTGGATCCATGGTGGA
CGTCGCCCAGCAGTCATCTAAGACCAGCTTGATGTTGGGGTCATCCCTGTTTAGGACTCTCACT
TCCATGTAAATTGGTTGGCGGAGGAATCTCACTAGAGGGTACTCGTTTTCCCCATAAGGTTG

And the Sanger sequencing results containing part of the pGEM-5Zf(+) vector sequences:

GAGGCATTTTTTGGCGATTGGGCCGACGTCGCATGCTCCCGGCCGCCATGGCGGCCGCG
GGAATTCGATTTGATCAGGATGGGTCACAGAGGAGCCGACTGGATGGAAGGTGGTCTGGTAG
TTGTCCAGGTCATATGCACAGCCATCCACGACAACGTTCCACTGGGGGAAAGAGTCTGGATCC
ATGGTGGACGTCGCCCAGCAGTCATCTAAGACCAGCTTGATGTTGGGGTCATCCCTGTTTAGG
ACTCTCACTTCCATGTAAATTGGTTGGCGGAGGAATCTCACTAGAGGGTACTCGTTTTCCCCAT
AAGGTTGAATCACTAGTGAATTCGCGGCCGCCTGCAGGTCGACCATATGGGAGAGCTCCCAAC
GCGTTGGATGCATAGCTTGAGTATTCTATAGTGTCACCTAAATAGCTTGGCGTAATCATGGTCA
TAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCA
TAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACT
GCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGG
GAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGT
CGTTCGGCTGCGGCGAGCGGTATCAGCT

CACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATG
TGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCAT
AGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCG
ACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGA
CCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGCGCTTTCTCATAGC
TCACGCTGTAGTATCTCAGTTCGGTGTAGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACC
CCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCACCCGGTAAGAA
CACGACTTATCGCCACTGGCA

Acknowledgement

Thanks. This study was supported by The Federal Targeted Programme for Research and Development in Priority Areas of Advancement of the Russian Scientific and Technological Complex for 2014-2020 (the Ministry of Education and Science of the Russian Federation), grant № 14.615.21.0002, the Unique identifier of the agreement: RFMEFI61515X0002

Disclosure statement

No potential conflict of interest was reported by the authors .

Notes on contributors

Ganjcai Xie, PhD, specialist at the Shanghai Institute for Biological Sciences, CAS, Shanghai, China.

Xi Jiang, specialist at the Shanghai Institute for Biological Sciences, CAS, Shanghai, China.

Ning Fu, PhD, specialist at the Skolkovo Institute for Science and Technology, Skolkovo, Russia.

Philip Khaytovich, PhD, specialist at the Shanghai Institute for Biological Sciences, CAS, Shanghai, China; Skolkovo Institute for Science and Technology, Skolkovo, Russia; The Immanuel Kant Baltic Federal University, Kaliningrad, Russia.

References

BLAST. (2016). NCBI National Center for Biotechnology Information. http://blast.ncbi.nlm.nih.gov/Blast.cgi.

BLAT. (2016). UCSC Genomics Institute. http://genome.ucsc.edu.

Chen, J., Sun, M., Hurst, L. D., Carmichael, G. G., Rowley, J. D. (2005). Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet., *21*, 326-329.

Chu, J., Dolnick, B. J. (2002). Natural antisense (rTSalpha) RNA induces site-specific cleavage of thymidylate synthase mRNA. *Biochim. Biophys. Acta*, 1587, 183-193.

ClustalW2. (2016). EMBL-EBI. http://www.ebi.ac.uk/Tools/msa/clustalw2/.

Dan, I., Watanabe, N. M., Kajikawa, E., Ishida, T., Pandey, A., Kusumi, A. (2002). Overlapping of MINK and CHRNE gene loci in the course of mammalian evolution. *Nucleic Acids Res.*, *30*, 2906-2910.

DATASTATISTICSBROWSER: HUMAN. (2016). GENCODE. http://www.gencodegenes.org/.

Ensembl. (2016). EMBL-EBI. http://www.ensembl.org/.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res., 33*, 121-124.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M. (2005). Antisense transcription in the mammalian transcriptome. *Science*, *309,* 1564-1566.

Liftover. (2016). UCSC Genome Bioinformatics. http://genome.ucsc.edu/util.html.

Liu, C., Bai, B., Skogerbo, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., Chen, R. (2005). NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, *33*, 112-115.

Markham, N.R., Zuker, M. (2005). DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res., 33*, 577-581.

Oeder, S., Mages, J., Flicek, P., and Lang, R. (2007). Uncovering information on expression of natural antisense transcripts in Affymetrix MOE430 datasets. *BMC Genomics*, *8*, 200-203.

Pelletier, J., Sonenberg, N. (1985). Insertion mutagenesis to increase secondary structure within the 5' noncoding region of a eukaryotic mRNA reduces translational efficiency. Cell, *40* (3), 515-526.

Prescott, E. M., Proudfoot, N. J. (2002). Transcriptional collision between convergent genes in budding yeast. *Proc. Natl. Acad. Sci. USA., 99*, 8796-8801.

Q05996 (ZP2_HUMAN). (2016). UniProt. http://www.uniprot.org/uniprot/Q05996.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language. Cell, *146*, 353-358.

Smalheiser, N. R. (2012). The search for endogenous siRNAs in the mammalian brain. Exp. Neurol., *235*, 455-463.

Spielmann, M., Brancati, F., Krawitz, P. M., Robinson, P. N., Ibrahim, D. M., Franke, M., (2012). Homeotic arm-to-leg transformation associated with genomic rearrangements at the PITX1 locus. Am. J. Hum. Genet., *91*, 629-635.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26 (1), 148-153.

Storz, P., Doppler, H., Toker, A. (2005). Protein kinase D mediates mitochondrion-to-nucleus signaling and detoxification from mitochondrial reactive oxygen species. *Mol. Cell. Biol.*, *25,* 8520-8530.

Wang, W., Zheng, H., Yang, S., Yu, H., Li, J., Jiang, H. (2005). Origin and evolution of new exons in rodents. *Genome Res.*, *15*, 1258-1264.

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A. (2003). Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, *21*, 379-386.

Zhu, Z., Zhang, Y., Long, M. (2009). Extensive structural renovation of retrogenes in the evolution of the Populus genome. *Plant Physiol.*, *151,* 1943-1951.