# Supplementary Materials

Dual Eigen-modules of Cis-element Regulation Profiles and Selection of Cognition-language Eigen-direction along Evolution in Hominidae

Liang Li, Sheng Zhang, Lei M. Li

## Content

# Top eigen-components of cis-element frequency matrices

SVD is a powerful dimension-reduction method that enables us to extract useful information from fairly large matrices. The top eigenvalue and eigenvector of each species-specific SVD are essentially the adjusted averages, and these baselines provide little interesting biological explanations to the binding profiles. Hence we exclude the baseline component when we refer to the first and second eigenvalues or eigenvectors throughout the article. For all three species, the top six eigen-components contribute about 42 percent of the information to the cis-element regulation profiles. Table S1 shows the top six singular values of the three cis-element frequency matrices and the percentages of their cumulative contributions. The distance between two adjacent singular values is measured by the ratio of their difference over the larger one.

**Table S1**. **The top six singular values of the cis-element frequency matrices of human, chimpanzee and orangutan, and their percentages.**

| | **Human** | | | **Chimpanzee** | | | **Orangutan** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Value (x10^3) | Distance (%) | Cumulative percentage(%) | Value (x10^3) | Distance (%) | Cumulative percentage (%) | Value (x10^3) | Distance (%) | Cumulative percentage (%) |
| ρ1 | 77.92 | | 24.77 | 69.92 | | 25.19 | 69.59 | | 25.30 |
| ρ2 | 19.43 | 75.058 | 30.95 | 16.89 | 75.85 | 31.28 | 16.37 | 76.47 | 31.26 |
| ρ3 | 12.35 | 36.461 | 34.88 | 10.77 | 36.25 | 35.15 | 10.59 | 35.30 | 35.11 |
| ρ4 | 8.11 | 34.335 | 37.46 | 7.07 | 34.31 | 37.70 | 7.03 | 33.63 | 37.67 |
| ρ5 | 7.86 | 3.041 | 39.96 | 6.79 | 4.00 | 40.15 | 6.45 | 8.23 | 40.02 |
| ρ6 | 5.92 | 24.662 | 41.84 | 5.11 | 24.70 | 41.99 | 5.13 | 20.53 | 41.88 |

# The probabilistic model for quantification of cis-trans binding strength

A key problem in the study of global cis-regulation is how to quantitatively measure the genome-wide binding strength or probability. Our study design is partially motivated by the probabilistic model proposed by (Feng, et al. 2019). The model shows that motif frequency is a fair measure of binding strength. Specifically, given a factor $T$ and its motif position weight matrix, all potential binding sites in the proximal regulatory sequence $S$ of a gene can be identified by searching the high-scoring matches to its motif. Suppose $N$ binding sites are found in $S$. The model assumes that the binding events of $T$ with these $N$ sites independent, and identical distributed Bernoulli trials with a binding probability $p_0$. Then the probability that $T$ binds to at least one site is given by $1-(1-p_0)^N$. If $p_0$ is sufficiently small, it can be approximated by

$$1-(1-p_0)^N \approx Np_0.$$

The motif occurrences in the promoter of a certain gene are positively related to the binding strength of its trans-acting factor.

# SVD and dual eigen-analysis

Crucial to the understanding of the impact of regulatory sequences on the phenotypic changes from apes to humans is a good mathematical representation of the cis-element frequency profile $\tilde{C}$. The dual eigen-analysis (Li, et al. 2017) that we originally proposed for the analysis of high-dimensional expression profiles is

turned out to be ideal for unravelling the structure of the regulatory profile. The initial computation of the dual eigen-analysis is a robust version of the singular value decomposition (SVD) (Golub and Loan 1996; Lin, et al. 2010; Cand, et al. 2011).

$$\min_{C,S} \ \|C\|_* + \lambda\|S\|_1 \quad subject\ to\ \tilde{C} = C + S,$$

$$C = \sum_{k=1}^{m} \rho_k u_k v_k^T$$

Importantly, this eigen-representation stratifies the cis-element frequencies into dual eigen-spaces at levels from high to low. Instead of aligning protein sequences, the learning of regulatory evolution is achieved by aligning the dual eigen-spaces of cis-element frequencies across species. We note that dual eigen-analysis, though based on SVD, is different from the classical multivariate analysis tools such as PCA or factor analysis, which are based on covariance structures of joint normal distributions.

Dual eigen-analysis is used to interpret the eigenvectors from SVD of relatively large matrices. It has several elements. First, $u_k$ pairs up with $v_k$ as SVD indicates. Second, sort the loadings of $u_k$ in the descending order, and denote the resulting polarized gene eigenvector by $\tilde{u}_k$. Similarly, we have the polarized motif eigenvector denoted by $\tilde{v}_k$. The two ends of $\tilde{u}_k$ respectively correspond to the two ends of $\tilde{v}_k$. Third, each pair $(u_k, v_k)$ represent an association structure of the row attributes and the column attributes. The structure is identified by associating $\tilde{u}_k$ with the row attributes, and by associating $\tilde{v}_k$ with the column attributes. For example, in the Type 2 diabetes project (Li, et al. 2017), we consider the mouse expression profiles whose rows corresponded to samples and columns corresponded to genes. For each principal eigen-component, $\tilde{u}_k$ was translated into experimental factors such as diet or age while $\tilde{v}_k$ reflected the underlying molecular mechanisms that can be identified by gene set enrichment analysis. Consequently, the relationship between the macro-factors of samples and micro-biology of genes was revealed by the dual-eigen-analysis.

In the current situation of a cis-element frequency matrix $C$, the rows represent genes and the columns represent the binding elements of regulators. Each eigen-triplet, $\langle\rho_k|\tilde{u}_k|\tilde{v}_k\rangle$, forms a dual eigen-module of cis-regulatory element frequency (CREF) $C_k$, in which $\rho_k$ indicates its frequency level. The two ends of the polarized regulatory eigenvector $\tilde{v}_k$ correspond to the two ends of the polarized gene eigenvector $\tilde{u}_k$, whose biological meanings can be inferred by the Wilcoxon scoring method. We note the SVD is not unique in the sense that we can swap the signs of $u_k$ and $v_k$ simultaneously without changing the product $u_k v_k^T$. In our analysis, once the sign of the ape motif eigenvector $v_k$ is set to be positively correlated with that of the human's , the sign of the ape gene eigenvector $u_k$ is determined accordingly.

## Comparisons of motif eigenvectors among hominidae

In Fig. 2A, we show the scatter plots of human's and chimpanzee's motif eigenvectors. Here, we further show their scatter plots with respect to orangutan's motif eigenvectors. In the case of humans versus orangutans, a similar pattern is observed as in the case of humans versus chimpanzees (Fig. S1A). The top three pairs and the sixth pair are highly consistent whereas the fourth and fifth pairs are less correlated. In the case of chimpanzees versus orangutans, which serves a good reference for the comparison of human versus apes, all top six pairs are highly correlated (Fig. S1B).

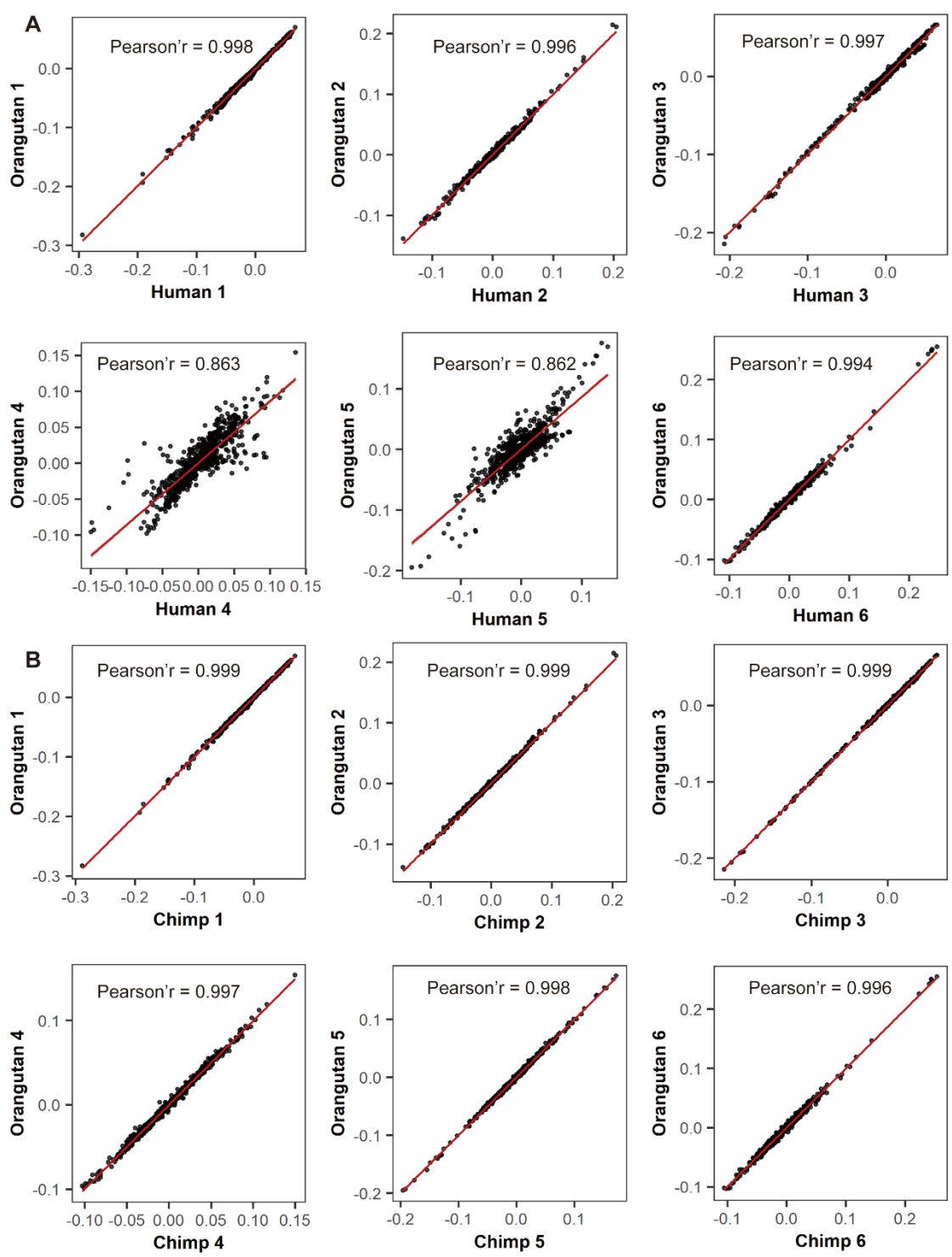
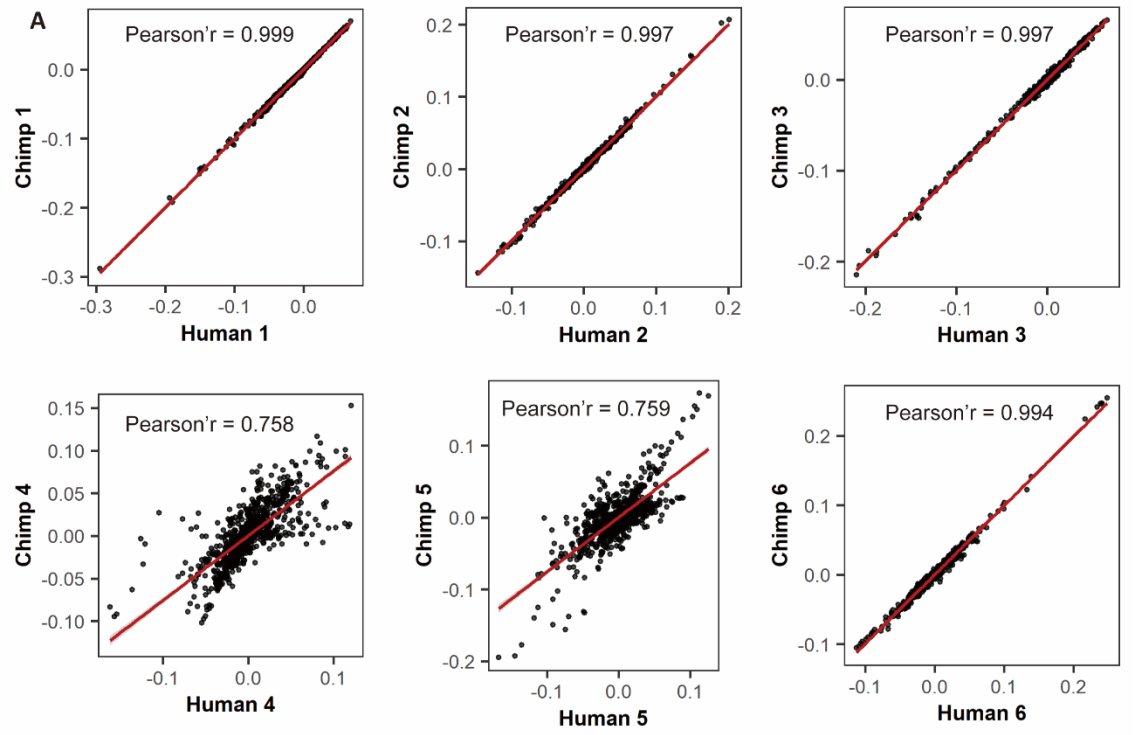



**Fig. S1**. **(A) The scatter plots of human's top six motif eigenvector loadings versus orangutan's.** Like the case of human's versus chimpanzee's, the top three pairs and the sixth pair are highly correlated (Pearson correlation > 0.99). The fourth and fifth pairs are less correlated. **(B) The scatter plots of chimpanzee's top six motif eigenvector loadings versus orangutan's.** All the top six pairs are highly correlated (Pearson correlation > 0.99).

## Dual eigen-analysis on cis-element frequency profiles in the proximal regulatory regions of orthologous genes

To remove the effects of gene annotation variations, we focus exclusively on motif occurrences in the proximal regulatory regions of orthologous genes across the three species. A set of 11454 1:1:1 orthologous genes among human, chimpanzee and orangutan are selected directly from Ensembl database. We perform the dual eigen-analysis based on the matrices containing motif frequencies in their proximal regulatory regions. The scatter plots of motif eigenvectors between two species are shown in Fig. S2. As expected, the top three and the sixth human-apes motif eigenvectors are highly conserved (Pearson correlation > 0.99) whereas the human fourth and fifth motif eigenvectors rotate from those of apes'.



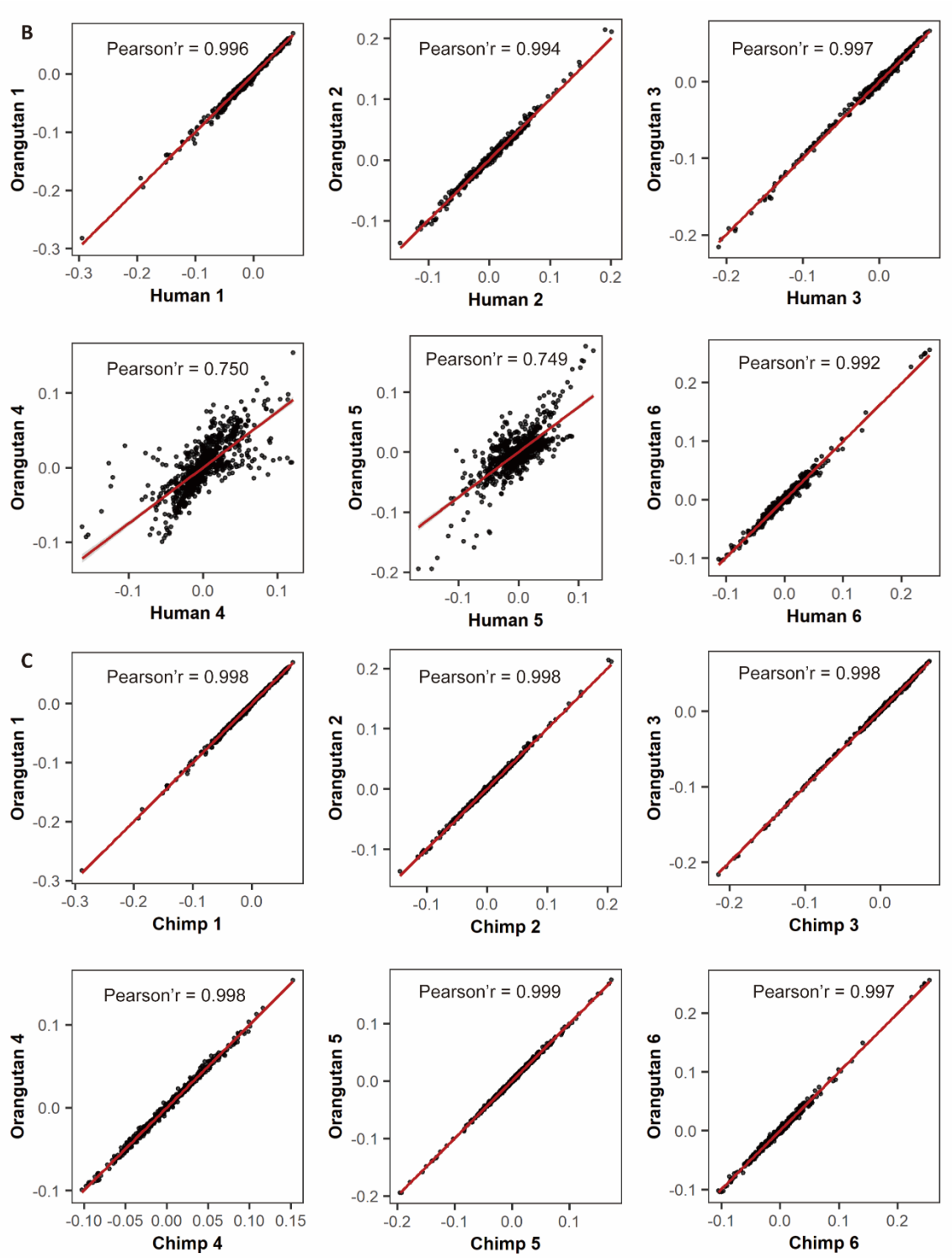



**Fig. S2**. **The scatter plots of motif eigenvectors based on motif frequencies in the orthologous promoters.** The results support that the conservation and reorganization of the CREF modules is robust with respect to the effects of gene repertoire variations (the second last paragraph in Discussion). **(A)** Human's top six motif eigenvector loadings versus chimpanzee's. **(B)** Human's top six motif eigenvector loadings versus orangutan's. **(C)** Chimpanzee's top six motif eigenvector loadings versus orangutan's.

## Stability analysis of the top singular values by random sampling

The perturbation theory demonstrates that the stability of CREF eigen-modules hinges on the pairwise distances between adjacent singular values. We need to evaluate the reliability of the distances. To address this issue, we introduce random perturbations by re-sampling the motifs, and examine the sampling distribution. We note that the bootstrap method, a sampling with replacement, generate a fairly large portion of duplicate motifs. Instead, we adopt subsampling without replacement. The re-sampling process consists of three steps: (i) randomly select 80% motifs from 1403 motifs without replacement; the corresponding columns of those motifs within $C$ are extracted to form a perturbed matrix denoted by $\hat{C}^i$; (ii) calculate the SVD of $\hat{C}^i$ and collect its singular values $\hat{\rho}_k^i$; (iii) repeat the first two steps 100 times. The sampling distribution of each singular value was obtained based on $\{\hat{\rho}_k^i\}_{i=1}^{100}$. Similarly, the sampling distribution of the distance between the fourth and fifth levels was obtained based on $\{(\hat{\rho}_4^i - \hat{\rho}_5^i)/\hat{\rho}_4^i\}_{i=1}^{100}$.

The sampling distributions of chimpanzee's top six singular values are more similar to those of human's than to those of orangutan's (Fig. S3A-B and Fig. 2C). That is, in the case of chimpanzees, the sampling distributions of the top three and the sixth singular values are well separated from adjacent ones while the fourth and fifth overlap by a portion. With a close look at the distributions of orangutans, we find a subtle difference from those of the other two species, see Fig. S3B. That is, the 95 percent confidence interval of the 5-th singular value does not overlap with that of the 4-th value. This is in consistency with the larger distance between $\rho_4$ and $\rho_5$ of orangutan's cis-regulation profile, 8.2%, compared with those of humans and chimpanzees, which are 3.0 % and 4.0 % respectively (Table S1).

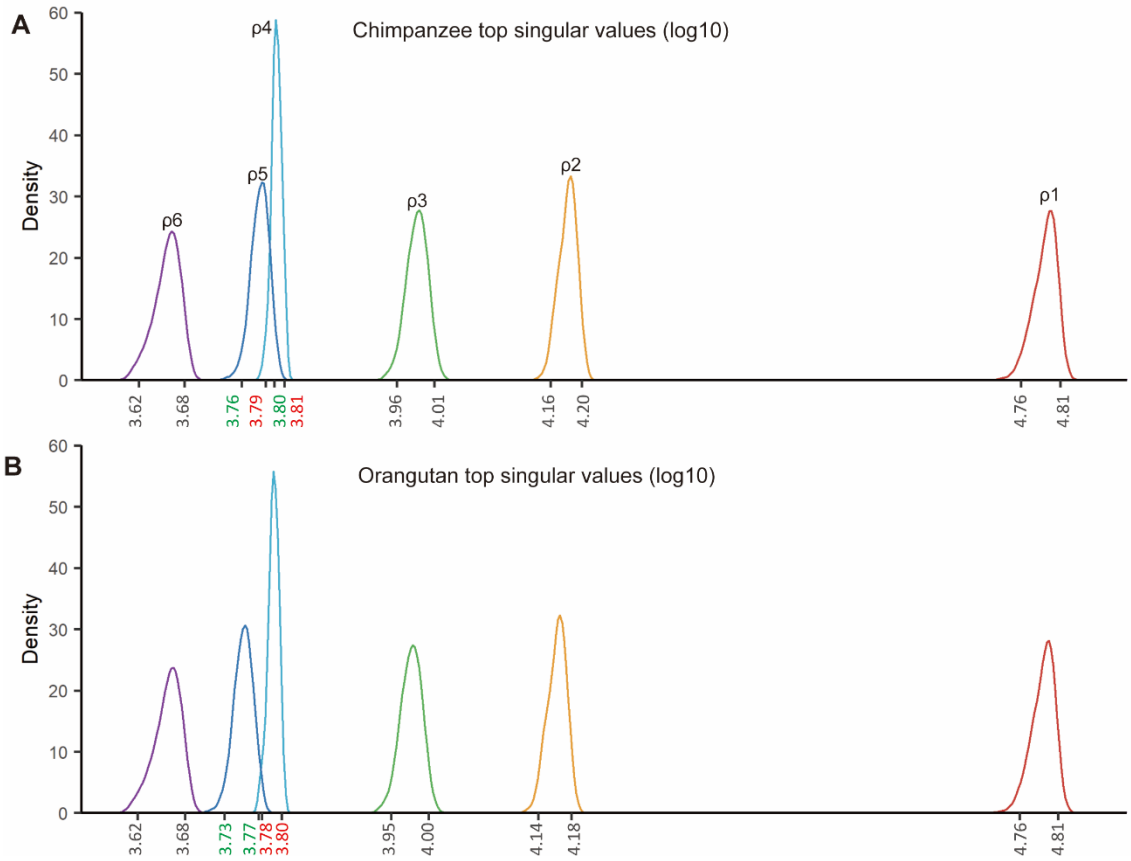


**Fig. S3**. **The sampling distributions of chimpanzee's and orangutan's top six singular values.** Their densities are shown from right to left. The 2.5% and 97.5% sample quantiles are labelled along the x-axis. Particularly, the sample quantiles of the fourth and fifth singular values are marked by red and green color respectively. (**A**) In the case of chimpanzees, the sampling distributions of the top three and the sixth singular values are well separated from adjacent ones while the fourth and fifth overlap by a large portion. **(B)** In the case of orangutans, the 95 percent confidence intervals of the fourth and fifth singular values do not overlap.

# Comparisons of the distance between the fourth and fifth levels among species by sampling and tests

We would test if the distance between the fourth and fifth levels increases as the divergence time increases. First, based on the SVD results of each resampled sub-matrix as described in the above section, the matched distance of each species is obtained. In total, we have 100 subsamples. Second, a paired two-sample t-test (one-side) is used to compare the sampling distributions of two species, e.g., human versus chimpanzee. The sample distributions and statistical significance are shown in Fig. S4A. We find that human's distance is significantly smaller than chimpanzee's (p-value = 3.96e-7), and, that chimpanzee's is significantly smaller than orangutan's (p-value = 1.23e-78). These results confirm that as the divergence time between species increases, the distance between the 4-th and 5-th levels increases as well.

We further compare the distance between level 4 and 5 of each species in reference to that generated at its fusion point, when the fourth and fifth levels are equal. First, we need to generate the distribution of the distance at the fusion point. We generate a new cis-element frequency matrix, denoted by $C^*$, by forcing the fourth and fifth levels equal in each species to mimic the fusion state,

$$C^* = \sum_{k=1}^{m} \rho_k^* u_k v_k^T,$$

$$\text{where } \rho_k^* = \begin{cases} \rho_k, & k \neq 4,5 \\ (\rho_4 + \rho_5)/2, & k = 4,5 \end{cases}$$

Second, The re-sampling method described above was then carried out for $C^*$ to obtain the perturbed distances.

Third, we compared them against the distances generated from the original matrices by matching the motif subsamples. Among the 100 random simulations, the former distances are smaller than the latter in 77, 85 and 97 cases respectively in human, chimpanzee and orangutan. We further applied a two-sample paired t-test (one-side) to test the significance of deviance. The two kinds of distributions and their statistical significance for comparison are shown in Fig. S4B for each species. The discrepancy increases in the order of human (p-value = 7.17e-12), chimpanzee (p-value = 1.73e-17) and orangutan (p-value = 3.83e-44). It confirms that human is the closest to the fusion point.

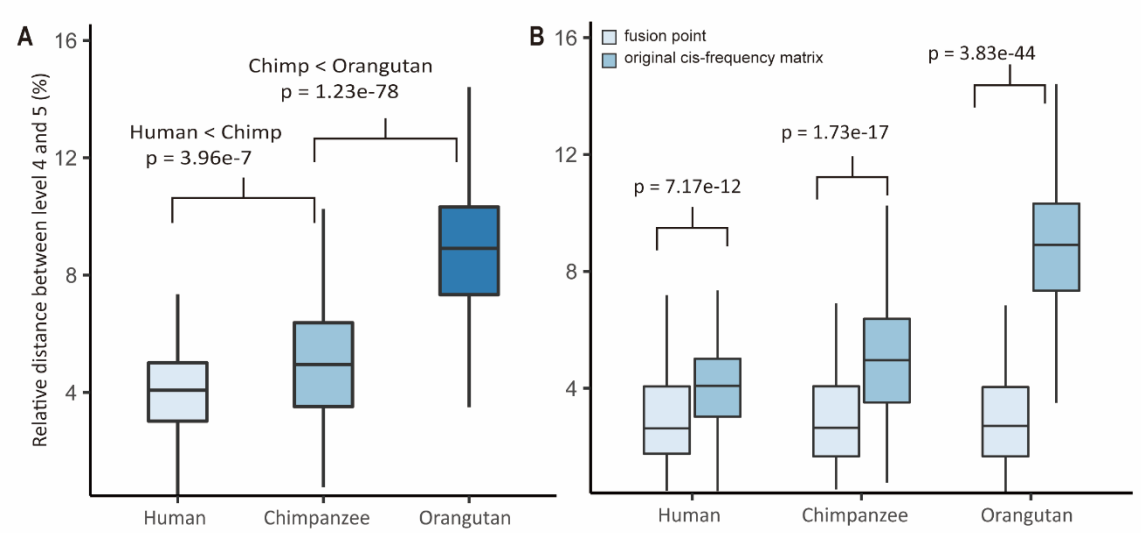



**Fig. S4 (A) The sampling distributions of the distance between level 4 and 5**. Comparison of the distributions from two species is performed using a paired two-sample t-test (one-side). The statistical significances are labelled on the top of each box. **(B) The sampling distributions at the fusion point versus those in (A)**. For each species, the two distributions are compared using a paired two-sample t-test (one-side).

# Enrichment analysis by the Wilcoxon rank sum scoring method

In the dual eigen-analysis, one key step of analyzing CREF modules is to infer biological activities around the poles of the polarized gene eigenvectors using enrichment analysis, which requires the definition of relevant gene subsets. In this report, we consider the following gene subsets: KEGG pathways (Kanehisa and Goto 2000; Kanehisa, et al. 2016; Kanehisa, et al. 2017), Gene Ontology gene sets (Ashburner, et al. 2000; Carbon, et al. 2017), which are classified into biological processes, molecular functions, and cellular components, and the REACTOMEs database (Fabregat, et al. 2018). In the dual eigen-analysis, the enrichment analysis is based on the gene loadings. To address the robustness with respect to the scale, we take the rank-based Wilcoxon scoring method as the tool for enrichment analysis. The method was proposed in Cheng et al (Cheng, et al. 2007). For each polarized gene vector, we identify what subsets are significantly enriched at the two ends. Suppose we have $n$ gene subsets $S_1, S_2, \dots, S_n$. Denote the union of these gene subsets by $G = \cup_{i=1}^{n} S_i$. The strategy consists of two steps. In the first step, for each subset $S_i$, we compare their loadings against those in the complement of $S_i$ in $G$ denoted by $G - S_i$. This is a typical two-sample problem in statistics. We use the Wilcoxon rank test to calculate p-value (one-sided) for each comparison. In the second step, we rank these subsets according to their significances.

To characterize the biological relevance of each eigen-module simply and clearly, we organized the significant subsets according to the three rules: 1. we prioritized gene categories that were most significant at each level; 2. if a gene subset was enriched at multiple levels, only the highest level was marked unless it was extremely significant at the lower level; 3. among candidate pathways at a certain level, we emphasized those gene subsets whose biological relevance matched the importance.

**Table S2** The portions of significant gene subsets shared by both human and chimpanzee at each level.

| | **GO biological process** | **KEGG** |
|---|---|---|
| **Level 1** | 0.85 | 0.84 |
| **Level 2** | 0.85 | 0.82 |

| | | |
|---|---|---|
| **Level 3** | 0.70 | 0.83 |
| **Level 4** | 0.39 | 0.46 |
| **Level 5** | 0.81 | 0.90 |
| **Level 6** | 0.78 | 0.75 |

# Details of the enrichment analysis of top gene eigenvectors

## The human/chimpanzee first gene eigenvector

### Positive end: reproduction

The positive end of the human first gene eigenvector is marked by reproduction (Table S3). Two related biological processes, gametogenesis and fertilization, are enriched around the pole. Gene categories associated with gametogenesis include meiosis and oogenesis. Gene categories associated with fertilization include sperm preparation and sperm getting through the zona pellucida. In addition, two other groups, immune defense and rRNA silencing, appear with reproduction around this pole.

### Negative end: embryogenesis

The negative end of this vector is marked by a series of biological events during embryogenesis, including embryonic placenta development, mesoderm and endoderm formation, neural tube development, heart and lung morphogenesis (Table S4). They are known to occur at different stages of human early embryonic development. At the stage of gastrulation, the three germ layer, ectoderm, mesoderm and endoderm, are differentiated from the epiblast. They will finally give rise to all the structures and organs of the body. The neural tube develops from the ectoderm cells at the stage of neurulation, and ultimately transforms into the nervous system. The hearts derive from the layer of mesoderm, and is the first functional organ to develop during embryogenesis. The lungs arise from the layer of endoderm and will develop to maturity until early adulthood.

**Table S3**. In the sorted loadings of the human first gene eigenvector $u_1$, gene subsets enriched at the positive end.

| | Description | Source | P-value |
|---|---|---|---|
| Reproduction | single fertilization | GO BP | 6.79E-03 |
| | binding of sperm to zona pellucida | GO BP | 7.21E-03 |
| | plasminogen activation | GO BP | 8.72E-04 |
| | regulation of acrosome reaction | GO BP | 1.99E-04 |
| | oogenesis | GO BP | 9.75E-03 |
| | cell wall macromolecule catabolic process | GO BP | 3.37E-04 |
| | homologous recombination | KEGG | 2.47E-02 |
| | meiotic recombination | REACTOME | 1.03E-05 |
| | separation of sister chromatids | REACTOME | 3.64E-02 |
| | meiosis | REACTOME | 2.06E-03 |
| rRNA silencing | SIRT1 negatively regulates rRNA Expression | REACTOME | 1.15E-05 |
| | DNA methylation | REACTOME | 3.42E-04 |
| | nucleosome assembly | REACTOME | 3.01E-03 |

| | | | |
|---|---|---|---|
| | PRC2 methylates histones and DNA | REACTOME | 2.22E-03 |
| | chromosome maintenance | REACTOME | 1.61E-02 |
| Immune defense | T cell activation involved in immune response | GO BP | 1.23E-10 |
| | natural killer cell activation involved in immune response | GO BP | 4.82E-10 |
| | B cell proliferation | GO BP | 1.20E-08 |
| | humoral immune response | GO BP | 8.90E-06 |
| | positive regulation of innate immune response | GO BP | 1.54E-04 |
| | monocyte chemotaxis | GO BP | 2.60E-04 |
| | regulation of type I interferon-mediated signaling pathway | GO BP | 1.40E-05 |
| | negative regulation of interferon-gamma production | GO BP | 5.34E-05 |
| | cellular response to interferon-gamma | GO BP | 2.48E-04 |
| | complement receptor mediated signaling pathway | GO BP | 1.10E-03 |
| | cellular defense response | GO BP | 9.38E-04 |
| | toll-like receptor signaling pathway | GO BP | 2.55E-03 |

**Table S4**. In the sorted loadings of the human first gene eigenvector $u_1$, gene subsets enriched at the negative end.

| | Description | Source | P-value |
|---|---|---|---|
| Implantation | embryonic placenta development | GO BP | 7.80E-03 |
| Gastrulation | mesoderm formation | GO BP | 2.32E-05 |
| | endoderm formation | GO BP | 1.77E-03 |
| Neurulation | neural crest cell migration | GO BP | 2.50E-02 |
| | planar cell polarity pathway involved in neural tube closure | GO BP | 2.10E-02 |
| Organogenesis | lung morphogenesis | GO BP | 2.13E-02 |
| | heart morphogenesis | GO BP | 1.67E-02 |
| | endocardial cushion morphogenesis | GO BP | 3.59E-05 |
| | embryonic digit morphogenesis | GO BP | 2.18E-03 |
| | axon extension | GO BP | 3.62E-02 |
| | branching involved in ureteric bud morphogenesis | GO BP | 4.70E-03 |
| Signaling pathway | Wnt signaling pathway | KEGG | 1.26E-08 |
| | Hedgehog signaling pathway | KEGG | 1.24E-05 |
| | Signaling by FGFR | REACTOME | 1.37E-06 |
| | positive regulation of Notch signaling pathway | GO BP | 5.41E-04 |
| Others | developmental biology | REACTOME | 6.87E-15 |
| | vasculogenesis | GO BP | 2.03E-02 |
| | cell fate commitment | GO BP | 2.20E-03 |

## The human/chimpanzee second gene eigenvector

### Positive end: fetal maturation

The positive end of the second gene eigenvector is marked by fetal maturation, characterized by a series of organ and tissue development (Table S5). This process follows early embryogenesis, which lies at one end of the first gene eigenvector. Around this pole, we observe the signals of development of heart, pancreas, thymus, cartilage, muscle, eye, digestive tract, and forebrain.

**Negative end: immune response**

The negative pole is marked by immune system (Table S6). Related significant categories include immune cell chemotaxis, complement system, humoral response, innate immune response, GPCR signaling. Although several immune pathways such as immune cell activation, are enriched at one end of the first eigenvector too, we observe stronger signals of cell chemotaxis at this eigenvector, such as monocyte chemotaxis, lymphocyte chemotaxis, macrophage chemotaxis, neutrophil chemotaxis and leukocyte chemotaxis. Signal pathways involved in immune cell chemotaxis are enriched too, such as
the KEGG pathway of cytokine-cytokine receptor interaction and the REACTOME pathway of chemokine receptors bind chemokines.

**Table S5**. In the sorted loadings of the human second gene eigenvector $u_2$, gene subsets enriched at the positive end.

| | Description | Source | P-value |
|---|---|---|---|
| Organ and tissue development | embryonic organ development | GO BP | 1.39E-03 |
| | digestive tract development | GO BP | 4.69E-03 |
| | cartilage development | GO BP | 3.97E-02 |
| | cardiac muscle tissue development | GO BP | 1.78E-03 |
| | forebrain development | GO BP | 7.33E-03 |
| | pancreas development | GO BP | 6.68E-04 |
| | thymus development | GO BP | 5.34E-03 |
| | eye development | GO BP | 1.93E-02 |
| | proximal/distal pattern formation | GO BP | 4.49E-03 |
| | embryo development | GO BP | 5.46E-03 |
| | lung alveolus development | GO BP | 1.30E-02 |
| Neural system | neuron development | GO BP | 1.29E-02 |
| | neuronal stem cell population maintenance | GO BP | 7.59E-05 |
| Others | cell fate specification | GO BP | 2.02E-03 |
| | cell fate commitment | GO BP | 3.07E-02 |

**Table S6**. In the sorted loadings of the human second gene eigenvector $u_2$, gene subsets enriched at the negative end.

| | Description | Source | P-value |
|---|---|---|---|
| Chemotaxis | monocyte chemotaxis | GO BP | 1.19E-07 |
| | lymphocyte chemotaxis | GO BP | 4.56E-08 |
| | macrophage chemotaxis | GO BP | 4.86E-04 |
| | positive regulation of neutrophil chemotaxis | GO BP | 6.74E-03 |
| | leukocyte chemotaxis | GO BP | 3.92E-03 |
| | cytokine-cytokine receptor interaction | KEGG | 1.08E-05 |
| | chemokine receptors bind chemokines | REACTOME | 5.44E-07 |
| Innate immune | antimicrobial humoral response | GO BP | 1.30E-11 |
| | cellular response to interferon-gamma | GO BP | 6.11E-10 |
| | innate immune response in mucosa | GO BP | 7.00E-08 |

| | | | |
|---|---|---|---|
| | antibacterial humoral response | GO BP | 3.78E-06 |
| | acute-phase response | GO BP | 1.17E-04 |
| Complement system | complement and coagulation cascades | KEGG | 2.59E-05 |
| | initial triggering of complement | REACTOME | 3.68E-07 |
| | creation of C4 and C2 activators | REACTOME | 3.29E-06 |
| | lectin pathway of complement activation | REACTOME | 9.63E-05 |
| | complement cascade | REACTOME | 4.24E-03 |
| Homeostasis | retina homeostasis | GO BP | 2.71E-05 |
| | triglyceride homeostasis | GO BP | 2.33E-04 |
| GPCR signaling | GPCR ligand binding | REACTOME | 3.91E-09 |
| | G alpha (i) signalling events | REACTOME | 4.14E-03 |

## The human/chimpanzee third gene eigenvector

### Positive end: stress response

The positive pole of the third gene eigenvector reflects cellular responses to various stimuli including metal ion, hormone and other activities that result in changes in the state of a cell (Table S7). Cell responses include cell-cell adhesion, substance secretion (regulated exocytosis), ATP and protein metabolism, cytolysis, cellular homeostasis, etc.

### Negative end: mitosis

The negative pole of the third gene eigenvector is marked by cell-division cycle (Table S8). Almost all cell cycle phases and important cellular events during mitosis cluster around this end, such as G1/S transition, DNA replication at S phase, G2/M DNA replication checkpoint, condensation of prophase chromosomes and separation of sister chromatids at M phase, and M/G1 transition.

**Table S7**. In the sorted loadings of the human third gene eigenvector $u_3$, gene subsets enriched at the positive end.

| | Description | Source | P-value |
|---|---|---|---|
| Stress response | detection of chemical stimulus involved in sensory perception of bitter taste | GO BP | 1.85E-06 |
| | response to peptide | GO BP | 1.11E-03 |
| | response to activity | GO BP | 2.86E-02 |
| | cellular response to copper ion | GO BP | 2.56E-02 |
| | regulation of MAPK cascade | GO BP | 6.60E-03 |
| | cellular response to hormone stimulus | GO BP | 4.24E-02 |
| | stimuli-sensing channels | REACTOME | 1.88E-03 |
| | Signal regulatory protein (SIRP) family interactions | REACTOME | 9.98E-03 |
| | regulation of vesicle fusion | GO BP | 1.43E-04 |
| | cell-cell junction assembly | GO BP | 3.60E-02 |
| | positive regulation of cell adhesion | GO BP | 2.16E-02 |
| | cilium movement involved in cell motility | GO BP | 3.51E-03 |
| | cilium movement | GO BP | 5.01E-03 |
| | activation of JUN kinase activity | GO BP | 8.47E-03 |

| | | | |
|---|---|---|---|
| | actin filament-based movement | GO BP | 9.25E-03 |
| | ATP metabolic process | GO BP | 1.05E-02 |
| | sensory perception | GO BP | 2.97E-02 |
| | positive regulation of proteolysis | GO BP | 2.34E-03 |
| | ion channel transport | REACTOME | 1.90E-02 |

**Table S8**. In the sorted loadings of the human third gene eigenvector $u_3$, gene subsets enriched at the negative end.

| | Description | Source | P-value |
|---|---|---|---|
| Cell growth and preparation | positive regulation of mitotic cell cycle | GO BP | 5.11E-05 |
| | cell growth | GO BP | 5.38E-03 |
| | gene expression | REACTOME | 3.61E-12 |
| | transcription | REACTOME | 7.62E-04 |
| | translation | REACTOME | 2.47E-02 |
| | spliceosome | KEGG | 9.72E-04 |
| | synthesis of DNA | REACTOME | 4.41E-03 |
| | DNA replication | REACTOME | 2.10E-03 |
| | assembly of the pre-replicative complex | REACTOME | 3.57E-04 |
| | regulation of mitotic cell cycle | REACTOME | 2.22E-03 |
| | APC/C-mediated degradation of cell cycle proteins | REACTOME | 2.22E-03 |
| | activation of the pre-replicative complex | REACTOME | 5.40E-03 |
| | FOXO signaling pathway | KEGG | 2.89E-04 |
| Cell cycle phase | cell cycle, mitotic | REACTOME | 1.37E-06 |
| | cell cycle | REACTOME | 3.11E-07 |
| | mitotic G1-G1/S phases | REACTOME | 1.99E-06 |
| | M phase | REACTOME | 2.94E-05 |
| | Hedgehog 'on' state | REACTOME | 4.65E-05 |
| | G1/S transition | REACTOME | 2.32E-05 |
| | mitotic metaphase and anaphase | REACTOME | 5.36E-05 |
| | mitotic anaphase | REACTOME | 3.56E-05 |
| | separation of sister chromatids | REACTOME | 6.15E-05 |
| | M/G1 transition | REACTOME | 1.99E-04 |
| | mitotic prometaphase | REACTOME | 1.55E-03 |
| | condensation of prophase chromosomes | REACTOME | 3.55E-03 |
| | cell cycle checkpoints | REACTOME | 1.98E-04 |
| | S phase | REACTOME | 3.42E-03 |
| | cell cycle checkpoints | REACTOME | 1.98E-04 |
| | cell cycle | KEGG | 2.54E-06 |

## The human fourth gene eigenvector

### Positive end: neurotransmission, long-term memory, behavior and language development

The positive pole of the human fourth gene eigenvector provides the richest physiological elements and developmental processes involved in the formation of high-level intelligence and cognition (Table S9). They

can be divided into five groups: 1. neurochemical processes and signaling pathways involved in long-term memory formation; 2. neuron-neuron interaction by cell adhesion molecules (CAM); 3. neurotransmission; 4. language and complex behavior; 5. brain and central nervous system development.

### Negative end: Golgi, mitochondrion, lysosome and phospholipid metabolism

The negative pole of the human fourth gene eigenvector is marked by the activities of three important membrane-bound organelles: Golgi, mitochondrion and lysosome, and by the metabolism of lipid (Table S10). Gene subsets associated with mitochondrial organization and functions include cristae formation, regulation of calcium transport, ATP production through respiration, etc. Golgi-related enriched pathways include integral component of Golgi membrane and protein targeting to Golgi. The Golgi plays important roles in the vesicular transport and post-translational modification of protein, as well as lipid transport and lysosome formation. Furthermore, the metabolic pathways of various lipids such as phospholipid, sphingolipid, fatty acid, triglyceride and steroid are enriched near this pole.

**Table S9**. In the sorted loadings of the human fourth gene eigenvector $u_4$, gene subsets enriched at the positive end.

| | Description | Source | P-value |
|---|---|---|---|
| Long-term memory | regulation of synaptic plasticity | GO BP | 6.58E-04 |
| | long-term memory | GO BP | 2.00E-02 |
| | voltage-gated calcium channel complex | GO CC | 1.76E-03 |
| | guanylate cyclase complex, soluble | GO CC | 3.28E-02 |
| | guanylate cyclase activity | GO MF | 2.16E-02 |
| | adenylate cyclase activity | GO MF | 2.56E-02 |
| | guanylate kinase activity | GO MF | 2.38E-02 |
| | neurexin family protein binding | GO MF | 3.65E-02 |
| | cGMP-PKG signaling pathway | KEGG | 4.64E-03 |
| | cAMP signaling pathway | KEGG | 4.31E-02 |
| | long-term memory | GO BP | 2.00E-02 |
| Neuron-neuron interaction | neuron cell-cell adhesion | GO BP | 5.53E-03 |
| | neuromuscular process | GO BP | 3.19E-02 |
| | calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules | GO BP | 5.81E-03 |
| | cell adhesion molecules (CAMs) | KEGG | 1.96E-02 |
| | NCAM1 interactions | REACTOME | 1.90E-03 |
| | L1CAM interactions | REACTOME | 1.76E-02 |
| Neurotransmission | positive regulation of synaptic transmission, glutamatergic | GO BP | 5.92E-03 |
| | presynaptic membrane assembly | GO BP | 7.35E-04 |
| | dendrite development | GO BP | 1.40E-02 |
| | synapse organization | GO BP | 1.78E-02 |
| | synaptic transmission, glutamatergic | GO BP | 4.10E-02 |
| | GABA-A receptor complex | GO CC | 5.07E-04 |
| | excitatory synapse | GO CC | 2.83E-02 |
| | inhibitory synapse | GO CC | 1.03E-02 |

| | | | |
|---|---|---|---|
| | GABA-A receptor activity | GO MF | 1.45E-04 |
| | ionotropic glutamate receptor binding | GO MF | 3.82E-03 |
| | GABAergic synapse | KEGG | 2.58E-04 |
| | glutamatergic synapse | KEGG | 7.07E-04 |
| | axon guidance | REACTOME | 9.57E-05 |
| | GABA receptor activation | REACTOME | 4.40E-04 |
| | transmission across chemical synapses | REACTOME | 2.44E-04 |
| | GABA A receptor activation | REACTOME | 3.47E-04 |
| | neurotransmitter receptor binding and downstream transmission in the postsynaptic cell | REACTOME | 2.33E-03 |
| Language and complex behavior | vocalization behavior | GO BP | 8.09E-04 |
| | adult behavior | GO BP | 6.10E-03 |
| | social behavior | GO BP | 1.49E-03 |
| | visual learning | GO BP | 9.23E-03 |
| | lens development in camera-type eye | GO BP | 5.10E-03 |
| | eyelid development in camera-type eye | GO BP | 4.65E-03 |
| | cochlea morphogenesis | GO BP | 4.29E-04 |
| | middle ear morphogenesis | GO BP | 3.86E-03 |
| Brain and CNS development | dentate gyrus development | GO BP | 1.50E-03 |
| | hippocampus development | GO BP | 7.77E-03 |
| | spinal cord development | GO BP | 4.37E-02 |
| | midbrain development | GO BP | 4.65E-03 |
| | sympathetic nervous system development | GO BP | 1.70E-02 |
| | cell proliferation in forebrain | GO BP | 7.31E-03 |

**Table S10**. In the sorted loadings of the human fourth gene eigenvector $u_4$, gene subsets enriched at the negative end.

| | **Description** | **Source** | **P-value** |
|---|---|---|---|
| Ca2+ transport | regulation of calcium ion transport | GO BP | 3.51E-04 |
| | organic cation transport | REACTOME | 1.01E-02 |
| Protein package and processing in Golgi | protein targeting to Golgi | GO BP | 1.32E-04 |
| | regulation of vesicle fusion | GO BP | 9.08E-05 |
| | integral component of Golgi membrane | GO CC | 1.20E-02 |
| Mitochondrion | mitochondrial transport | GO BP | 9.43E-03 |
| | regulation of mitochondrion organization | GO BP | 3.81E-03 |
| | positive regulation of mitochondrial fission | GO BP | 4.72E-02 |
| | mitochondrial respiratory chain complex III | GO CC | 3.60E-02 |
| | integral component of mitochondrial outer membrane | GO CC | 1.76E-02 |
| | cristae formation | GO BP | 4.08E-02 |
| | acyl-CoA metabolic process | GO BP | 8.04E-03 |
| | cytoplasmic microtubule organization | GO BP | 3.73E-02 |
| Lysosome | membrane fusion | GO BP | 1.40E-02 |

| | | | |
|---|---|---|---|
| | lysosome localization | GO BP | 1.91E-02 |
| | lysosome | KEGG | 1.22E-02 |
| Lipid metabolism | phosphatidic acid biosynthetic process | GO BP | 3.48E-02 |
| | glycosphingolipid metabolic process | GO BP | 1.43E-02 |
| | phosphatidylethanolamine acyl-chain remodeling | GO BP | 1.05E-02 |
| | phosphatidylcholine acyl-chain remodeling | GO BP | 1.53E-02 |
| | glycerophospholipid metabolism | KEGG | 2.19E-02 |
| | glycerophospholipid biosynthesis | REACTOME | 2.97E-04 |
| | phospholipid metabolism | REACTOME | 2.96E-04 |
| | hydrolysis of LPC | REACTOME | 5.89E-03 |
| | synthesis of PG | REACTOME | 1.95E-02 |
| | Acyl chain remodelling of PI | REACTOME | 3.78E-02 |
| | Acyl chain remodelling of PG | REACTOME | 2.19E-02 |
| | Acyl chain remodelling of PC | REACTOME | 1.01E-02 |
| Other lipid metabolism | arachidonic acid metabolic process | GO BP | 2.67E-03 |
| | fatty acid metabolic process | GO BP | 4.65E-02 |
| | fatty acid biosynthesis | KEGG | 4.02E-04 |
| | linoleic acid metabolism | KEGG | 2.26E-02 |
| | triglyceride biosynthesis | REACTOME | 3.71E-02 |
| | steroid metabolic process | GO BP | 4.02E-02 |
| | steroid hormone biosynthesis | KEGG | 4.37E-03 |
| | sphingolipid metabolic process | GO BP | 2.49E-02 |
| | sphingolipid metabolism | REACTOME | 1.90E-02 |
| | lipid particle organization | GO BP | 2.95E-02 |
| | fat digestion and absorption | KEGG | 4.12E-02 |
| | metabolism of lipids and lipoproteins | REACTOME | 6.92E-05 |

## The chimpanzee fourth gene eigenvector

### Positive end: epigenetic regulation of gene expression, RNA processing and transport

The gene subsets enriched at this pole of the chimpanzee fourth gene eigenvector are totally different from human's (Table S11). A prominent example is epigenetic regulation of gene expression, accompanied by gene pathways involved in histone modification and chromatin modelling. Other enriched groups include gene transcription, RNA (especially mRNA) processing and transport, and protein translation.

### Negative end: lipid metabolism, ion transport, synaptic activity and secretion

Compared with humans, the negative pole of the chimpanzee fourth gene eigenvector provides quite different groups of enriched gene subsets (Table S12). Lipid metabolism is the only common group between humans and chimpanzees. Several synaptic activities, which appear at the positive end of the human fourth eigenvector, are enriched at the opposite end in chimpanzees. Despite this, the pathways such as long-term memory, social behavior and visual learning, do not appear here. It's worth noting that grooming behavior is uniquely enriched in chimpanzees. Grooming is a major social activity widely observed in many animals, including human. Primates provide perhaps the richest examples of such behavior. One of the most critical functions of social grooming is to bond and reinforce social relationships among animals who live in close

proximity. Grooming in non-human primates is characterized by louse removal or scratching, and is a time-consuming activity to establish and maintain alliances through dominance hierarchies or pre-existing coalitions. However, in human society, language has taken over the social functions of grooming to manage more complex and extensive social networks.

**Table S11**. In the sorted loadings of the chimpanzee fourth gene eigenvector $u_4$, gene subsets enriched at the positive end.

| | Description | Source | P-value |
|---|---|---|---|
| Epigenetic regulation | positive regulation of gene expression, epigenetic | GO BP | 8.86E-07 |
| | epigenetic regulation of gene expression | REACTOME | 1.07E-08 |
| | histone acetyltransferase binding | GO MF | 2.53E-03 |
| | methyltransferase activity | GO MF | 9.22E-03 |
| | DNA helicase activity | GO MF | 2.99E-03 |
| | telomeric DNA binding | GO MF | 1.45E-02 |
| | histone deacetylation | GO BP | 6.75E-03 |
| | histone H3-K4 demethylation | GO BP | 8.73E-03 |
| | histone H4 acetylation | GO BP | 1.09E-02 |
| | telomere maintenance | GO BP | 1.13E-02 |
| | chromatin silencing at rDNA | GO BP | 1.80E-03 |
| | rRNA methylation | GO BP | 1.43E-03 |
| | negative epigenetic regulation of rRNA expression | REACTOME | 8.98E-09 |
| | NoRC negatively regulates rRNA expression | REACTOME | 1.96E-08 |
| RNA processing and transport | snRNA processing | GO BP | 6.53E-05 |
| | RNA metabolic process | GO BP | 1.18E-03 |
| | RNA transport | KEGG | 6.21E-06 |
| | RNA degradation | KEGG | 5.92E-03 |
| | post-elongation processing of the transcript | REACTOME | 8.02E-06 |
| | regulation of alternative mRNA splicing, via spliceosome | GO BP | 3.06E-04 |
| | RNA splicing, via transesterification reactions | GO BP | 2.31E-03 |
| | 7-methylguanosine mRNA capping | GO BP | 6.32E-03 |
| | regulation of RNA splicing | GO BP | 7.37E-03 |
| | spliceosome | KEGG | 9.91E-14 |
| | t-circle formation | GO BP | 1.98E-03 |
| | maturation of SSU-rRNA | GO BP | 3.14E-03 |
| | mRNA surveillance pathway | KEGG | 4.73E-05 |
| | mRNA splicing - major pathway | REACTOME | 8.79E-11 |
| | mRNA Splicing | REACTOME | 8.79E-11 |
| Transcription | gene expression | REACTOME | 7.48E-25 |
| | transcription | REACTOME | 2.80E-09 |
| | RNA polymerase | KEGG | 9.81E-04 |
| | RNA polymerase II transcription | REACTOME | 7.99E-06 |
| | RNA polymerase II repressing transcription factor binding | GO MF | 7.20E-04 |
| | RNA polymerase I, RNA polymerase III, and mitochondrial | REACTOME | 1.22E-05 |

| | | | |
|---|---|---|---|
| | transcription | | |
| | regulation of transcription from RNA polymerase III promoter | GO BP | 6.52E-05 |
| | RNA polymerase I activity | GO MF | 2.89E-04 |
| | RNA polymerase I promoter escape | REACTOME | 1.88E-05 |
| | RNA polymerase I transcription initiation | REACTOME | 2.98E-05 |
| | RNA polymerase I chain elongation | REACTOME | 3.39E-05 |
| | RNA polymerase I promoter clearance | REACTOME | 9.87E-05 |
| | processing of capped intronless pre-mRNA | REACTOME | 1.07E-04 |
| Translation | translation | REACTOME | 1.81E-05 |
| | ribosome | KEGG | 3.02E-07 |
| | ribosome biogenesis | GO BP | 7.45E-03 |
| | ribosome biogenesis in eukaryotes | KEGG | 7.56E-06 |
| | eukaryotic translation initiation | REACTOME | 4.24E-06 |
| | cap-dependent translation initiation | REACTOME | 4.24E-06 |
| | eukaryotic translation termination | REACTOME | 3.24E-05 |

**Table S12**. In the sorted loadings of the chimpanzee fourth gene eigenvector $u_4$, gene subsets enriched at the negative end.

| | **Description** | **Source** | **P-value** |
|---|---|---|---|
| Phospholipid metabolism | phosphatidic acid biosynthetic process | GO BP | 1.48E-03 |
| | phosphatidylglycerol acyl-chain remodeling | GO BP | 4.11E-03 |
| | phosphatidylinositol acyl-chain remodeling | GO BP | 9.82E-03 |
| | phospholipid biosynthetic process | GO BP | 1.31E-02 |
| | glycerophospholipid metabolism | KEGG | 1.67E-03 |
| | phospholipase D signaling pathway | KEGG | 1.89E-03 |
| | phospholipid metabolism | REACTOME | 1.55E-03 |
| | Acyl chain remodelling of PG | REACTOME | 3.83E-03 |
| | Acyl chain remodelling of PI | REACTOME | 1.09E-02 |
| | synthesis of PG | REACTOME | 1.35E-02 |
| | Acyl chain remodelling of PC | REACTOME | 1.37E-02 |
| | Acyl chain remodelling of PS | REACTOME | 1.48E-02 |
| | Acyl chain remodelling of PE | REACTOME | 2.24E-02 |
| | hydrolysis of LPC | REACTOME | 3.65E-02 |
| Other lipid metabolism | high-density lipoprotein particle | GO CC | 2.77E-02 |
| | fatty acid metabolic process | GO BP | 6.82E-04 |
| | linoleic acid metabolism | KEGG | 5.29E-04 |
| | arachidonic acid metabolism | KEGG | 7.27E-04 |
| | glycerolipid metabolism | KEGG | 2.57E-02 |
| | sphingolipid metabolism | REACTOME | 1.57E-02 |
| | metabolism of lipids and lipoproteins | REACTOME | 3.90E-06 |
| Response to ion | response to zinc ion | GO BP | 3.38E-03 |
| | cation transmembrane transport | GO BP | 6.22E-03 |

| | | | |
|---|---|---|---|
| | response to iron ion | GO BP | 7.16E-03 |
| | regulation of sodium ion transport | GO BP | 1.42E-02 |
| | regulation of calcium ion transport | GO BP | 1.83E-02 |
| | response to lead ion | GO BP | 2.32E-02 |
| | positive regulation of calcium ion import | GO BP | 2.59E-02 |
| Short-term synaptic plasticity and transmission | regulation of short-term neuronal synaptic plasticity | GO BP | 6.98E-03 |
| | calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules | GO BP | 2.17E-02 |
| | neuromuscular synaptic transmission | GO BP | 2.37E-02 |
| | regulation of neuronal synaptic plasticity | GO BP | 2.69E-02 |
| | neuronal action potential | GO BP | 3.02E-02 |
| | neuromuscular junction | GO CC | 4.39E-02 |
| | calcium signaling pathway | KEGG | 8.10E-04 |
| | neuroactive ligand-receptor interaction | KEGG | 2.25E-02 |
| Behavior | adult locomotory behavior | GO BP | 2.66E-02 |
| | grooming behavior | GO BP | 3.09E-02 |
| Secretion | regulation of protein secretion | GO BP | 1.47E-02 |
| | bile secretion | KEGG | 1.34E-03 |
| | insulin secretion | KEGG | 1.57E-03 |
| | pancreatic secretion | KEGG | 3.76E-03 |
| | endocrine resistance | KEGG | 8.00E-03 |
| | salivary secretion | KEGG | 1.54E-02 |

## The human fifth gene eigenvector

### Positive end: muscle contraction, synaptic activity and lipid transport

Three groups of gene subsets are enriched at this pole: muscle contraction, synaptic activity and lipid transport (Table S13). All of them can be found at the same end of chimpanzee fifth gene eigenvector too (Table S15). We previously proposed that the human 4-th level provided most abundant physiological elements underlying cognitive functions and language development. Although many of them are shown at the human and chimpanzee 5-th levels too, several important pathways such as long-term memory, social behavior, vocalization behavior and visual learning are not enriched here. On the other hand, they do not appear at the chimpanzee 4-th level either.

### Negative end: translation and RNA regulation

The negative end of the human fifth gene eigenvector is marked by protein synthesis, folding and transport, mRNA processing in gene expression and non-coding RNA regulation (Table S14).

**Table S13**. In the sorted loadings of the human fifth gene eigenvector $u_5$, gene subsets enriched at the positive end.

| | Description | Source | P-value |
|---|---|---|---|
| Muscle contraction | muscle filament sliding | GO BP | 1.11E-07 |
| | skeletal muscle contraction | GO BP | 4.16E-04 |
| | cardiac muscle contraction | GO BP | 2.41E-06 |

| | | | |
|---|---|---|---|
| | sarcomere organization | GO BP | 8.52E-06 |
| | striated muscle contraction | GO BP | 3.42E-04 |
| | cardiac muscle contraction | KEGG | 6.41E-05 |
| | striated muscle contraction | REACTOME | 2.39E-07 |
| | muscle contraction | REACTOME | 2.08E-05 |
| Neurotransmission and synaptic plasticity | neuromuscular synaptic transmission | GO BP | 1.90E-05 |
| | membrane depolarization during action potential | GO BP | 2.50E-05 |
| | regulation of long-term neuronal synaptic plasticity | GO BP | 7.56E-04 |
| | regulation of neuronal synaptic plasticity | GO BP | 5.72E-04 |
| | neuronal action potential | GO BP | 1.17E-04 |
| | positive regulation of excitatory postsynaptic potential | GO BP | 3.11E-03 |
| | regulation of calcium ion-dependent exocytosis | GO BP | 1.63E-03 |
| | regulation of neurotransmitter secretion | GO BP | 1.61E-03 |
| | synaptic vesicle exocytosis | GO BP | 2.60E-03 |
| | regulation of synaptic plasticity | GO BP | 2.18E-03 |
| | neuroactive ligand-receptor interaction | KEGG | 5.49E-07 |
| | cholinergic synapse | KEGG | 1.00E-03 |
| | serotonergic synapse | KEGG | 1.91E-03 |
| | glutamatergic synapse | KEGG | 9.33E-03 |
| | calcium signaling pathway | KEGG | 9.37E-03 |
| | transmission across chemical synapses | REACTOME | 2.32E-06 |
| | neurotransmitter receptor binding and downstream transmission in the postsynaptic cell | REACTOME | 3.75E-05 |
| | ion channel transport | REACTOME | 2.01E-03 |
| | NCAM signaling for neurite out-growth | REACTOME | 7.46E-04 |
| Lipid transport | high-density lipoprotein particle remodeling | GO BP | 4.59E-05 |
| | triglyceride catabolic process | GO BP | 8.75E-05 |
| | chylomicron remodeling | GO BP | 9.86E-05 |
| | reverse cholesterol transport | GO BP | 1.58E-04 |
| | lipoprotein transport | GO BP | 5.09E-03 |
| | cholesterol efflux | GO BP | 2.61E-03 |
| | low-density lipoprotein particle remodeling | GO BP | 1.94E-03 |
| | phospholipid efflux | GO BP | 2.37E-04 |
| | HDL-mediated lipid transport | REACTOME | 1.68E-05 |

**Table S14**. In the sorted loadings of the human fifth gene eigenvector $u_5$, gene subsets enriched at the negative end.

| | **Description** | **Source** | **P-value** |
|---|---|---|---|
| Translation | translation | REACTOME | 4.12E-20 |
| | mitochondrial translation | GO BP | 3.16E-07 |
| | mitochondrial translation initiation | REACTOME | 3.11E-09 |
| | mitochondrial translation termination | REACTOME | 1.50E-09 |
| | mitochondrial translation elongation | REACTOME | 7.05E-09 |
| | cytoplasmic translation | GO BP | 8.95E-05 |
| | positive regulation of translational initiation | GO BP | 1.17E-02 |

| | | | |
|---|---|---|---|
| | cap-dependent translation initiation | REACTOME | 1.14E-17 |
| | eukaryotic translation elongation | REACTOME | 2.46E-16 |
| | eukaryotic translation termination | REACTOME | 1.20E-16 |
| | tRNA binding | GO MF | 7.68E-06 |
| | tRNA aminoacylation for protein translation | GO BP | 3.22E-04 |
| | peptide chain elongation | REACTOME | 2.45E-16 |
| Protein fold and transport | protein import into nucleus | GO BP | 1.03E-05 |
| | chaperone-mediated protein folding | GO BP | 8.35E-03 |
| | 'de novo' protein folding | GO BP | 2.13E-02 |
| | retrograde protein transport, ER to cytosol | GO BP | 8.50E-03 |
| | protein targeting to mitochondrion | GO BP | 3.51E-04 |
| | protein localization to cilium | GO BP | 2.50E-04 |
| | protein import into mitochondrial matrix | GO BP | 9.11E-03 |
| | protein folding in endoplasmic reticulum | GO BP | 1.97E-02 |
| | protein methylation | GO BP | 5.74E-03 |
| | protein export | KEGG | 5.14E-05 |
| RNA processing | RNA metabolic process | GO BP | 4.00E-04 |
| | RNA secondary structure unwinding | GO BP | 5.35E-04 |
| | mRNA polyadenylation | GO BP | 1.17E-04 |
| | tRNA methylation | GO BP | 1.11E-02 |
| | production of siRNA involved in RNA interference | GO BP | 3.62E-02 |
| | RNA transport | KEGG | 1.22E-11 |
| | RNA degradation | KEGG | 2.47E-05 |
| | mRNA surveillance pathway | KEGG | 3.66E-08 |
| RNA regulation | regulation of gene silencing by miRNA | GO BP | 3.30E-05 |
| | chromatin silencing | GO BP | 9.59E-05 |
| | primary miRNA processing | GO BP | 7.61E-03 |
| | snoRNA binding | GO MF | 1.56E-03 |
| | regulatory RNA pathways | REACTOME | 6.28E-15 |
| | chromatin modifying enzymes | REACTOME | 3.33E-10 |
| | chromatin organization | REACTOME | 3.33E-10 |

## The chimpanzee fifth gene eigenvector

### Positive end: muscle contraction, synaptic transmission, lipid transport and ECM organization

Most enriched gene groups around the positive end in chimpanzee are found in human too, except for the group ECM organization (Table S15).

### Negative end: translation and RNA regulation

The groups of enriched gene subsets near the negative end of the chimpanzee fifth eigenvector are almost like those at the same end of the human fifth eigenvector (Table S16).

**Table S15**. In the sorted loadings of the chimpanzee fifth gene eigenvector $u_5$, gene subsets enriched at the positive end.

| | Description | Source | P-value |
|---|---|---|---|
| Muscle contraction | muscle contraction | REACTOME | 3.26E-07 |
| | muscle filament sliding | GO BP | 3.87E-07 |
| | skeletal muscle contraction | GO BP | 1.36E-05 |
| | cardiac muscle contraction | GO BP | 5.15E-05 |
| | striated muscle contraction | REACTOME | 2.84E-07 |
| Lipid transport | phospholipid efflux | GO BP | 8.39E-06 |
| | chloride transport | GO BP | 2.21E-05 |
| | chylomicron remodeling | GO BP | 1.52E-04 |
| | high-density lipoprotein particle remodeling | GO BP | 3.20E-04 |
| | lipoprotein metabolic process | GO BP | 3.56E-04 |
| | positive regulation of cholesterol efflux | GO BP | 3.96E-04 |
| | high-density lipoprotein particle assembly | GO BP | 8.80E-04 |
| | lipoprotein transport | GO BP | 5.19E-03 |
| | low-density lipoprotein particle remodeling | GO BP | 6.16E-03 |
| | lipid transporter activity | GO MF | 5.76E-03 |
| | lipid digestion, mobilization, and transport | REACTOME | 4.86E-06 |
| | HDL-mediated lipid transport | REACTOME | 4.04E-05 |
| Synaptic transmission and synapse formation | membrane depolarization during action potential | GO BP | 1.59E-04 |
| | neuromuscular synaptic transmission | GO BP | 1.69E-04 |
| | membrane depolarization | GO BP | 1.12E-03 |
| | calcium ion-regulated exocytosis of neurotransmitter | GO BP | 2.22E-03 |
| | regulation of synaptic plasticity | GO BP | 2.47E-03 |
| | regulation of neuronal synaptic plasticity | GO BP | 4.09E-03 |
| | synaptic transmission, cholinergic | GO BP | 5.09E-03 |
| | positive regulation of calcium ion-dependent exocytosis | GO BP | 6.14E-03 |
| | GABA-A receptor complex | GO CC | 7.44E-03 |
| | postsynapse | GO CC | 7.47E-03 |
| | dopaminergic synapse | KEGG | 1.23E-04 |
| | cholinergic synapse | KEGG | 8.85E-04 |
| | glutamatergic synapse | KEGG | 6.74E-04 |
| | neuroactive ligand-receptor interaction | KEGG | 2.79E-03 |
| | GABAergic synapse | KEGG | 6.52E-03 |
| | long-term potentiation | KEGG | 8.65E-03 |
| | long-term depression | KEGG | 3.53E-02 |
| | neurotransmitter receptor binding and downstream transmission in the postsynaptic cell | REACTOME | 2.47E-04 |
| | regulation of long-term neuronal synaptic plasticity | GO BP | 1.83E-02 |
| | synapse organization | GO BP | 2.32E-03 |
| | positive regulation of dendrite extension | GO BP | 2.33E-03 |
| | NCAM signaling for neurite out-growth | REACTOME | 1.68E-05 |
| Homeostasis | triglyceride homeostasis | GO BP | 2.88E-04 |
| | homeostasis | REACTOME | 5.90E-07 |

| | | | |
|---|---|---|---|
| ECM organization | ECM-receptor interaction | KEGG | 4.94E-05 |
| | extracellular matrix organization | REACTOME | 2.00E-12 |
| | degradation of the extracellular matrix | REACTOME | 2.79E-08 |
| | ECM proteoglycans | REACTOME | 3.78E-07 |
| | activation of matrix metalloproteinases | REACTOME | 1.82E-05 |
| | collagen degradation | REACTOME | 2.01E-05 |

**Table S16**. In the sorted loadings of the chimpanzee fifth gene eigenvector $u_5$, gene subsets enriched at the negative end.

| | **Description** | **Source** | **P-value** |
|---|---|---|---|
| Translation | translation | REACTOME | 4.66E-19 |
| | mitochondrial translation | GO BP | 2.23E-06 |
| | mitochondrial translation | REACTOME | 1.07E-11 |
| | mitochondrial translation elongation | REACTOME | 3.06E-10 |
| | mitochondrial translation initiation | REACTOME | 7.96E-10 |
| | cytoplasmic translation | GO BP | 2.03E-04 |
| | eukaryotic translation initiation | REACTOME | 1.02E-13 |
| | eukaryotic translation termination | REACTOME | 1.51E-10 |
| | translesion synthesis | GO BP | 9.38E-05 |
| | peptide chain elongation | REACTOME | 6.24E-10 |
| | tRNA aminoacylation for protein translation | GO BP | 6.50E-05 |
| | eukaryotic translation initiation factor 3 complex | GO CC | 1.09E-05 |
| | tRNA binding | GO MF | 1.42E-08 |
| | rRNA binding | GO MF | 1.02E-06 |
| | ribosome binding | GO MF | 2.65E-06 |
| | translation initiation factor binding | GO MF | 4.69E-03 |
| | regulation of translational initiation | GO BP | 9.25E-05 |
| | regulation of translation | GO BP | 8.63E-03 |
| RNA processing, degradation and transport | RNA secondary structure unwinding | GO BP | 5.47E-06 |
| | tRNA modification | GO BP | 2.26E-05 |
| | histone mRNA metabolic process | GO BP | 2.54E-05 |
| | tRNA export from nucleus | GO BP | 2.98E-04 |
| | RNA catabolic process | GO BP | 3.01E-04 |
| | rRNA methylation | GO BP | 6.93E-04 |
| | mRNA cleavage | GO BP | 6.92E-04 |
| | RNA metabolic process | GO BP | 1.51E-03 |
| | mRNA polyadenylation | GO BP | 1.84E-03 |
| | tRNA processing | GO BP | 1.26E-02 |
| | tRNA methylation | GO BP | 1.40E-02 |
| | RNA transport | KEGG | 3.59E-12 |
| | mRNA surveillance pathway | KEGG | 4.26E-04 |
| | RNA degradation | KEGG | 6.93E-04 |
| Protein processing | protein import into nucleus | GO BP | 2.03E-05 |

| | | | |
|---|---|---|---|
| and transport | 'de novo' protein folding | GO BP | 7.78E-05 |
| | protein targeting to mitochondrion | GO BP | 2.06E-04 |
| | protein targeting to lysosome | GO BP | 2.16E-04 |
| | protein import into mitochondrial matrix | GO BP | 4.55E-04 |
| | chaperone-mediated protein folding | GO BP | 1.05E-03 |
| | protein destabilization | GO BP | 2.31E-03 |
| | protein methylation | GO BP | 3.77E-03 |
| | protein folding in endoplasmic reticulum | GO BP | 6.63E-03 |
| | chaperone mediated protein folding requiring cofactor | GO BP | 7.63E-03 |
| | negative regulation of protein ubiquitination | GO BP | 7.12E-03 |
| | protein processing in endoplasmic reticulum | KEGG | 6.32E-09 |
| | protein export | KEGG | 4.35E-07 |
| | mitochondrial protein import | REACTOME | 2.52E-07 |
| RNA regulation | regulation of gene silencing by miRNA | GO BP | 1.07E-04 |
| | chromatin silencing | GO BP | 8.28E-03 |
| | gene silencing by RNA | GO BP | 9.86E-03 |
| | snoRNA binding | GOMF | 2.80E-04 |
| | Transcriptional regulation by small RNAs | REACTOME | 1.49E-05 |
| | chromosome maintenance | REACTOME | 1.81E-10 |

## The human/chimpanzee sixth gene eigenvector

### Positive end: mesenchymal stem cell differentiation

At the positive pole of the human sixth eigenvector, we observe the differentiation of mesenchymal stem cells into several cell linages including osteoblasts, chondrocytes which give rise to cartilage, myocytes and adipocytes (Table S17).

### Negative end: DNA repair

**Table S17**. In the sorted loadings of the human sixth gene eigenvector $u_6$, gene subsets enriched at the positive end.

| | Description | Source | P-value |
|---|---|---|---|
| Mesenchymal stem cell differentiation | positive regulation of mesenchymal cell proliferation | GO BP | 3.03E-03 |
| | skin development | GO BP | 1.28E-03 |
| | white fat cell differentiation | GO BP | 3.98E-03 |
| | chondrocyte differentiation | GO BP | 1.23E-02 |
| | positive regulation of cartilage development | GO BP | 1.45E-02 |
| | positive regulation of cell division | GO BP | 4.69E-03 |
| | myoblast differentiation | GO BP | 3.64E-02 |
| | osteoblast development | GO BP | 2.89E-02 |
| | chondrocyte proliferation | GO BP | 2.05E-02 |
| | bone remodeling | GO BP | 1.29E-02 |
| | myoblast fusion | GO BP | 4.10E-02 |
| | collagen trimer | GO CC | 4.84E-05 |
| | collagen formation | REACTOME | 2.89E-01 |

**Table S18**. In the sorted loadings of the human sixth gene eigenvector $u_6$, gene subsets enriched at the negative end.

| | Description | Source | P-value |
|---|---|---|---|
| DNA repair | interstrand cross-link repair | GO BP | 3.37E-05 |
| | nucleotide-excision repair, DNA incision | GO BP | 1.78E-04 |
| | nucleotide-excision repair | GO BP | 1.08E-03 |
| | global genome nucleotide-excision repair | GO BP | 1.71E-04 |
| | nucleotide-excision repair, DNA incision, 5'-to lesion | GO BP | 6.97E-04 |
| | nucleotide-excision repair, preincision complex stabilization | GO BP | 6.37E-03 |
| | nucleotide-excision repair, DNA incision, 3'-to lesion | GO BP | 5.54E-03 |
| | nucleotide-excision repair, DNA duplex unwinding | GO BP | 2.92E-03 |
| | nucleotide excision repair | KEGG | 1.35E-02 |
| | p53-Independent DNA damage response | REACTOME | 9.42E-03 |
| | DNA repair | REACTOME | 1.01E-03 |
| | base excision repair | REACTOME | 4.75E-03 |

## Human cognition-language gene eigenvector with respect to the chimpanzee's fourth and fifth eigenvectors

The rotation of the human fourth and fifth motif eigenvectors leads to a gene eigenvector encoding human-specific phenotypes. The uniqueness of this human module is supported by a group of gene subsets which are enriched near the positive pole of the human fourth eigenvector, but not or less significant along the ape fourth eigenvector. We find that those gene subsets are closely associated with high-order cognitive functions, language development and complex behavior, see Table S19 and Fig. 6B.

**Table S19.** gene subsets that are significantly enriched near the pole of the human fourth polarized gene eigenvector, but are not or are less significantly enriched along the fourth ape polarized gene eigenvector.

| | Description | Source | Human 4 | Chimp 4 | Chimp 5 |
|---|---|---|---|---|---|
| Synaptic activity and synapse development | transmission across chemical synapses | REACTOME | 2.44E-04 | 6.62E-01 | 2.98E-05 |
| | neurotransmitter receptor binding and downstream transmission in the postsynaptic cell | REACTOME | 2.33E-03 | 6.12E-01 | 2.47E-04 |
| | excitatory synapse | GO CC | 2.83E-02 | 2.16E-01 | 1.71E-01 |
| | presynaptic membrane assembly | GO BP | 7.35E-04 | 4.56E-01 | 1.93E-01 |
| | neurexin family protein binding | GO MF | 3.65E-02 | 4.99E-01 | 4.00E-01 |
| | neurofascin interactions | REACTOME | 2.03E-03 | 2.68E-01 | 1.95E-03 |
| | NCAM signaling for neurite out-growth | REACTOME | 2.13E-02 | 6.31E-01 | 1.68E-05 |
| | axon guidance | REACTOME | 9.57E-05 | 2.53E-01 | 3.99E-07 |
| | dendrite development | GO BP | 1.40E-02 | 2.00E-01 | 3.23E-02 |
| | neuron cell-cell adhesion | GO BP | 5.53E-03 | 3.54E-01 | 3.96E-02 |
| | neuromuscular process | GO BP | 3.19E-02 | 8.15E-02 | 1.40E-01 |

| | | | | | |
|---|---|---|---|---|---|
| | positive regulation of synaptic transmission, glutamatergic | GO BP | 5.92E-03 | 7.84E-02 | 2.71E-01 |
| | ionotropic glutamate receptor binding | GO MF | 3.82E-03 | 1.09E-01 | 2.69E-01 |
| | extracellular-glutamate-gated ion channel activity | GO MF | 4.80E-02 | 5.22E-01 | 1.20E-01 |
| | glutamatergic synapse | KEGG | 7.07E-04 | 2.57E-01 | 6.74E-04 |
| | GABA receptor activation | REACTOME | 4.40E-04 | 5.77E-02 | 3.09E-02 |
| | dopaminergic synapse | KEGG | 2.41E-02 | 1.34E-01 | 1.23E-04 |
| Language and complex behavior | adult behavior | GO BP | 6.10E-03 | 1.19E-01 | 6.00E-02 |
| | social behavior | GO BP | 1.49E-03 | 2.88E-01 | 3.22E-01 |
| | visual learning | GO BP | 9.23E-03 | 1.72E-01 | 1.16E-01 |
| | vocalization behavior | GO BP | 8.09E-04 | 6.60E-02 | 7.31E-03 |
| | middle ear morphogenesis | GO BP | 3.86E-03 | 2.33E-02 | 5.11E-01 |
| | cochlea morphogenesis | GO BP | 4.29E-04 | 5.33E-01 | 1.30E-02 |
| | eyelid development in camera-type eye | GO BP | 4.65E-03 | 1.27E-01 | 1.27E-01 |
| | lens development in camera-type eye | GO BP | 5.10E-03 | 8.70E-02 | 1.69E-02 |
| Organ and Neural development | midbrain development | GO BP | 4.65E-03 | 1.42E-01 | 1.42E-01 |
| | sympathetic nervous system development | GO BP | 1.70E-02 | 1.05E-01 | 5.14E-02 |
| | cell proliferation in forebrain | GO BP | 7.31E-03 | 3.34E-01 | 4.38E-02 |
| | embryonic cranial skeleton morphogenesis | GO BP | 2.09E-02 | 2.07E-01 | 7.74E-01 |
| | spinal cord development | GO BP | 4.37E-02 | 7.84E-02 | 5.59E-03 |
| | neuronal system | REACTOME | 7.62E-05 | 9.36E-01 | 8.01E-06 |
| Memory | long-term memory | GO BP | 2.00E-02 | 5.50E-02 | 2.25E-01 |
| | regulation of synaptic plasticity | GO BP | 6.58E-04 | 2.59E-01 | 2.47E-03 |
| | guanylate cyclase activity | GO MF | 2.16E-02 | 9.12E-01 | 5.10E-01 |
| | adenylate cyclase activity | GO MF | 2.56E-02 | 9.21E-01 | 4.83E-01 |
| | cGMP-PKG signaling pathway | KEGG | 4.64E-03 | 9.23E-01 | 3.84E-02 |
| | cAMP signaling pathway | KEGG | 4.31E-02 | 8.11E-01 | 8.87E-03 |
| | guanylate cyclase complex, soluble | GO CC | 3.28E-02 | 9.65E-01 | 4.09E-01 |
| | dentate gyrus development | GO BP | 1.50E-03 | 1.71E-02 | 1.23E-01 |
| | hippocampus development | GO BP | 7.77E-03 | 3.50E-02 | 9.86E-01 |

# Regulators unveiled by the dual motif eigenvectors

## Embryogenesis regulators near one pole of the human first motif eigenvector

The first motif eigenvector $v_1$ is coupled with the first gene eigenvector $u_1$, whose key functions are

embryogenesis and reproduction at the two ends (Fig. 3A). In this vector, the cis-elements near the positive end correspond to the genes associated with reproduction whereas those near the negative end correspond to the genes involved in embryogenesis.

Embryogenesis starts with the germinal stage, when the early embryo successfully is implanted in the uterus, continues with the next stage of gastrulation, when the three germ layers of the embryo form in a process, and the processes of neurulation and organogenesis follow. We indeed find a group of motifs at the dual end whose binding factors have been reported to participate in these stages (Table S20). Other functional studies in murine models have proposed the important roles of ZIC family transcription factors in neuroectodermal development and neural crest cell induction, indicated by their temporal-spatial expression pattern (Houtmeyers, et al. 2013). ZIC2 is the earliest of the murine ZIC genes to be expressed at the stage of blastulation. The expression of ZIC2 and ZIC3 is maintained in the ectoderm, newly-formed mesoderm and pre-somitic populations during gastrulation. Later, during neurulation, their expression is limited to in most dorsal neuroectoderm where neural crest cell and dorsal neuron will produce. ZIC1 are first expressed at the somite stage of development in the neuroectoderm and somitic mesoderm. We also found other regulators at the somite stage including MRF4 and MEIS2. MRF4, cooperating with other myogenic factors, plays an abundant role in somitic myogenesis (Zhang, et al. 1995). Inactivation of MRF4 results in rib malformation. MEIS2, a member of TALE-family homeodomain proteins, serves as a cell marker in the dorso-ectodermal region and potentially induce myogenesis in the underlying paraxial mesoderm (Cecconi, et al. 1997; Machon, et al. 2015). Additionally, it is also highly expressed in the differentiating regions of forebrain. Another member of TALE-family, MEIS3, directs neural anteroposterior (AP) patterning as a direct target of canonical Wnt signaling (Elkouby, et al. 2010). HNF1B functions before gastrulation and specifies visceral endoderm (Barbacci, et al. 1999). CHCH functions as an important switch between gastrulation and neurulation via FGF signaling (Sheng, et al. 2003). Loss of NKX2-9, the murine homolog of NKX2-8, results in impaired floor plate, the important structure located on the ventral midline of neural tube mediating the proper left-right body axis by commissural axon guidance (Holz, et al. 2010). The heart is the first functional organ to develop in human embryo. TBX5 plays a critical role in heart morphogenesis, and its dysregulation leads to abnormal cardiac morphogenesis and congenital heart diseases (Horb and Thomsen 1999).

**Table S20.** Motifs of embryogenesis regulators at the negative end of the first motif eigenvector. The last column provides the PubMed ID for each reference.

| Motif symbol | Transcription factor | Rank | Function | Reference |
|---|---|---|---|---|
| ZIC3_05 | ZIC3 | 3 | ectoderm and mesoderm formation | 23443491 |
| ZIC2_05 | ZIC2 | 4 | neuroectodermal development | |
| ZIC1_01 | ZIC1 | 7 | somite formation | |
| MYF6_04 | MRF4 | 11 | myotome differentiation | 7797078 |
| LFA1_Q6 | HNF1B | 13 | visceral endoderm specification | 10518496 |
| TBX5_Q5 | TBX5 | 15 | heart morphogenesis | 10079235 |
| CHCH_01 | CHCH | 26 | regulation of gastrulation and neurulation | 14651851 |
| NKX29_01 | NKX2-8 | 42 | floor plate development | 21068056 |
| MRG2_01 | MEIS3 | 47 | neural plate organization | 20356957 |
| MEIS2_01 | MEIS2 | 51 | forebrain differentiation<br>somitic mesoderm differentiation | 9337138<br>26545946 |

| | | | | |
|---|---|---|---|---|
| | | | neural crest development | |

## Cell cycle regulators near one pole of the human third motif eigenvector

The third motif eigenvector $v_3$ is coupled with the third gene eigenvector $u_3$. In this eigenvector, the cis-elements near the negative end correspond to the genes near the mitosis end of the coupling gene eigenvector. In consistency with the natural regulatory correspondence between them, several cis-elements whose binding factors are known to participate in the cell cycle regulation such as E2F1/DP2 and FOXO1/3/4, are found to cluster at this end (Table S21).

**Table S21. Motifs of cell cycle regulators near the negative end of the third motif eigenvector.** The last column provides the PubMed IDs for the related references.

| Motif symbol | Transcription factor | Rank | Consensus sequence | Phase | Reference |
|---|---|---|---|---|---|
| E2F_01 | E2F1 | 106 | GCGCCAAAA | G1-S,S-G2 | 11799067 |
| FOXO3A_Q1 | FOXO3A | 33 | TGTAAACAA | S-G2,G2-M | 11964479 |
| FOXO1_Q5 | FOXO1 | 42 | AAAACA | G1 | 10783894 |
| FOXO4_02 | FOXO4 | 55 | KTTGTTTAC | G1 | 10783894 |
| E2F1DP2_01 | E2F1:DP2 | 102 | TTTSSCGC | G1-S | 7880534<br>7739537 |
| NANOG_02 | NANOG | 34 | AACAAAGG | core reprograming factor | 18029452<br>18035408 |
| OCT4_02 | POU5F1 | 69 | ATTGTNATGCTAAT | | |
| SOX2_01 | SOX2 | 71 | CCTTTGTNNTGCAAA | | |

## Cis-trans regulation of the human cognitive eigenvector

The human fourth motif eigenvector $v_4$ results from the rotation of the ape eigenvectors, and is coupled to the cognitive gene eigenvector (Table S19). Among the top motifs at the positive end, quite a few motifs, referred as new-comers in this paper, are boosted from the fifth eigenvector of chimpanzees. In other words, they rank high at the fifth level but rank quite low at the fourth level in chimpanzees, but are now prominent at the fourth eigenvector in humans. To classify the old-timers and newcomers clearly, a linear regression of the motif loadings at level 4 in human (*y*) on their loadings at the same level in chimpanzee (*x*) is performed. The regression residuals (*e*) reflect the influence of chimpanzee level 5 since the human fourth and fifth eigenvectors are totally linearly dependent on those of chimpanzee's (Fig.2B). For each motif, we take the ratio (*e/y*) of the regression residual to its current loading at human level 4 as the percentage of contribution by chimpanzee level 5. If the ratio is greater than 0.5, the motif is identified as a new-comer. We discover that nearly one quarter of the top motifs, 47 out 200, are new-comers. Several transcription factors, whose functions in long-term memory storage have been studied extensively, stand around the two ends. Two of them, EGR1 and CREB, are new-comers. Three of them, CREB, CEBP and AP1, are MPA (Table S22). The seven cis-elements in the SP1-centered network, which is composed by six new-comers and one old-timer (Fig.4C), are also included in Table S22.

**Table S22. Regulators of human cognitive gene eigenvector.** Two motifs sitting at the negative end are marked by negative ranks. The last column provides the PubMed IDs for the related references.

| Motif symbol | Transcription factor | Human 4 | Chimp 4 | Chimp 5 | Reference |
|---|---|---|---|---|---|
| **Non-MPA & new-comer motifs** | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| GKLF_Q4 | KLF4 | 32 | 220 | 11 | 9651398 |
| GKLF_02 | | 84 | 416 | 14 | |
| MAZ_Q6 | MAZ | 35 | 165 | 18 | 12425938 |
| MAZ_Q6_01 | | 66 | 244 | 20 | |
| EGR1_04 | EGR1 | 68 | 203 | 30 | 19126756 |
| EGR2_Q6 | EGR2 | 114 | 514 | 19 | |
| EGR_Q6 | EGR1/2 | 203 | 440 | 69 | |
| SP3_Q3 | SP3 | 140 | 317 | 57 | 12736330 |
| LRF_Q2 | ZBTB7A(LRF) | 39 | 381 | 7 | 12004059; 15917220 |
| BTEB3_Q5 | KLF13 | 67 | 226 | 22 | 11477107; 12004059 |
| **MPA & new-comer motifs** | | | | | |
| SP1_03 | SP1 | 28 | 262 | 9 | 15081116 |
| GC_01 | | 123 | 413 | 25 | |
| SP1_Q2_01 | | 131 | 457 | 26 | |
| SP1_02 | | 129 | 393 | 31 | |
| SP1_Q4_01 | | 168 | 551 | 34 | |
| CREBATF_Q6 | CREB | -184 | 415 | -38 | 19126756 |
| TAXCREB_02 | CREB1 | -112 | -288 | -71 | 19126756 |
| **Non-MPA & old-timer motifs** | | | | | |
| ELF1_Q6 | ELF1 | 38 | 55 | 154 | 8797822 |
| P50_Q6 | NFKB1 | 81 | 149 | 45 | 19126756 |
| **MPA & old-timer motifs** | | | | | |
| CEBP_Q3 | CEBPA/ CEBPB/ CEBPD | 16 | 5 | -245 | 19126756 |
| CEBPA_Q6 | | 57 | 35 | -273 | |
| CEBP_Q2_01 | | 74 | 42 | -229 | |
| CEBPB_Q6 | | 93 | 54 | -249 | |
| CEBPD_Q6_01 | | 190 | 155 | -259 | |
| AP1_02 | AP1 | -197 | -187 | 398 | 19126756 |

# Comparative dual eigen-analysis between human and orangutan

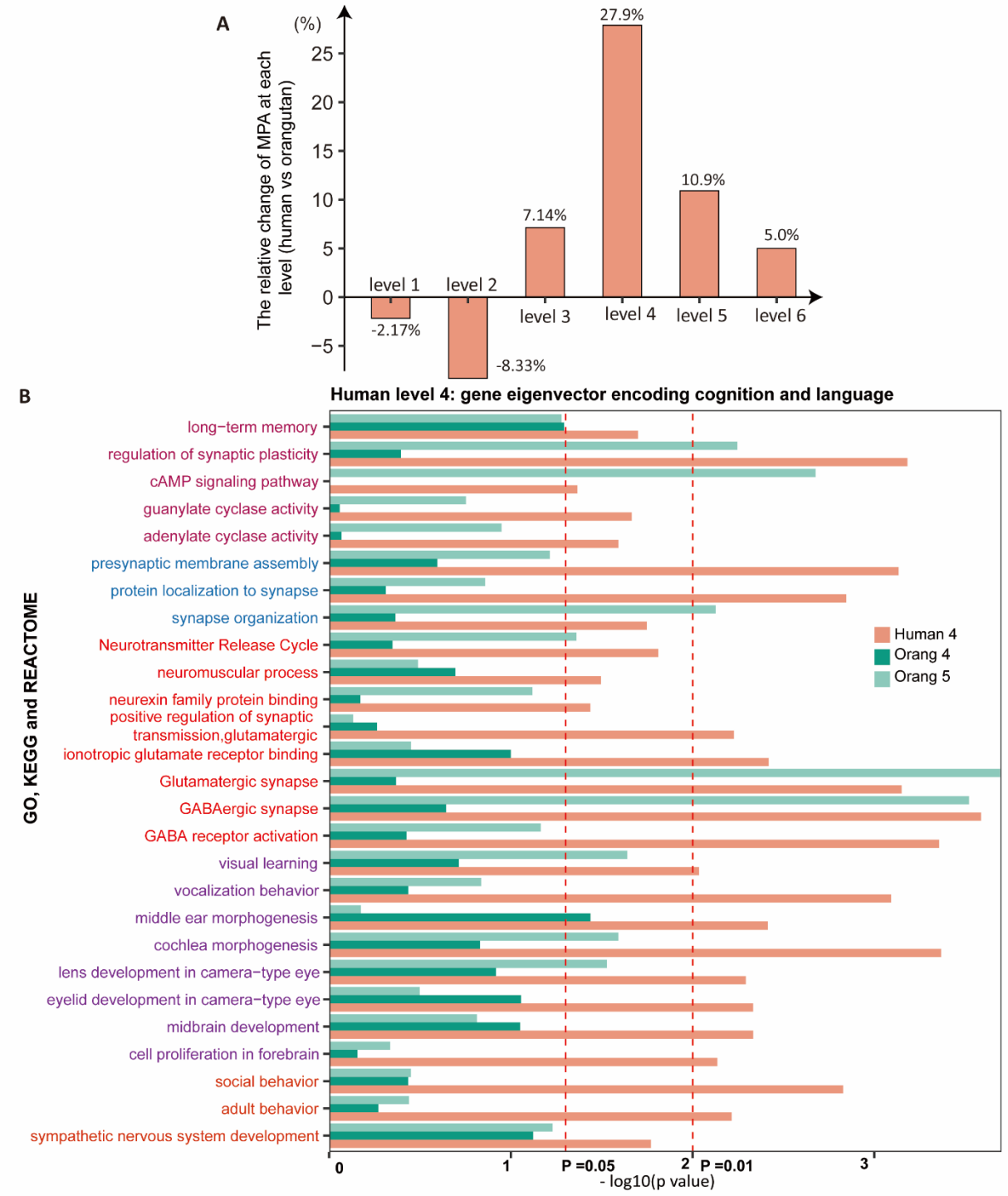



**Fig. S5 (A) The relative change of MPA (motifs present on Alu) in percentages at each level from orangutan to human.** The number of MPA increases most significantly at level 4 by 27.9%, followed by 10.9% at the fifth level. **(B) Gene subsets that are significantly enriched near the positive pole of the human fourth polarized gene eigenvector, but are not or less significantly enriched along the fourth orangutan polarized gene eigenvector.** The color scheme is the same as that in Fig.6.

## Human specific Alu insertions in the proximal regulatory regions

8,817 human specific Alu elements were identified recently by comparing the genomes of human and nine nonhuman primates including chimpanzee (Tang, et al. 2018). We filtered out those falling out of the proximal regulatory regions, namely, -1000bp to 500bp around TSS of the human 5'-most transcripts, mostly principal ones. This resulted in 47 genes, each of which contains a clean human specific Alu insertion. A list of the 47 genes is shown in Table S24. Each line describes one gene, including its name, the type of Alu insertions, the distance to TSS, the difference between its rank in the fourth eigenvector of human and that of chimpanzee, and its gene function. The gene ranks in the two species were based on the subset of genes whose annotations are available for both species. The gene functions are obtained from UniProt (The UniProt Consortium 2019), if not specified otherwise.

**Table S23 Statistical comparisons of the occurrences of a MPA in the 47 human specific Alu insertions versus those in their neighboring regulatory DNA sequences.** The comparisons were quantified by significances of statistical tests. The significances were calculated for each MPA across the 47 genes with human specific Alu insertions as follows. First, for a given MPA, we computed its frequency $p_0$ in the regulatory DNA sequence of one gene excluding the associated human specific Alu insertion. Then we counted its occurrences in the Alu insertion. Second, the count is assumed to be a random variable following a Binomial distribution $Bin(n, p)$, where $n$ is the length of the Alu insertion. We tested the hypothesis $H_0$: $p = p_0$ versus $H_a$: $p > p_0$, and a p-value was calculated using the exact binomial test. Third, a total of 47 p-values were obtained as above, and they were combined by the Fisher's method (Fisher 1925; Fisher 1948). Shown in the table are the integrated p-values of 17 MPAs, which are all newcomers around the poles of the fourth polarized human motif eigenvector.

| Motif Symbol | P value |
|---|---|
| SP1_03 | 6.42E-159 |
| IK_Q5 | 2.61E-07 |
| CACD_01 | 1.39E-88 |
| LYF1_01 | 1.43E-14 |
| GC_01 | 3.33E-114 |
| SP1_02 | 9.87E-141 |
| SP1_Q2_01 | 5.19E-202 |
| SP1_Q4_01 | 3.05E-194 |
| TAXCREB_01 | 9.13E-43 |
| CREBATF_Q6 | 1.45E-38 |
| TAXCREB_02 | 1.87E-14 |
| ATF6_01 | 3.46E-18 |
| E2F_Q2 | <6.42E-159 |
| PAX5_02 | 3.98E-48 |
| E2F1_Q3_01 | 1.65E-179 |
| PAX3_B | 3.27E-20 |
| ZF5_B | 2.54E-141 |

**Table S24 List of 47 genes with human specific Alu insertions in their proximal regulatory regions.** These human specific Alu insertions were obtained from a recent study (Tang, et al. 2018). The third column provides the locations of

Alu insertions from the TSSs. Negative and positive values respectively correspond to Alu insertion events upstream and downstream the TSSs. For each gene, the fourth column provides the difference between its rank in the fourth gene eigenvector of human and that of chimpanzee. A negative value of difference means the gene ranks higher in human; whereas a positive value means it ranks lower (ties could occur). The fifth column provides short descriptions of the gene functions.

| Gene | Alu | Location to TSS (bp) | Rank Difference | Gene Function |
|---|---|---|---|---|
| SPTY2D1 | AluYa5 | -584 | -8644.5 | CREB-modulated gene upon stimulus in the hippocampus (Lemberger, et al. 2008) |
| SEMA4F | AluYg6 | -737 | -7279.5 | axon guidance |
| BSG | AluSg7 | -819 | -5989 | neural network formation |
| GRK7 | AluYi6 | -565 | -5693.5 | visual perception |
| CDT1 | AluYe5 | -167 | -4274.5 | Meier-Gorlin syndrome 4 |
| CEBPG | AluYa5 | -427 | -4233 | key regulator of synaptic plasticity and memory formation (Alberini 2009) |
| VDAC3 | AluYh3 | -982 | -1035 | synaptic transmission; behavioral fear response; learning |
| TRIB3 | AluY | -213 | -504 | programmed neuronal cell death |
| CTHRC1 | AluY | -519 | -493 | cochlea morphogenesis |
| MKS1 | AluYa5 | -846 | -6766.5 | Meckel syndrome 1; head development |
| ATG7 | AluYa5 | -789 | -1661.5 | axonal homeostasis |
| ATXN7L3B | AluYa5 | -843 | -8285 | mutation causes neurodevelopmental delay and cerebellar ataxia |
| TMSB4Y | AluYa5 | -54 | -8062 | Sex-biased expressed in human brains (Reinius and Jazin 2009) |
| DHODH | AluYb8 | 420 | -152 | highly active in neocortex (Schaefer Ch, et al. 2010) |
| UBE2T | AluYa5 | 374 | -7794.5 | MEF2-regulated genes in human neural progenitor cell (Chan, et al. 2015) |
| LDAH | AluYb8 | -505 | -4090.5 | Loss of LDAH associated with hearing loss (Currall, et al. 2018) |
| INTS5 | AluYb8 | -623 | -11551 | component of INT complex that prevents dedifferentiation of intermediate neural progenitors (Zhang, et al. 2019) |
| DBNDD2 | AluYb8 | -629 | -1088 | highest expression level in C1 segment of cervical spinal cord |
| PDP2 | AluYa5 | 393 | -5774 | A biomarker of Parkinson's Disease (Fan and Xiao 2018) |
| TMEM181 | AluYb8 | -990 | -795.5 | enhanced RNA expression in brain (https://www.proteinatlas.org/ENSG00000146433-TMEM181/tissue) |
| CEP44 | AluYa5 | -596 | -1540 | highest expression level in corpus callosum |
| KRTAP9-1 | AluYg6 | -978 | -61 | Keratin-associated protein in the hair cortex |
| UGGT2 | AluYa5 | -447 | -8841.5 | ER-associated misfolded protein catabolic process |
| NKTR | AluYa5 | -909 | -5063 | NK-tumor recognition |
| MED31 | AluY | 494 | -1739.5 | Mediator of RNA polymerase II transcription |
| TSSK4 | AluYa5 | -100 | 2784 | positive regulation of CREB transcription factor activity that regulate memory development (Alberini 2009) |
| PUM3 | AluY | 364 | 4120.5 | eye development |
| FRS2 | AluYb8 | -862 | 1117.5 | forebrain development, axon guidance |
| SHCBP1L | AluYk11 | -708 | 223 | spermatogenesis |
| GGACT | AluYa5 | -293 | 329 | catabolic process |
| GTF2H3 | AluYa5 | -776 | 871.5 | component of TFIIH core complex |
| LEO1 | AluSx1 | -348 | 909.5 | RNA polymerase-associated protein |
| OSBPL11 | AluYa5 | -751 | 1033 | adipocytes differentiation |

| | | | | |
|---|---|---|---|---|
| CFAP45 | AluYh3 | 346 | 1084.5 | cilia- and flagella-associated protein 45 |
| RPL17 | AluYa5 | -975 | 1110 | component of the large ribosomal subunit |
| IDNK | AluYb8 | -677 | 1387.5 | cellular modified amino acid catabolic process |
| HOOK2 | AluYe5 | -250 | 1387.5 | endocytosis; centrosome establishment |
| RABEPK | AluYb8 | -444 | 1596.5 | trans-Golgi network |
| MGAT4C | AluY | -757 | 2280 | protein N-linked glycosylation |
| TBPL2 | AluSc8 | -134 | 3106 | myoblasts differentiation |
| PLCD3 | AluYa5 | -161 | | calcium signaling pathway (KEGG: hsa04020) |
| OCLM | AluYg6 | -283 | | visual perception |
| OR12D2 | AluYa8 | -437 | | expressed in hypothalamus |
| C17orf100 | AluY | -774 | | highest expression level in temporal lobe |
| WDFY4 | AluYb8 | -270 | | autophagy |
| OR5AC2 | AluYc | -943 | | odorant receptor |
| OR11H2 | AluYb8 | -29 | | olfactory receptor activity |

**Table S25 Details of 6 other genes used in the main text to explain the influence of Alu elements.** ADAM10 has a human specific AluJr4 insertion 700 bp upstream of its TSS, but not reported in Table S23. DRD3 has an existing Alu element downstream from the most upstream TSS in human. MYH14 has two Alu elements upstream of its most upstream TSS in human. The other three genes have human specific Alu insertions around their TSSs. The AluYa5 insertion in the case of GABRP does not show in the human reference genome, but was validated to be fixed in human by a sequencing data set (Hormozdiari, et al. 2013). The fifth column lists the appropriate references (PubMed ID) for their gene functions. The last column provides the UCSC browser URL links to these human specific Alu insertions.

| **Gene** | **Alu** | **Location to TSS (bp)** | **Gene Function** | **PubMed ID** | **UCSC session link** |
|---|---|---|---|---|---|
| ADAM10 | AluJr4 | -700 | neuronal plasticity<br>cochlea development | 23676497<br>30639848 | https://genome.ucsc.edu/s/liliang/hg38-ADAM10 |
| PAX2 | AluYa5 | -8278 | cochlea and brain development | 8943028 | https://genome.ucsc.edu/s/liliang/hg38-PAX2 |
| NRXN1 | AluYg6 | -1600 | vocal learning<br>social behavior | 18057082<br>20468056 | https://genome.ucsc.edu/s/liliang/hg38-NRXN1 |
| MYH14 | AluJo<br>AluYm1 | -830<br>-1250 | vocalization and hearing<br>axon guidance | 21480433 | https://genome.ucsc.edu/s/liliang/hg38-MYH14 |
| GABRP | AluYa5 | 1765 | GABA receptor subunit | 9182563 | Not available |
| DRD3 | AluJr | 500 | dopaminergic synaptic transmission, visual learning, social behavior | 29276054<br>16524466 | https://genome.ucsc.edu/s/liliang/hg38-DRD3 |

# UCSC browser snapshots of the human specific Alu insertions around ADAM10, DRD3 and PAX2

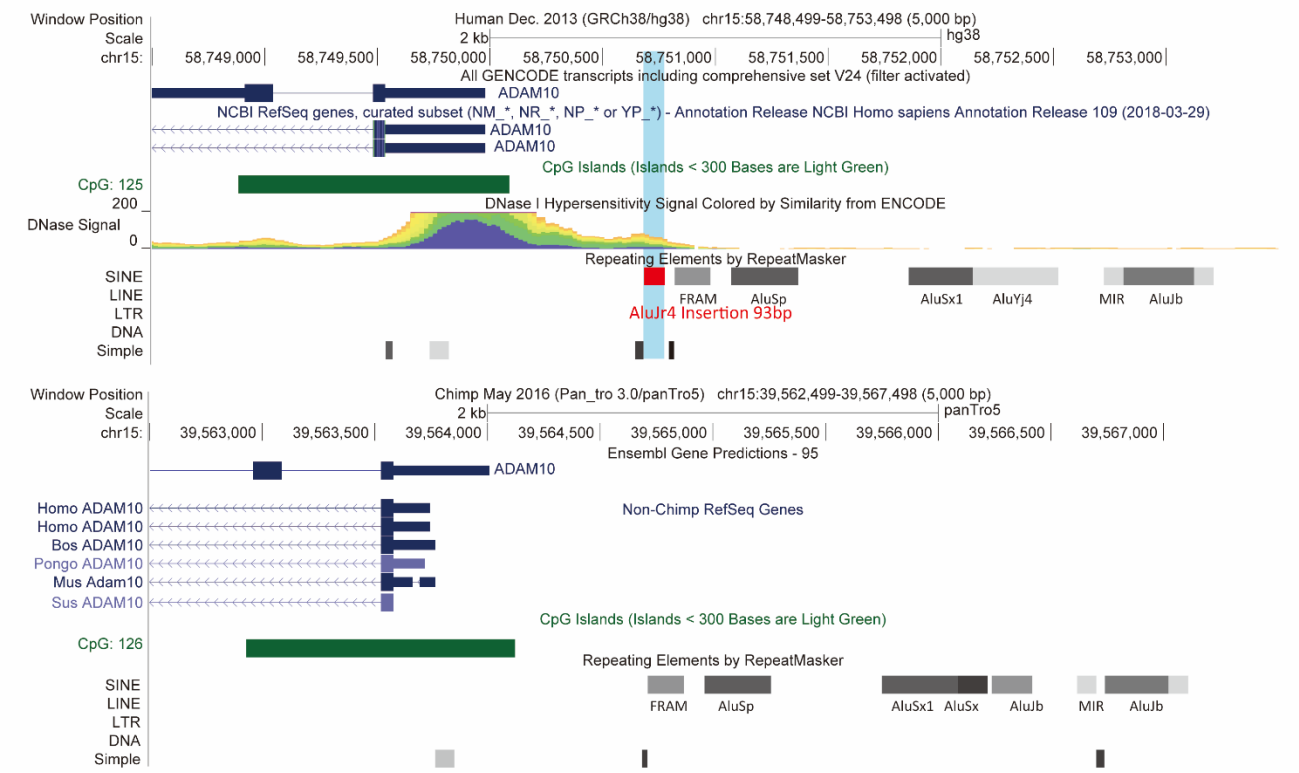



**Fig. S6 A human specific insertion event, AluJr4, in the proximal regulatory region of the gene ADAM10.**

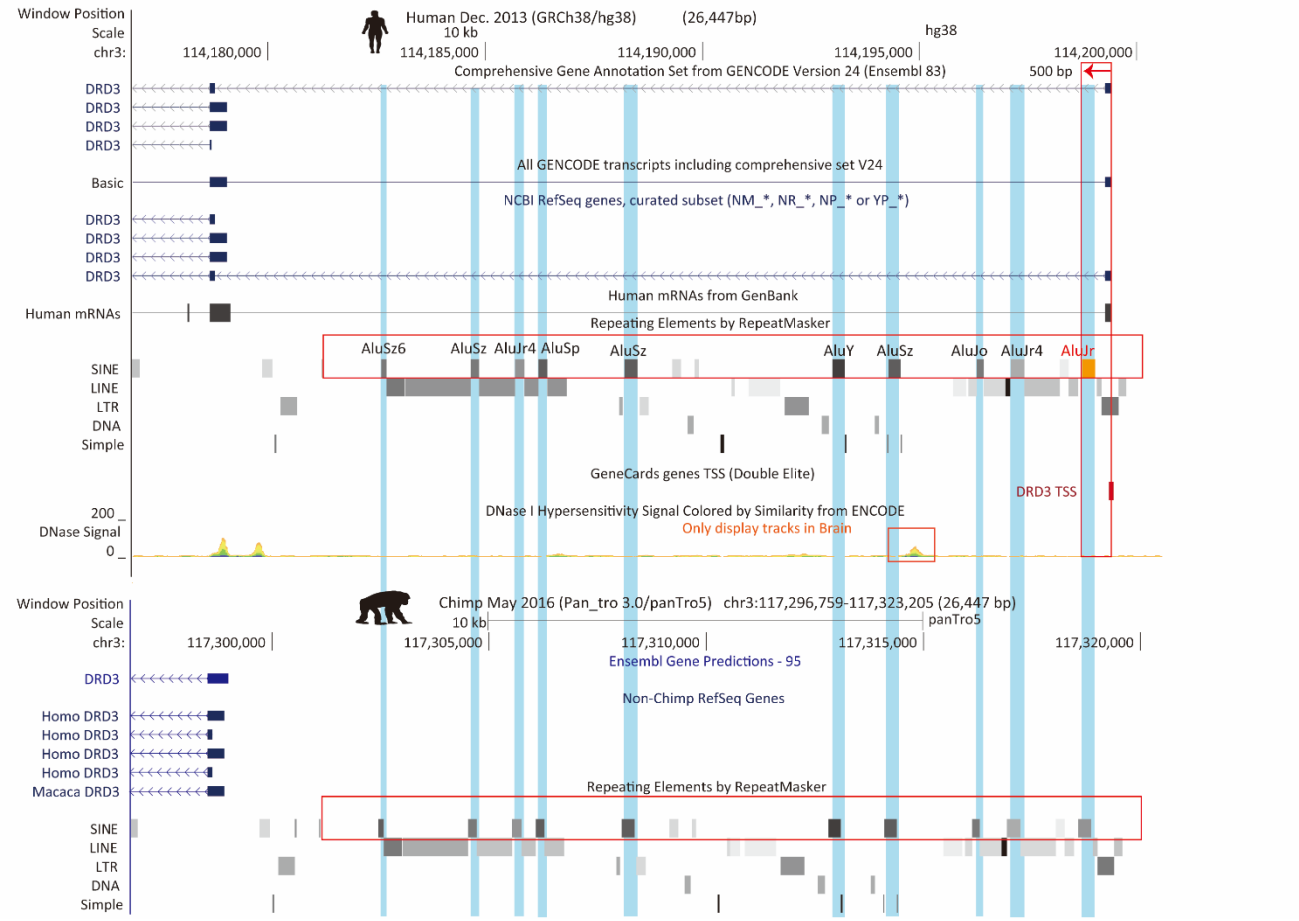



**Fig. S7. An AluJr4, located 500bp downstream of the TSS of a human specific DRD3 transcript.** Human has a specific transcript extending towards the 5' end so that it contains 10 existing Alu elements in its first intron. However, no such transcript is found in chimpanzee

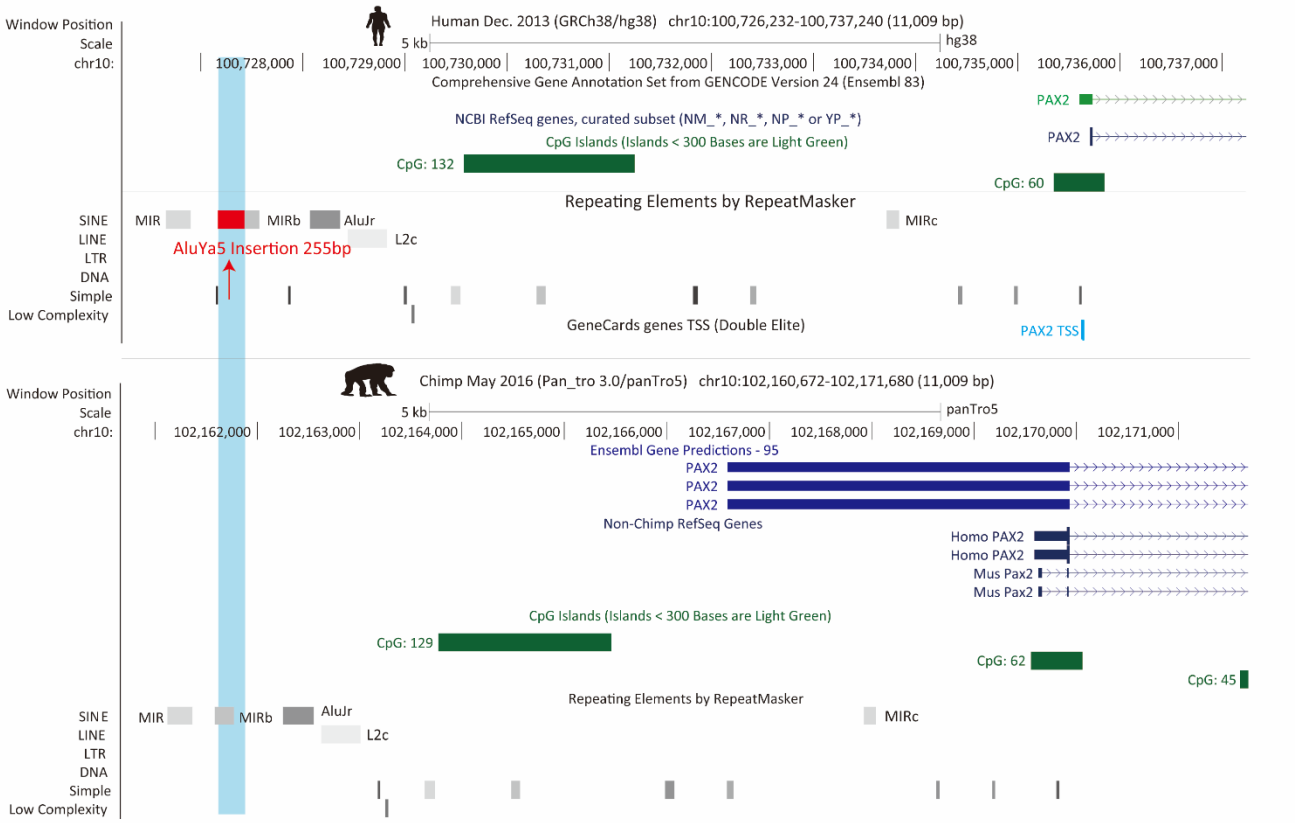



**Fig. S8 A human specific insertion event, AluYa5, found upstream of the TSS of PAX2.** The TSS in chimpanzee is predicted to be in the left CpG island, which is much closer to the Alu insertion.

## Dual eigen-analysis with the option of the APPRIS principal transcripts

The proximal regulatory regions for motif searching rely on the TSS selection. When multiple annotated transcripts are available, it is expected to select the principal transcript. We calculated the CREF modules for human and chimpanzee based on the principal transcript tagged as the main functional isoform by APPRIS. The module reorganization was observed too (Fig. S9), indicated by an approximate 52° rotation between the fourth and fifth eigenvectors. The enrichment analysis as well as the change in the MPA numbers are, by and large, similar to what are reported in the main text (Fig. S10). However, the APPRIS annotation is not available for orangutan.

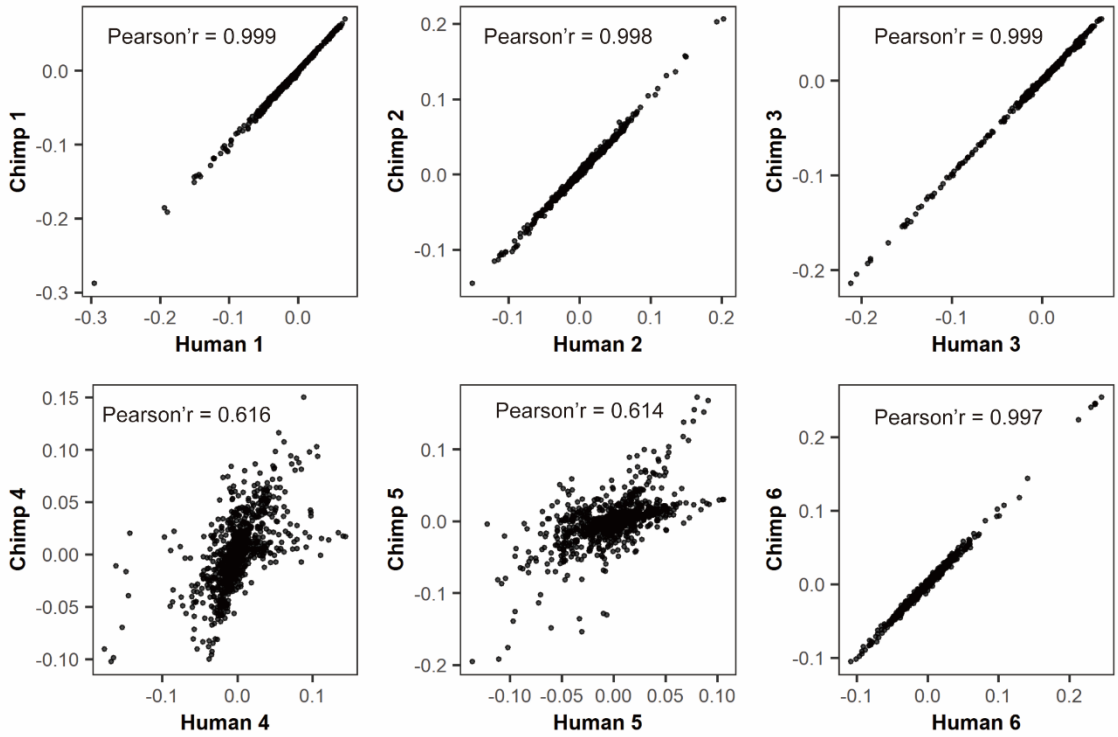



**Fig S9. The scatter plots of human's top six motif eigenvector loadings versus chimpanzee's using the APPRIS principal transcripts**.

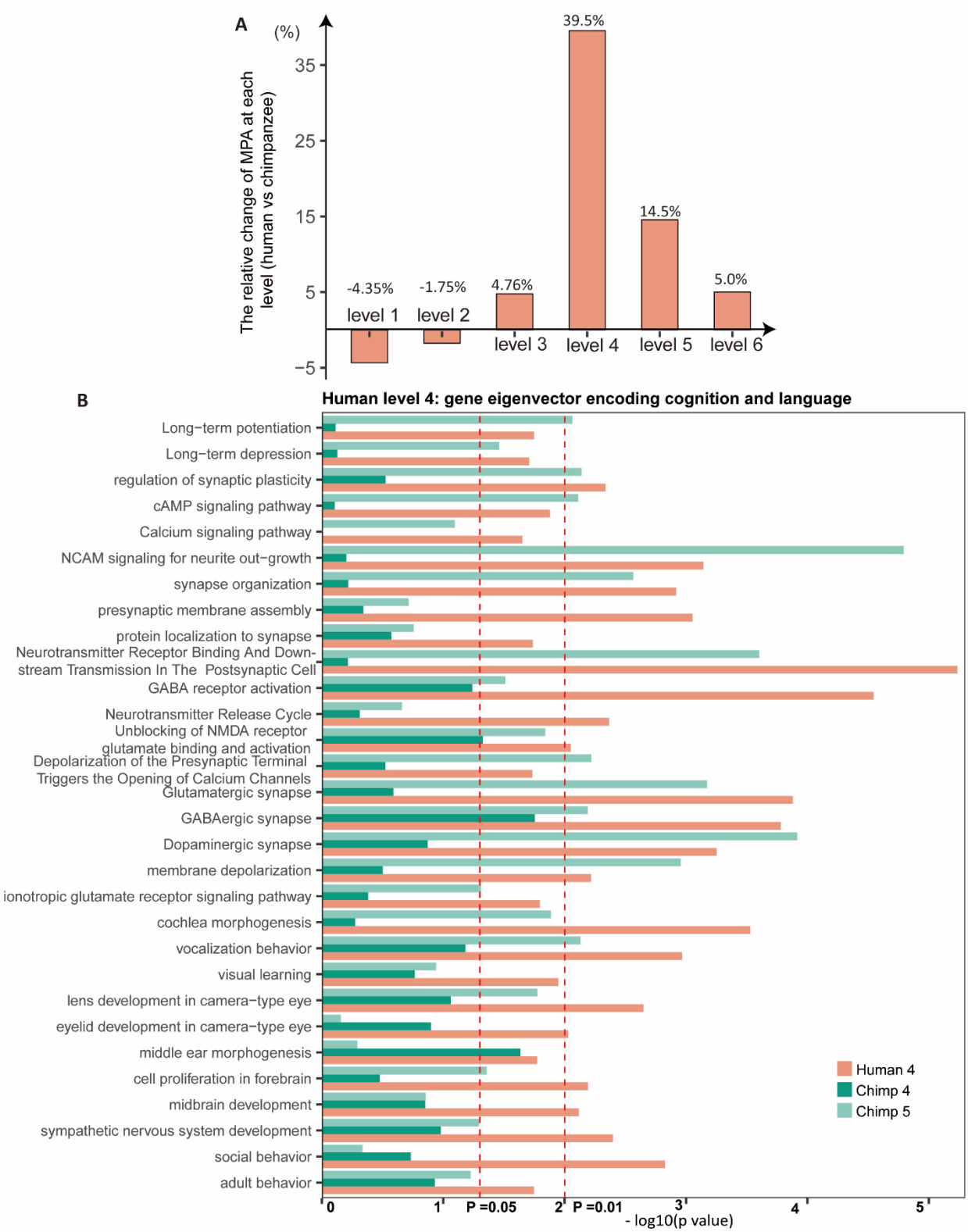



**Fig. S10 Comparative analysis of human and chimpanzee motif/gene eigenvectors.** The results suggest that our inferences on MPA changes underlying module reorganization, as well as our functional interpretations for the fourth CREF module, are robust with respect to TSS selection. **(A) The relative change of MPA in percentages at each level from chimpanzee to human by choosing the APPRIS principal transcripts.** The number of MPA increases most significantly at level 4 by 39.5%, followed by 14.5% at the fifth level. **(B) Comparison of enrichments results between human and chimpanzee by choosing the APPRIS principal transcripts.** Gene subsets that are significantly enriched near

the positive pole of the human fourth polarized gene eigenvector, but are not or less significantly enriched along the fourth chimpanzee polarized gene eigenvector. The color scheme is the same as that in Fig.6.

# Reference


Alberini CM. 2009. Transcription factors in long-term memory and synaptic plasticity. Physiol Rev 89:121-145.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25-29.

Barbacci E, Reber M, Ott MO, Breillat C, Huetz F, Cereghini S. 1999. Variant hepatocyte nuclear factor 1 is required for visceral endoderm specification. Development 126:4795-4805.

Cand EJ, Li X, Ma Y, Wright J. 2011. Robust principal component analysis? J. ACM 58:1-37.

Carbon S, Chan J, Kishore R, Lee R, Muller H-M, Raciti D, Van Auken K, Sternberg P. 2017. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res 45:D331-D338.

Cecconi F, Proetzel G, Alvarez-Bolado G, Jay D, Gruss P. 1997. Expression of Meis2, a Knotted-related murine homeobox gene, indicates a role in the differentiation of the forebrain and the somitic mesoderm. Developmental Dynamics 210:184-190.

Chan SF, Huang X, McKercher SR, Zaidi R, Okamoto SI, Nakanishi N, Lipton SA. 2015. Transcriptional profiling of MEF2-regulated genes in human neural progenitor cells derived from embryonic stem cells. Genom Data 3:24-27.

Cheng C, Fabrizio P, Ge H, Wei M, Longo VD, Li LM. 2007. Significant and Systematic Expression Differentiation in Long-Lived Yeast Strains. PLOS ONE 2:e1095.

Currall BB, Chen M, Sallari RC, Cotter M, Wong KE, Robertson NG, Penney KL, Lunardi A, Reschke M, Hickox AE, et al. 2018. Loss of LDAH associated with prostate cancer and hearing loss. Hum Mol Genet 27:4194-4203.

Elkouby YM, Elias S, Casey ES, Blythe SA, Tsabar N, Klein PS, Root H, Liu KJ, Frank D. 2010. Mesodermal Wnt signaling organizes the neural plate via Meis3. Development:dev.044750.

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al. 2018. The Reactome Pathway Knowledgebase. Nucleic Acids Res 46:D649-d655.

Fan Y, Xiao S. 2018. Progression Rate Associated Peripheral Blood Biomarkers of Parkinson's Disease. J Mol Neurosci 65:312-318.

Feng Y, Zhang S, Li L, Li LM. 2019. The cis-trans binding strength defined by motif frequencies facilitates statistical inference of transcriptional regulation. BMC Bioinformatics 20:201.

Fisher RA. 1948. Questions and Answers. The American Statistician 2:30-31.

Fisher RA. 1925. Statistical Methods for Research Workers. Oliver and Boyd (Edinburgh).

Golub GH, Loan CFV. 1996. Matrix computations (3rd ed.): Johns Hopkins University Press.

Holz A, Kollmus H, Ryge J, Niederkofler V, Dias J, Ericson J, Stoeckli ET, Kiehn O, Arnold HH. 2010. The transcription factors Nkx2.2 and Nkx2.9 play a novel role in floor plate development and commissural axon guidance. Development 137:4249-4260.

Horb ME, Thomsen GH. 1999. Tbx5 is essential for heart development. Development 126:1739-1751.

Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker JA, Nelson B, Alkan C, Sudmant PH, Huddleston J, et al. 2013. Rates and patterns of great ape retrotransposition. Proc Natl Acad Sci U S A 110:13457-13462.

Houtmeyers R, Souopgui J, Tejpar S, Arkell R. 2013. The ZIC gene family encodes multi-functional proteins essential for patterning and morphogenesis. Cell Mol Life Sci 70:3791-3811.

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353-d361.
Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27-30.
Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457-462.
Lemberger T, Parkitna JR, Chai M, Schutz G, Engblom D. 2008. CREB has a context-dependent role in activity-regulated transcription and maintains neuronal cholesterol homeostasis. FASEB J 22:2872-2879.
Li LM, Liu X, Wang L, Wang Y, Liu X, Tian X, Gong F, Shen L, Peng X-d. 2017. A Novel Dual Eigen-Analysis of Mouse Multi-Tissues' Expression Profiles Unveils New Perspectives into Type 2 Diabetes. Scientific Reports 7:5044.
Lin Z, Chen M, Ma Y. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055.
Machon O, Masek J, Machonova O, Krauss S, Kozmik Z. (Machon2015 co-authors). 2015. Meis2 is essential for cranial and cardiac neural crest development. BMC Developmental Biology 15:40.
Reinius B, Jazin E. 2009. Prenatal sex differences in the human brain. Mol Psychiatry 14:987, 988-989.
Schaefer Ch M, Schafer MK, Lofflerr M. 2010. Region-specific distribution of dihydroorotate dehydrogenase in the rat central nervous system points to pyrimidine de novo synthesis in neurons. Nucleosides Nucleotides Nucleic Acids 29:476-481.
Sheng G, dos Reis M, Stern CD. 2003. Churchill, a Zinc Finger Transcriptional Activator, Regulates the Transition between Gastrulation and Neurulation. Cell 115:603-613.
Tang W, Mun S, Joshi A, Han K, Liang P. 2018. Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. DNA Res 25:521-533.
The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506-D515.
Zhang W, Behringer RR, Olson EN. 1995. Inactivation of the myogenic bHLH gene MRF4 results in up-regulation of myogenin and rib anomalies. Genes Dev 9:1388-1399.
Zhang Y, Koe CT, Tan YS, Ho J, Tan P, Yu F, Sung WK, Wang H. 2019. The Integrator Complex Prevents Dedifferentiation of Intermediate Neural Progenitors back into Neural Stem Cells. Cell Rep 27:987-996 e983.