Tissue-specific expression of young small-scale duplications in human central nervous system regions

Solène Brohard-Julien,[*,1,2,3] Vincent Frouin,[2] Vincent Meyer,[1] Smahane Chalabi,[1] Jean-François Deleuze,[1] Edith Le Floch[*,1,4], and Christophe Battail[*,1,4]

[1]Centre National de Recherche en Génomique Humaine (CNRGH), Institut François Jacob, CEA, Evry, France

[2]UNATI, Neurospin, Institut Joliot, CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette, France

[3]Université Paris-Sud, Université Paris-Saclay, Orsay, France

[4] These authors contributed equally to this work.

*Corresponding authors: brohard@cng.fr; lefloch@cng.fr; christophe.battail@cea.fr

Key words

paralog, small-scale duplication, tissue-specific, human, central nervous system, co-expression network, RNA-seq, brain disease, whole-genome duplication

Abstract

Gene duplication, a mechanism that has been occurring in the genome throughout evolution and has led to refined patterns of spatial expression and biological functions, is a major contributor to tissue complexity in humans. The accumulating evidence, from inter-species and inter-organ comparisons, points to the relevance of exploring tissue expression through the lens of duplication type and date. In this study, we use the transcriptional profiles of human central nervous system (CNS) tissues to identify evolutionary features associated with the spatial expression of paralogs and gene families. We show that paralogs, especially those originating from small-scale duplications younger (ySSD) than whole-genome duplication (WGD), are significantly implicated in tissue-specific expression in CNS territories. Interestingly, we observe that both the young age of the ySSD and the duplication type are associated with their tissue-specificity. The exploration of these paralog properties at the family level of organization shows that the families composed of a majority of genes co-expressed across CNS tissues are enriched in tandem duplications, ySSDs and tissue-specific paralogs. Moreover, these families are strongly enriched in tissue-specific families, suggesting that co-expression analysis is able to capture shared tissue-specificity, especially of ySSDs, within paralog families. Overall, our study provides new evidence for the major involvement of ySSDs in the differentiation of CNS territories, that extend ySSDs properties already established from comparative expression analyses across organs and species.

Introduction

The human central nervous system (CNS), and especially the brain, is made up of billions of neuronal

and glial cells from hundreds of different subtypes. Its complexity can, however, be explored by the underlying patterns of spatial and temporal gene expression which reveal anatomical regions, biologically active processes and brain function. Over the past 10 years, transcriptional profiling of neuroanatomic regions has been performed to better understand the molecular complexity of this organ. One of the first studies used expression data generated from gene microarray profiling of hundreds of human postmortem samples to estimate the cellular composition of brain tissues (Oldham et al. 2008). Another study based on exon-level microarray analysis of more than 50 postmortem human brains identified sex-biased expression patterns and compared spatial expression patterns between developmental and adult brains (Kang et al. 2011). Molecular profiling of the brain has also been advanced by the constitution of the Allen Human Brain Atlas, a multi-modal atlas that integrates magnetic resonance imaging with histological and gene microarray expression data across neuroanatomical regions (Hawrylycz et al. 2012). Finally, in a recent landmark contribution, the GTEx (Genotype-Tissue Expression) consortium used RNA sequencing technology to establish the landscape of human gene expression across a large collection of postmortem biopsies from hundreds of normal individuals, and particularly for 13 brain-related tissues (Ardlie et al. 2015).

Gene duplication is considered as one of the major contributors to tissue expression and has resulted in paralogs, which are defined as homologous genes within the same species. The retention of paralogs in the genome is associated with the divergence of spatial expression and molecular function (Stephens 1951; Ohno et al. 1968; Ohno 1970; McLysaght et al. 2002; Dehal and Boore 2005; Dixon et al. 2012). Around 60% of human genes are considered to be paralogs while remainder are considered as singletons (Chen et al. 2013). Paralogous genes originate from two main types of molecular evolutionary events, namely whole-genome (WGD) and small-scale duplications (SSD), with the latter predominantly involving a single gene locus (Hakes et al. 2007). Two rounds of WGD events are known to have occurred in early vertebrate lineage some 500 million years ago (McLysaght et al. 2002; Nakatani et al. 2007) while SSD events have taken throughout evolutionary history. Retained duplicates define the paralogs in current day genomes.

Several mechanistic events have been proposed to explain the retention of paralogs in the genome, depending on the type of duplication. First, this retention can be favored by dosage balance of protein abundance (Zhang 2003; Lan and Pritchard 2016). Indeed, once genes were duplicated by WGD, whole biological processes are impacted by a uniform increase in protein abundance. Consequently, among groups of genes implicated in the same processes, the subsequent removal of individual genes after a WGD event would not be favored as it would disrupt the balance between these genes (Papp et al. 2003). This property does not hold for genes duplicated by SSD as the duplication of a single gene would naturally disturb the biological processes in which it is involved (Papp et al. 2003; Innan and Kondrashov 2010). Therefore, for SSD pairs, dosage balance can consist of the sharing of ancestral protein abundance by down-regulation of one or both duplicates (Li et al. 2005; Ganko et al. 2007; Qian

et al. 2010). Afterwards, a divergence process can occur and lead to long-term SSD retention, resulting in two main outcomes: (i) The sub-functionalization process which corresponds to the distribution between the two gene copies of either the ancestral gene functions or spatial expression (Prince and Pickett 2002; Assis and Bachtrog 2015), and (ii) The neo-functionalization process which leads to the accumulation of mutations in one of the copies that affect its ancestral function, resulting in the gain of a new function, or extending its spatial expression into a new territory not targeted by the ancestral gene (Stephens 1951; Force et al. 1999; Teshima and Innan 2008; Innan and Kondrashov 2010).

The hypothesis of paralog retention in the genome, favored by divergence of spatial expression, is related to the concept of gene tissue-specificity, which indicates whether a gene has a broad or narrow expression pattern across a collection of tissues (Zhang 2003; Lan and Pritchard 2016). Following an SSD event, often the duplicate with the lower expression tends to rapidly become more tissue-specific than the ancestral gene. Moreover, when both paralogs in an SSD pair are tissue-specific, they tend to be expressed in the same tissue. On the contrary, when both paralogs of a WGD pair are tissue-specific, they are often expressed in different tissues (Kryuchkova-Mostacci and Robinson-Rechavi 2016). Previous studies on co-expression of paralog pairs have shown that SSD pairs, and especially those younger than WGD, showed a higher level of co-expression than WGD pairs (Acharya and Ghosh 2016; Lan and Pritchard 2016). Moreover, this higher level of co-expression seems to be related to the fact that these WGD-younger SSD pairs are more frequently composed of tandem duplicates, suggesting that they may thus share common genetic regulatory sequences (Lan and Pritchard 2016). These previous results are all derived from inter-species or inter-organ expression analyses, and have not yet been validated using gene expression data restricted to related tissue, such as those constituting the CNS.

Beyond the classification of paralogs into duplication type or age categories, another level of organization worth investigating is the gene family. Gene families, which correspond to sets of genes generated by duplications from a common ancestor, can originate from a WGD or an SSD event, or from a mixture of both. Furthermore, the identification of duplicated genes that emerged following the human-chimpanzee divergence has attracted important attention because of their potential contribution to human-specific traits, and in particular to brain complexity (Sudmant et al. 2010; Dennis and Eichler 2016; Dennis et al. 2017). Notably, the study of the interplay between several genes of the *SRGAP2* (Slit-Robo Rho GTPase-activating protein 2) and *ARHGAP11* (Rho-type GTPase-activating protein 11) families has revealed a functional association of the human-specific duplicated genes *SRGAP2C* and *ARHGAP11B* with an increase in both dendritic spine density and cortical neuron population (Charrier et al. 2012; Florio et al. 2015). A powerful way to identify expression patterns associated with groups of genes is to study gene co-expression networks (Zhang and Horvath 2005; Parikshak et al. 2015). These approaches, which consist of finding groups or modules of genes with highly correlated expression values across samples, are effective ways to explore gene associations

with tissue differentiation. To date, cross-analysis between modules co-expressed across CNS regions and their organization into gene families has not been done.

A better comprehension of the biology of paralogs will also support our understanding of diseases. Recent evidence suggests that the proportion of essential genes is higher in paralogs than in singletons (Acharya et al. 2015). Furthermore, when the origin of duplication is considered, essential genes are more frequent in WGD than in SSD paralogs (Makino et al. 2009; Acharya and Ghosh 2016). Disease-associated genes (i.e. genes whose sequences or expression are often altered in disease states), have been found to be over-represented in paralogs compared to singletons (Makino and McLysaght 2010; Dickerson and Robertson 2012; Chen et al. 2013; Singh et al. 2014). Whether considering all or only Mendelian diseases, disease-associated genes tend to be enriched in paralogs from WGD (Chen et al. 2014; Singh et al. 2014; Acharya et al. 2015). As for SSD genes, it has been shown that there is an enrichment in WGD-older SSD and a depletion in WGD-younger SSD in disease-associated genes compared to non-disease genes (Chen et al. 2014). These associations between the evolutionary characteristics of paralogs and disease-related genes were carried out with no restrictions on the human pathologies considered. Therefore, it is relevant to investigate whether these paralog properties are also true if we consider only the population of genes related to brain diseases.

In the present study, we explored in detail the expression patterns of paralogs in CNS tissues, using the GTEx resource, associated with their evolutionary characteristics and gene families. Firstly, we assessed whether duplicated genes were significantly associated with differences in expression between CNS tissues. Secondly, we investigated the association of paralogs with the tissue-specificity and the distribution of tissue-specific genes across CNS territories. Thirdly, we studied the association between the tissue-specific expression of paralogs and their evolutionary characteristics. More precisely we investigated whether the associations previously observed between tissue-specificity and duplication types and ages, from studies comparing species or organs, were confirmed by the expression analysis on human CNS tissues alone. Fourthly, we studied the organization of paralogs in families using co-expression to define co-expressed gene families and to analyse their tissue-specificity and evolutionary characteristics. Finally, we explored the association of paralogs and their evolutionary characteristics with brain diseases.

Results

1/ Association of paralog expression with CNS differentiation

We considered all human protein coding genes and the information collected on duplication events in order to split the gene population into paralogs and singletons (Chen et al. 2013) (see Materials and Methods). Gene expression data for hundreds of individuals from 13 normal brain-related tissues (see Materials and Methods) were obtained from the GTEx consortium (Ardlie et al. 2015). After filtering out low information content genes, abundance values of 16,427 protein-coding genes, including 10,335

paralogs and 6,092 singletons were conserved. Previous work by GTEx established the relevance of using gene expression data to cluster samples obtained from the same tissues, even though assigning samples to the correct CNS region was more difficult than for other organs (Ardlie et al. 2015). We extended this analysis by focusing specifically on CNS tissues and assessing whether paralog expression could better classify samples into tissues than singletons or all protein-coding genes. Our unsupervised hierarchical classification of human CNS samples, based on their pair-wise similarity in terms of correlation across gene expression values, was able to group together most samples belonging to the same tissue (see Materials and Methods, fig. 1). The choice of color gradients for tissues that anatomically overlap confirmed the ability of gene expression profiles to classify these tissues into neurologically relevant groups. Therefore we pooled together some of the 13 initial tissues that showed similar expression profiles in order to define a short list of 7 CNS regions (see Materials and Methods). The relevance of our experimental classification was evaluated according to the expected groups of brain-related tissues using the adjusted rand index (ARI) (Hubert and Arabie 1985). We observed that globally, the sample classification based on paralog expression (ARI = 0.197) was slightly better than the classification obtained using all protein-coding genes (ARI = 0.175) or singletons (ARI= 0.182). It should be noted that the quality of a clustering is likely to be influenced by the number of genes used in the analysis. Therefore, the better ARI score obtained with the paralogs compared to singletons could be partly due to the higher number of paralogs in relation to singletons. However, we also obtained a greater ARI with the paralogs in comparison to the ARI calculated from all protein-coding genes, thus suggesting a particular biological relationship between paralogs expression and CNS tissue differentiation.

In addition to this clustering analysis, we carried out another assessment by performing differential expression analysis of gene count data between all pairs of CNS tissues (see Materials and Methods). We obtained a list of significantly differentially expressed genes (DEGs) for each pair of tissues (see Supplementary Materials, table S3). By comparing the relative proportion of DEGs in paralogs and singletons, we observed that DEGs were significantly enriched in paralogs for 75 out of the 78 tissue-pairs tested (Chi-squared test, and threshold p-value = 6.41E-04 with Bonferroni correction to account for the number of tissue pairs). Furthermore, in order to assess the potential bias of expression level in these results, we calculated the overall expression of paralogs averaged over brain-related tissues and found it to be significantly lower than that of singletons (12 versus 37 RPKM respectively, t-test, p-value=5E-16). This observation, which implies less power in the DE tests for the former group, makes the enrichment of the DEGs in paralogs even more reliable.

Overall, these complementary analyses on tissue clustering and differential expression illustrate the strong biological contribution of paralogous genes to expression differences between CNS territories.

2/ Tissue-specific expression of paralogs in CNS regions

We further investigated these expression differences of paralogs between CNS territories by looking at their tissue-specificity. The detection of tissue-specific genes was performed using expression profiles quantified across the 7 CNS regions previously defined. From the collection of methods developed to measure tissue specificity, we selected the method based on Tau score because of its high sensitivity to detect tissue-specific genes (Yanai et al. 2005; Kryuchkova-Mostacci and Robinson-Rechavi 2017). The Tau score ranges from 0 for broadly expressed genes, to 1 for highly tissue-specific genes (see Materials and Methods). Contrary to Tau score distributions reported in a previous study on different organs (Kryuchkova-Mostacci and Robinson-Rechavi 2017), the distribution of Tau scores in the CNS regions in the present study was not bi-modal and had a unique mode centered on low values (fig. 2A). Consequently, the Tau threshold for declaring a gene tissue-specific could not be visually defined. We thus developed an approach based on permutations to adapt this threshold choice to the case of similar tissues within a single organ system. We calculated an empirical p-value for each gene, based on permutations of the tissue labels, and then performed a False Discovery Rate (FDR) correction on the p-values for the multiple genes tested (Benjamini-Hochberg corrected p-value < 0.01) (fig. 2A). Applying this approach, we selected a Tau threshold of 0.525 and found that 17% (2,829) of protein-coding genes expressed in the CNS regions were tissue-specific (see Supplementary Materials, fig. S1). Moreover we established that paralogs were significantly enriched in tissue-specific genes compared to singletons (19.2% of paralogs were tissue-specific, versus 13.9% of singletons, p-value = 2.045E-18, using a Chi-squared test) (table 1).

The overall expression of paralogs averaged over CNS tissues being significantly lower than that of singletons, the association of duplication status with tissue-specificity could be influenced by the expression level. In order to confirm the impact of paralogy on the tissue-specificity of genes in CNS regions, in addition to the potential effect of expression level, we used a multivariate linear regression (see Materials and Methods). The model was fitted on the Tau scores of all protein coding genes with explanatory variables corresponding to the maximal expression of each gene across the 7 CNS regions and its duplication status (table 2a). We observed that the maximal expression had a significant negative effect on the Tau score, while the duplication status had a smaller but significant positive effect. We also assessed that the use of the average expression instead of the maximal expression led to similar results (see Supplementary materials, result S1 and table S16a). These results indicate that paralogous genes are associated with tissue-specific properties, which may therefore contribute to the differences in their expression between CNS territories.

Although this method based on the Tau score can identify tissue-specific genes, it does not indicate which CNS region is targeted by this specificity (Yanai et al. 2005). In order to study the regional distribution of tissue-specific genes, we mapped each tissue specific gene to one CNS region (see Supplementary Materials, table S4). Therefore, for each tissue-specific gene, we considered the anatomical region associated with the highest expression value to be the specific region (fig. 2B). We

discovered that the distribution of tissue-specific genes across CNS regions was very heterogeneous (see Supplementary Materials, table S6) compared to an almost constant proportion of expressed genes across these regions (see Supplementary Materials, table S5). The highest proportions of tissue-specific genes were found in the cerebellum (40.2%), spinal cord (20.9%) and hypothalamus (16.4%). The remaining tissue-specific genes (22.5%) were scattered over the last four brain-related regions. The distribution of tissue-specific paralogs across CNS territories was also highly heterogeneous and similar to the distribution obtained for all tissue-specific protein-coding genes (see Supplementary Materials, table S6).

In summary, we found that paralogs were more tissue-specific than other genes and that tissue-specific paralogs were concentrated in a limited number of CNS regions similarly to the other tissue-specific genes. Then, we confirmed on CNS tissues that the low expression level was the main factor explaining the tissue-specificity of genes. However, we also observed that the paralogous status contributed to the tissue-specific property in addition of the expression value.

3/ Evolutionary and genomic properties of tissue-specific paralogs

The date of a SSD can be estimated in relation to the WGD events and attributed to one of the three duplication age categories: younger SSD (after WGD events - ySSD), older SSD (before WGD events- oSSD) and WGD-old SSD (around WGD events – wSSD) (see Materials and Methods) (Singh et al. 2014). Using our collection of paralogs with tissue-specific expression in the CNS regions , we performed statistical tests to determine if they were enriched in particular duplication events (WGD or SSD) or dates of SSD (oSSD, wSSD and ySSD categories). Genes can undergo both WGD and SSD duplication and can sometimes be retained after each duplication. Unless otherwise stated, when we refer to a duplication type from this point on in the paper, we are referring to genes that have been retained after this duplication type only (WGD or SSD), in order to make a clear distinction between the effects of the two duplication types. Of the 10,335 paralogs considered in our study, 5,114 are from WGD, 3,719 from SSD (1,192 from ySSD, 1,260 from wSSD and 1,267 from oSSD) and 1,502 unclassified (966 both WGD-SSD and 536 without annotation).

We first observed that, among paralogs, SSD genes were significantly enriched in tissue-specific genes (22.6% of SSDs were tissue-specific versus 17.3% of the other paralogs, p-value = 9.022E-11), while on the opposite WGDs are depleted in tissue-specific genes (table 1). However, we noticed that WGDs seemed slightly enriched in tissue-specific genes, compared to singletons (15.7% of WGDs were tissue-specific versus 14.4% of the singletons, p-value = 4.1E-02). Furthermore, when we performed the same analysis only on the WGD-old paralogs (WGDs and wSSDs), the WGD genes were still significantly depleted in tissue-specific genes (15.7% of WGDs were tissue-specific versus 24% of wSSDs, p-value = 5.185E-12) (table 1). These tests allowed us to conclude that SSD paralogs were enriched in tissue-specific genes, independently of the potential effect of the duplication date on tissue-specificity.

In addition to assessing the duplication type, we also tested the association between categorical duplication age and tissue-specificity, and found that ySSD were also enriched in tissue-specific paralogs (28.6% of ySSDs versus 18.0% of the remaining paralogs, p-value = 6.341E-18). Moreover, ySSDs were still enriched in tissue-specific paralogs when we performed the analysis on only the SSD paralogs (28.6% of ySSDs versus 19.8% of the remaining SSDs, p-value = 3.483E-09). On the other hand, oSSDs were depleted in tissue-specific genes compared to other SSD gene paralogs (15.6% of oSSDs versus 26.2% of the remaining SSDs, p-value = 2.729E-13) (table 1). In summary, we could conclude that ySSD genes were more tissue-specific than other paralogs, probably due to both their SSD origin and their duplication age.

We then estimated the respective effects of the expression level (the maximal expression over CNS regions), the phyletic age of the duplication and the type of duplication (SSD or WGD) on Tau scores using multivariate linear regression models (see Materials and Methods and table 2b). We observed that the maximal expression had the stronger significant negative effect (Supplementary Materials, result S1), while the duplication age had a weaker but still significant negative effect on tissue-specificity (the same results were obtained with the age categories, see Supplementary Materials, table S16b). The SSD effect did not appear, probably because the age variability of SSDs is much higher than that of WGDs, which may hide the SSD effect. We tested this last hypothesis by applying the multivariate linear model on the subset of genes with the the same age (WGD and wSSD) (table 2c). We found with this last model that indeed the SSD duplication type significantly contributed to a high tissue-specificity. Therefore, these results seem to confirm that ySSD genes are more tissue-specific than other genes, due to both their SSD origin and their duplication age.

To refine again the association between duplication age and tissue-specificity, we performed enrichment analyses using a short list of paralogs that came from human-specific duplication events (see Materials and Methods) (Dennis et al. 2017) and found no significant associations (see Supplementary Materials, table S19). To obtain a complementary view of this tissue-specificity loss for very recent duplications, we examined the distribution of the Tau scores of paralogs according to their phyletic age (see Supplementary Materials, fig. S4). We found that the maximum Tau scores were obtained for genes with phyletic ages around 0.12 which corresponds in most cases to ySSD duplication events that occurred around the separation of the Simians clade (Ensembl Compara GRCh37 p.13). This result seems to indicate that tissue-specific expression is not a property particularly associated with human-specific duplications, even though it seems to increase for slightly older ySSDs and to decrease afterwards.

In summary, we found that SSD genes and in particular ySSD genes were more often tissue-specific than other paralogs due to their duplication origin and to the age of ySSD genes. We also confirmed, using a multivariate analysis, that both the duplication age and the SSD duplication type contributed to their tissue-specificity in addition to the expression level effect. Finally we showed that this higher

tissue-specificity of ySSDs did not seem to come from human-specific ySSDs but from slightly older ySSDs.

4/ Tissue-specificity analysis of co-expressed gene families

We demonstrated in our above results that paralogs, and especially SSDs and ySSDs, were involved in territorial expression in the CNS, notably through tissue-specificity. We then tried to determine if the paralogs within gene families tended to share the same tissue-specificity across the CNS. We studied the potential expression similarity between paralogs across CNS regions by using a co-expression analysis without a priori knowledge on their tissue-specificity.

The study of co-expression allowed us to explore the higher level of organization of the paralogs into groups of genes with coordinated expression across CNS tissues and compare these modules of co-expressed paralogs across tissues against annotated gene families. The Weighted Gene Correlation Network Analysis (WGCNA) methodology (Langfelder and Horvath 2008) was used to infer the correlation-based co-expression network. Contrary to previous studies that inferred a network per tissue and then compared modules between networks (Oldham et al. 2008; Pierson et al. 2015), in this study we carried out co-expression network inference by simultaneously using all CNS tissue samples profiled by the GTEx consortium in order to explore gene associations with tissue differentiation. We optimized the WGCNA to generate highly correlated co-expression modules of small size in order to compare them with the annotated gene families (Supplementary Materials, fig. S2) (see Materials and Methods). Indeed, out of our 3,487 gene families, 1,644 (47%) were constituted of only two genes. Our WGCNA analysis extracted 932 modules of co-expressed paralogous genes. Only 104 genes were not included in a co-expression module. The module size ranged from 2 to 911 genes with 84% of small size modules (modules with less than 10 genes) (see Supplementary Materials, table S7). A high proportion of modules were enriched in molecular function and biological process GO terms indicating that our network inference approach captured shared biological functions among co-expressed paralogs (see Supplementary materials, result S4).

To first check the relationship between co-expression and shared tissue-specificity, we analysed the distribution of tissue-specific genes across the 932 modules of co-expressed paralogs and found that 177 modules included at least two tissue-specific genes. We then looked at whether within each of these modules the tissue-specific genes were expressed in the same or in different regions. We found that among these 177 modules, 66% and 92% consisted of tissue-specific genes associated respectively with the same region or at most two different regions (see Supplemental Materials, table S15). Therefore, gene modules identified from correlation-based co-expression networks also capture shared tissue-specificity.

We then explored further the collection of co-expressed paralog modules by asking whether paralogs were more co-expressed across tissues when they came from the same family and whether it could be

associated with some shared tissue-specificity. We initiated this analysis by classifying the gene families into two categories, homogeneous and heterogeneous families, based on their patterns of expression across CNS tissues (see Materials and Methods). A homogeneous family was defined by the property that the majority of its member genes were included in the same co-expression module. Out of the 3,487 gene families considered in this study, we identified 111 homogeneous families (with 257 co-expressed paralogs out of a total of 300 expressed paralogs in these families, the remaining 43 paralogs being removed from all tests on homogeneous family genes in the rest of article) and thus 3,376 heterogeneous families (10,035 paralogs) (see Supplementary Materials, tables S13 and S14). We showed by a permutation approach that this number of homogeneous families was significantly large, with an empirical p-value inferior to 10^{-3} (see Materials and Methods), suggesting that paralogs were more co-expressed across tissues when they came from the same family. The comparison of the average size of families between each category showed that homogeneous families were significantly smaller than heterogeneous ones (Welch statistical test, average size of homogeneous families = 2.89, average size of heterogeneous families = 3.84, p-value = 8.278E-10). A total of 53 of these homogeneous families were completely included in the same module of co-expression. Furthermore, some modules were found to comprise several homogeneous gene families (see Supplementary Materials, table S9). A biological pathway enrichment analysis of the homogeneous family genes revealed that they were notably enriched in transcription factors and signaling proteins involved in neural development (see Supplementary Materials, result S6 and table S10)

Before looking at shared tissue-specificity within homogeneous families, we investigated the association of tissue-specificity with these co-expressed families, and observed a significant enrichment of tissue-specific paralogs in genes coming from homogeneous families (4.7% of tissue-specific paralogs versus 2% of the other paralogs, p-value = 5.374E-12) (table 3). We then investigated the link between shared tissue-specificity and homogeneous gene families by categorizing families according to their tissue-specificity following the classification defined by Guschanski et al, 2017 (Guschanski et al. 2017). Families composed of a majority of genes tissue-specific to the same regions were classified as tissue-specific families. We identified 58 tissue-specific families and we found a significant enrichment of tissue-specific families in homogeneous families (45% of tissue-specific families versus 2.5% of other families, p-value = 1.691E-69) (table 3). We then studied whether homogeneous families were associated with a type of duplication event or with a duplication age. We found that SSD and ySSD genes were both enriched in genes coming from homogeneous families (3.3% of SSD versus 2.1% of the other paralogs, p-value= 2.777E-04; 5.2% of ySSD versus 2.1% of the other paralogs, p-value = 5.758E-10) (table 3). We also found a significant enrichment of human-specific genes in homogeneous families, using ySSD genes as reference group, suggesting that the recent ySSDs tend to be more co-expressed than the other ySSDs (p-value = 3.868E-04, OR = 19.58) (table 3 and Supplementary Materials, result S7). Similarly, SSD and ySSD genes were significantly enriched in genes

coming from tissue-specific families (see Supplementary Materials, table S17). Finally, we also analysed the shared tissue-specificity of SSDs and ySSDs at the pair level but the very low number of tissue-specific paralog pairs did not allow to get significant results (see Supplementary Materials, result S2).

It can be expected that co-expression between two duplicates in a paralog pair will be associated with their proximity on the genome, as epigenetic co-regulation of gene expression partly depends on the proximity between genes on the genome (Xie et al. 2016; Ibn-Salem et al. 2017; Lian et al. 2018). We thus investigated whether the genomic distance between paralog pairs (Supplementary materials, result S5) could be used to differentiate homogeneous from heterogeneous families. For homogeneous families, we considered only pairs in which both paralogs belonged to the main co-expression module (37 pairs), and removed the other pairs from the test. We found that homogeneous families were depleted in inter-chromosomal pairs (70.3% of homogeneous families versus 90.2% of heterogeneous families were spread across different chromosomes, p-value= 7.73E-04) and were enriched in tandem duplicated pairs (27% of homogeneous families and 6.7% of heterogeneous families were separated by less than 1 Mb, p-value = 1.743E-04) (see Supplementary Materials, table S21); this supports the idea that paralog co-expression is favored by proximity along the genome. Moreover, we showed that the genomic proximity of duplicates was associated with recent SSDs. Indeed, we found that the younger the SSD pair, the more the duplicate were found in tandem in the genome (see Supplementary Materials, result S5). These results explain why SSDs, and especially ySSDs, tend to be more co-expressed and to share more often the same tissue-specificity within their family than other paralogs.

In summary, the gene co-expression network analysis performed on the CNS tissues allowed us to find that when several tissue-specific genes were clustered in the same module of co-expression, they were often expressed in the same region or the same pair of regions. We showed that within gene families, the shared tissue-specificity of paralogs was associated with their co-expression across tissues and we classified gene families into two categories according to co-expression status. Heterogeneous families tended to be associated with WGD genes and enriched in paralogs resulting from inter-chromosomal duplications. Homogeneous families were enriched in paralog pairs which were closely located on the genome in tandem duplication, probably due to the specific trend of SSD pairs to be duplicated in tandem. Indeed, these homogeneous families were enriched in SSDs, especially in ySSDs, and were associated with a shared tissue-specificity.

5/ Exploration of brain disorder-associated genes

In addition to paralog implication in tissue-specific gene expression, another factor contributing to the importance of a gene is its potential association with disease. Indeed, disease-associated mutations preferentially accumulate in paralogous genes rather than singletons (Dickerson and Robertson 2012). In the case of duplication categories, it has been reported that the proportion of both Mendelian (monogenic) and complex (polygenic) disease genes are enriched in WGD genes in comparison to non-

disease genes (Chen et al. 2013). We decided to refine theses analyses by considering only the genes that are associated with brain diseases. We therefore used the ClinVar database to collect a list of genes that harbored a Single Nucleotide Variant (SNV) or were located within a Copy Number Variant (CNV) and related to a brain disorder (Landrum et al. 2016) (see Materials and Methods). We found that paralogs were enriched in brain disease genes (50.2% of paralogous genes, versus 46% of other genes, p-value = 3.740E-07) (see Supplementary Materials, table S18). We further focused on paralog categories and observed that neither WGDs or SSDs were enriched in brain disease genes (p-value = 0.555) but we noticed that ySSDs genes tended to be very slightly enriched in brain disease genes (53.1% of ySSD genes, versus 49.8% of other paralogs, p-value = 3.535E-02). However, brain disease genes tended to be depleted in tissue-specific genes and were neither enriched in genes coming from homogeneous families or in human-specific paralogs (see Supplementary Materials, table S18). In summary, brain disease genes are not enriched in WGDs and the paralogs associated with brain diseases do not seem to be the same ones that we found in the previous result sections associated with tissue-specificity and co-expressed gene families.

Discussion

As far as we are aware, this study is the first to focus specifically on the spatial expression of paralogs and gene families in the CNS based on post-mortem human tissues analysed by the GTEx consortium. Previous studies based on gene expression analysis between organs and species have already established the important association between paralogs and tissue differentiation (Freilich et al. 2006; Kryuchkova-Mostacci and Robinson-Rechavi 2016). We showed that paralog expression could separate CNS tissues better than singletons, despite their low expression compared to singletons. Therefore, the relationship between paralogs and tissue differentiation is also true for comparisons of the different anatomical regions of the CNS.

Paralogs are known to be more tissue-specific than other genes (Huminiecki and Wolfe 2004; Freilich et al. 2006; Huerta-Cepas and Gabaldón 2011; Guschanski et al. 2017), and ySSDs tend to be more tissue-specific than older paralogs when comparing tissues from different organs (Kryuchkova-Mostacci and Robinson-Rechavi 2016). In our study, we showed that the higher tissue-specificity of paralogs, especially ySSDs, was still observed between the tissues that composed the human CNS. In addition, we found that even wSSDs were enriched in tissue-specific genes compared to other paralogs of the same age (WGDs), thus suggesting that the tissue-specificity is not only associated with the young age of duplication but also with the type of duplication (i.e. with SSD duplications). Our results are consistent with the findings of previous studies which report that the proportion of WGDs in genes expressed in different brain regions is higher than for other organs, while the proportion of SSDs is lower (Satake et al. 2012; Roux et al. 2017). Furthermore, when considering the brain as a whole and

comparing it with other organs, it has been found that WGDs tend to be enriched in brain-specific genes compared to SSDs (Guschanski et al. 2017), which could seem contradictory given the higher tissue-specificity of the SSDs in the CNS. However the higher tissue-specificity of the SSDs, i.e. the fact a SSD gene tends to be more often specific to only one or just a few CNS anatomical regions than a WGD gene, implies that the average expression of SSD genes over the whole brain would be lower than the average expression of WGDs.

A previous study performed using gene expression profiles across mammalian organs established that most of tissue-specific expression variance was explained by the expression level with no significant contribution of the evolutionary time (Guschanski et al. 2017). Using multivariate linear models, we confirmed the major contribution of expression level to tissue-specificity in CNS territories. The paralogous status and the SSD duplication type explained also part of the tissue-specific expression variance. Regarding the evolutionary time, low phyletic ages were significantly associated with high tissue-specificity; a property potentially restricted to CNS tissues. However, we observed that tissue-specific expression did not seem to be enriched in human-specific duplications, but rather associated with less recent ySSDs.

We then studied the gene family level of organization, an area which has so far been poorly investigated by large-scale expression studies of paralogs. Using gene co-expression network analysis of paralogs across CNS tissues we showed that the tissue-specific paralogs were preferentially organized into modules of genes co-expressed in a single or at most a pair of tissues. The characterization of gene families according to the level of co-expression of their member genes has led to the identification of two categories of families: homogeneous families, which are composed of a majority of co-expressed genes, and heterogeneous families. We observed that homogeneous families were enriched in ySSD genes (particularly in human-specific genes) and tandem duplicate pairs, in agreement with a previous study showing that pairs of ySSD paralogous genes tend to be duplicated in tandem and co-expressed just after the duplication event (Lan and Pritchard 2016). Moreover, even among WGD-old duplications, we found that wSSDs were significantly enriched in tandem duplications compared to WGDs, in agreement with previous findings on co-expression of SSD pairs (Acharya and Ghosh 2016). A previous study established that when the two paralogs of an ySSD pair are tissue-specific, they tend to be specific to the same tissue more often than in the case for other paralog pairs (Kryuchkova-Mostacci and Robinson-Rechavi 2016). We confirmed that this property was also true for the CNS territories by showing the high co-expression of ySSD pairs and the enrichment of co-expressed families in tissue-specific families, where the majority of genes were tissue-specific to the same tissue. However, from our analyses on human-specific paralogs and on phyletic age, we noticed that the tissue-specificity would not seem to increase just after the duplication event but after a short time of divergence between SSDs.

With regard to the relationship between paralogs and human diseases, if we consider all the genes

involved in Mendelian or complex genetic diseases, it is known that the mutations accumulate preferentially in paralogs compared to singletons (Dickerson and Robertson 2012; Chen et al. 2013; Singh et al. 2014). Moreover, old paralogs (WGDs and oSSDs) tend to be more frequently associated with diseases (Makino and McLysaght 2010; Chen et al. 2014; Singh et al. 2014; Acharya and Ghosh 2016). Finally, in the case of SSD paralogs, disease genes are known to be enriched in oSSDs and depleted in ySSDs when compared to non-disease genes (Chen et al. 2014). Our study confirmed that paralogs were enriched in brain disease-associated genes. However, we observed that WGD or SSD duplications types were not enriched in these brain disease genes.

From the analysis of gene expression across human and mouse organs, Lan et al. 2016 proposed a model for the retention of SSD duplicates appearing in mammals that is consistent with the dosage-sharing hypothesis in which down regulation, to match expression of the ancestral genes, is the first step enabling the initial survival of young duplicates. Pairs of young paralogs are also often highly co-expressed probably because tandem duplicates are co-regulated by shared regulatory regions (Lan and Pritchard 2016). This model seems to fit the ySSD expression features found across CNS territories, indeed they seem to be weakly expressed and organized within small families of co-expressed genes. Furthermore, we can refine this model for the CNS tissues because our results seem to show that SSD paralogs are highly co-expressed just after their duplication and will become more tissue-specific but less co-expressed through chromosomal rearrangement, suggesting a subfunctionalization of expression as indicated in the Lan et al, 2016 model. A last hypothesis could be that the ySSD genes with shared tissue specificity that retain their co-expression over time, could point out to duplicates that should not be preserved in the long term.

Overall, this study indicates the major involvement of ySSDs (SSDs younger than WGDs) within the human CNS in terms of tissue-specificity and co-expression compared to multiple organ analyses. This suggests either intra-organ properties that could be tested on another complex organ or either a brain particularity.

Material and methods

Human genes, duplication events and families

A list of 21,731 human genes, with both their HGNC gene symbol and their Ensembl IDs (GRCh37, release 59), was collected based on the work of Chen and co-workers (Chen et al. 2013). Among these genes, 14,084 paralogs made up of 3,692 gene families, identified by TreeFam methodology (Ruan et al. 2008), were obtained from Chen and co-workers (Chen et al. 2013). These authors downloaded all gene families from the TreeFam v.8.0 database, which identifies duplicates based on gene family evolution. To define human gene families, Chen at al. excluded gene families that did not contain human genes, or that contained genes from less than four different species. In cases where a gene family or an orthologous group contained two or more human genes, they considered these genes as duplicates.

Moreover, for each paralog, they represented the phyletic age of its last duplication event by the total branch length from the node indicating where the duplication event had happened on the species tree to the human leaf node, and they assigned the associated duplicate (Chen et al. 2012; Chen et al. 2013). A second list of 20,415 genes was extracted from (Singh et al. 2014). This gene ID list was converted to HGNC gene symbols and intersected with the first list in order to annotate it (17,805 protein-coding genes in common). Thus, in the present study, we collected the duplication category for each paralog (Singh et al. 2014). Singh et al. obtained WGD annotations from (Tinti et al. 2012) and obtained their SSD annotations by running an all-against-all BLASTp using human proteins (Singh et al. 2012). Singh and co-workers defined genes as singletons if they were not classified as WGDs or SSDs and they obtained the duplication age for SSD genes from the Ensembl compara (Vilella et al. 2009). They classified paralogs into the following categories: WGD, SSD, ySSD (i.e. SSD with duplication date younger than WGD), oSSD (i.e. SSD with duplication date older than WGD) and wSSD (i.e. WGD-old SSD). There were 5,390 annotated paralogs originating from the WGD and 4,889 from SSD (2,104 from ySSD, 1,354 from oSSD and 1,431 from wSSD). Moreover, there were 2,607 paralogs without annotations and 1,198 paralogs annotated as both WGD and SSD (WGD-SSD). The WGD-SSD paralogs were not included into the WGD or the SSD duplication categories. However, the unannotated and WGD-SSD paralogs were both considered into the paralog group. We verified that these paralog duplication categories were consistent with the phyletic ages (duplication dates) collected from Chen and co-workers (Chen et al. 2012; Chen et al. 2013) (see Supplementary Materials, fig. S3). The list of our paralogous gene pairs and gene families is given in the Supplementary Materials, table S1. The evolutionary annotation of paralogous genes is indicated in the Supplementary Materials, table S2. The list of singleton genes is given in the Supplementary Materials, table S12. Furthermore for the analysis of the duplicate pairs, we considered only the 3,050 pairs which appeared twice in our paralog list (i.e. where the first paralog is associated with the second paralog and vice versa and where the duplication category annotation is the same for both paralogs); genomic distances between duplicate pairs were obtained from Ensembl (GRCh37/90). We also obtained a list of paralogous genes generated by human-specific duplication events (Dennis et al. 2017). From these human-specific duplications, 22 were in our list of paralogs and 8 were among the genes expressed in CNS.

Gene expression profiles in CNS tissues

We obtained gene counts and RPKM (Reads Per Kilobase Million) values for 63 to 125 individuals (1259 post-mortem samples – RNA integrity > 6) distributed over 13 CNS tissues (cerebellum, cerebellar hemisphere, cortex, frontal cortex, anterior cingulate cortex, hypothalamus, hippocampus, spinal cord, amygdala, putamen, caudate, nucleus accumbens and substantia nigra) from the GTEx consortium data release 6 (GRCh37) (Ardlie et al. 2015). The CNS tissue associated with each GTEx patient sample used in our study is indicated in the Supplementary Materials, table S11. These gene expression data, calculated by GTEx took into account only uniquely mapped reads

(https://gtexportal.org). After filtering out low-information content genes (genes with a null variance across samples and weakly expressed genes, with mean expression per tissue lower than 0.1 RPKM for all tissues), we kept for analyses a total 16,427 genes distributed across 10,335 paralogs (5,114 WGD, 3,719 SSD, 1,192 ySSD, 1,260 wSSD and 1,267 oSSD, 966 WGD-SSD and 536 without annotations) grouped in 3,487 families and 6,092 singletons. It should be noted that all analyses of the articles were performed on this list of expressed genes only, except for the analysis on brain disease genes. Moreover, the WGD-SSD paralogs were not included in the WGD or SSD categories. However, unannotated and WGD-SSD paralogs as well as all other duplication categories were considered to constitute the paralog group. Gene RPKM values were log-transformed (log2 (RPKM + 1)) and adjusted by linear regression for batch effects and various biological effects (platform, age, gender and the first 3 principal components of genetic data illustrating the population structure given by the GTEx Consortium; the intercept of the regression was not removed from the residuals in order to keep the mean differences between genes (https://www.cnrgh.fr/genodata/BRAIN_paralog). These filtered, log-transformed and adjusted RPKM values were used as input for unsupervised classification of brain tissues, as well as for gene co-expression network inference and for tissue-specificity analysis. Moreover, gene expression data for tissues considered to anatomically overlap were merged by calculating the average expression value across related tissues prior to the tissue-specificity analysis. Therefore, from an initial list of 13 tissues, we defined a short list of 7 CNS regions: cerebellum (cerebellum and cerebellar hemisphere), cortex (cortex, frontal cortex and anterior cingulate cortex), basal ganglia (putamen, nucleus accumbens and caudate), amygdala-hippocampus, hypothalamus, spinal cord and substantia nigra.

Unsupervised clustering of gene expression profiles

Gene expression profiles (filtered and adjusted RPKM values) generated by the GTEx Consortium for the 1,259 samples distributed across the 13 CNS tissues, were clustered by unsupervised hierarchical clustering using the pheatmap package of R version 3.4 (similarity measure: Pearson correlation, clustering method: average linkage). We estimated the relevance of the clustering according to the expected groups of CNS tissues. We evaluated, independently, the clusterings generated from protein-coding genes, paralogs and singletons, using adjusted rand index (Hubert and Arabie 1985) after cutting trees (so that we obtained 30 clusters for each gene category).

Differential gene expression analysis

Genes with low-information content were removed before differential gene expression (DGE) analysis. DGE analysis was performed by DESeq2 (Love et al. 2014) on count data for each pair of CNS tissues, with the “median ratio” between-sample normalization and using batch and biological effects as covariates. For each tissue pair, we then corrected gene p-values for the number of tested genes using FDR (Benjamini and Hochberg 1995) and obtained a list of significantly differentially expressed genes (DEGs) (FDR<0.05). Finally, we considered only the DEGs with a log2 fold-change greater than 0.5.

Inference of gene co-expression networks

The gene network inference was carried out using the Weighted Gene Correlation Network Analysis (WGCNA) methodology (Langfelder and Horvath 2008), which generates co-expression networks and identifies modules (groups) of co-expressed genes. We applied the WGCNA tool only to paralogous gene expression data (RPKM) across the GTEx samples of the 13 CNS tissues. Genes were grouped into modules according to their expression profile similarity. The module named "grey", which grouped genes that were not considered as co-expressed by WGCNA, was composed of genes with very low variability across all samples. Since we had removed the genes with no variance across tissue samples and those which were weakly expressed before performing the WGCNA analysis, the grey module was small in size (104 genes). Furthermore, if this filtering had not been performed, some of the genes with an overall weak expression might have been integrated into co-expression modules, thus creating a bias. One of our goals was to compare gene families to co-expression modules. Given that 47% of gene families have a size equal to 2, we optimized WGCNA parameters to obtain small highly co-expressed modules (see Supplementary Materials, result S3).

Homogeneous and heterogeneous families

Definition. A gene family was defined as homogeneous if the majority, more than 60%, of its member genes were included in the same co-expression module. It should be noted that the total size of gene families was used to compute this percentage, even if some member genes were not in the list of expressed paralogs. Gene families which did not respect this homogeneity rule, i.e. those with member genes scattered over different co-expression modules, were defined as heterogeneous.

Assessment of the significance of the number of homogeneous families. Starting from the paralog modules obtained with WGCNA, we used a permutation procedure (by permuting 1,000 times the module labels of paralogs and counting the number of falsely homogeneous families for each permutation) and were able to conclude that the number of homogeneous families was significantly large, since for each permutation the number of falsely homogeneous families was lower than the number that we obtained, leading to an empirical p-value inferior to 10^{-3}.

Tissue-specificity calculation

Tau score calculation. To select tissue-specific genes, we used the τ score (Yanai et al. 2005) to estimate the degree of tissue-specificity of each gene in our set of CNS tissues:

$$\tau = \frac{\sum_{i=1}^{n} \left(1 - \widehat{X}_i\right)}{n-1} \; ; \; \widehat{X}_i = \frac{x_i}{\max\limits_{1 \leqslant i \leqslant n} \left(x_i\right)} \qquad (2)$$

In this equation, x_i is the mean expression of a given gene in tissue i and n is the number of different tissues. τ varies from 0 to 1 where 0 indicates that the gene is broadly expressed and 1 that the gene is tissue-specific. For τ computation, genes must have a positive mean of expression in every CNS region. Although we log-normalized expression data with log2(RPKM+1) leading to positive expression values, the correction for batch and some biological effects induced some negative values in gene mean

expression. We replaced the negative values by zeros to keep all protein coding genes (16,427 genes) for the τ score computation.

We pooled expression data generated by GTEx for the 13 tissues into 7 CNS regions so that the τ score would not decrease artificially for genes specific to several close tissues.

Tau score threshold defined by permutations. The τ score was computed for each gene and for the 7 CNS regions. We then plotted the τ score distribution obtained from all protein coding genes (fig. 2A). However, there is no general τ score threshold at which a gene is considered to be tissue-specific.

To define a tissue-specificity threshold, we implemented a statistical method based on permutations. We applied 1000 permutations on the region labels assigned to the samples to shuffle the correspondence between samples and regions. For each permutation, τ scores were recomputed for each gene. The distribution of the 1000 X 16427 τ scores obtained from the permutations is given in the Figure 2. For each gene and its original τ score, a p-value was then calculated as the proportion of permutation-based τ scores higher than the original τ score. The Benjamini-Hochberg correction for the number of genes tested was applied to all p-values. Genes with a corrected p-value lower than 0.01 were declared tissue-specific, which corresponded to a τ score threshold of 0.525 (fig. 2A). Visualization of gene profiles across brain regions at different windows of the τ score showed tissue-specificity beyond the τ score threshold of 0.525 (see Supplementary Materials, fig. S1). However even for τ scores in the range [0.5-0.75] some genes were still expressed in two regions. Therefore for each tissue-specific gene, we considered that the CNS region with the highest expression value to be the specific region.

Multivariate linear models

To evaluate the contribution of different factors (expression level, duplication status (singleton or paralog) to tissue-specificity (Tau scores), we constructed multivariate linear models in R version 3.4 (package stats) that included all protein coding genes. Both the mean and the maximum expressions over CNS regions were tested to account for expression level. We also constructed multivariate linear models from paralogous genes only to evaluate the contribution of expression level, duplication age, duplication type (WGD or SSD) to tissue-specificity. Both the phyletic age and the categorical duplication age (older than WGD/WGD-old, younger than WGD) were used to account for the duplication age.

Brain disease genes

Genetically altered genes associated with brain disease. The ClinVar database was used to collect genes linked to brain diseases (Landrum et al. 2016). We considered brain disease-associated genes containing a pathogenic alteration (SNV or located within a CNV (duplication, amplification or deletion). We selected genes associated by ClinVar to the following “Disease/Phenotypes” : “Parkinson”, “Alzheimer”, “brain”, “Autism”, “Epilepsy”, “Aicardi”, “Angelman”, “Aphasia”, “Apraxia”, “Asperger”, “Behcet”, “spinal”, “Canavan”, “Charcot”, “Chorea”, “Dementia”, “Dyslexia”, “Fabry”, “Gaucher”,

“Gerstmann”, “Huntington”, “Refsum”, “Joubert”, “Kennedy”, “Klippel”, “Krabbe”, “learning”, “mental”, “Leigh”, “Leukodystrophy”, “migraine”, “Niemann”, “Rett”, “Sandhoff”, “syncope”, “Tay-sachs”, “Tourette”, “nervous”, “Schizophrenia”, “Narcolepsy”“neuro”, “cephal”, “cortico”, “crani”, “mening”, “psych”. We then removed genes associated with tumors or cancers. We obtained a total of 10,375 genes linked to brain diseases including 7,989 gene expressed in CNS tissues. Among them, 5,184 were expressed paralogous genes.

Acknowledgments

This study received funding from the Université Paris-Sud (support to S.B.J.) and the Fondation pour la Recherche Médicale (support to S.C.). We are grateful to Mark Robinson-Rechavi for his feedback on the methods and results. We thank Steven McGinn and Elizabeth May for English language editing. We also thank Carène Rizzon, Margot Coréa, Olivier Jaillon, François Artiguenave and Morgane Pierre-Jean for constructive discussions.

References

Acharya D, Ghosh TC. 2016. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics 17:71.

Acharya D, Mukherjee D, Podder S, Ghosh TC. 2015. Investigating Different Duplication Pattern of Essential Genes in Mouse and Human. PLoS ONE 10:e0120784.

Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348:648–660.

Assis R, Bachtrog D. 2015. Rapid divergence and diversification of mammalian duplicate gene functions. BMC Evol. Biol. 15:138.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. A Stat. Soc. 57:289–300.

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, Jin W-L, Vanderhaeghen P, Ghosh A, Sassa T, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell 149:923–935.

Chen W-H, Trachana K, Lercher MJ, Bork P. 2012. Younger Genes Are Less Likely to Be Essential than Older Genes, and Duplicates Are Less Likely to Be Essential than Singletons of the Same Age. Mol. Biol. Evol. 29:1703–1706.

Chen W-H, Zhao X-M, van Noort V, Bork P. 2013. Human monogenic disease genes have frequently functionally redundant paralogs. PLoS Comput. Biol. 9:e1003073–e1003073.

Chen W-H, Zhao X-M, van Noort V, Bork P. 2014. Comments on “Human dominant disease genes are enriched in paralogs originating from whole genome duplication”. PLoS Comput. Biol. 10:e1003758–e1003758.

Dehal P, Boore JL. 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLOS

Biol. 3:e314.

Dennis MY, Eichler EE. 2016. Human adaptation and evolution by segmental duplication. Curr. Opin. Genet. Dev. 41:44–52.

Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, Antonacci F, Penewit K, Denman L, Raja A, et al. 2017. The evolution and population diversity of human-specific segmental duplications. Nat. Ecol. Evol. 1:69.

Dickerson JE, Robertson DL. 2012. On the origins of Mendelian disease genes in man: the impact of gene duplication. Mol. Biol. Evol. 29:61–69.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. Nature 485:376–380.

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, et al. 2015. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science 347:1465–1470.

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151:1531–1545.

Freilich S, Massingham T, Blanc E, Goldovsky L, Thornton JM. 2006. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. Genome Biol. 7:R89.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in Expression between Duplicated Genes in Arabidopsis. Mol. Biol. Evol. 24:2298–2309.

Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. Genome Res. 27:1461–1474.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol. 8:R209.

Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, et al. 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 489:391–399.

Hubert L, Arabie P. 1985. Comparing partitions. J. Classif. 2:193–218.

Huerta-Cepas J, Gabaldón T. 2011. Assigning duplication events to relative temporal scales in genome-wide studies. Bioinforma. Oxf. Engl. 27:38–45.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. 14:1870–1879.

Ibn-Salem J, Muro EM, Andrade-Navarro MA. 2017. Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res. 45:81–91.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11:97–108.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, et al. 2011. Spatiotemporal transcriptome of the human brain. Nature 478:483–489.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2016. Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. PLOS Comput. Biol. 12:e1005274.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. Brief. Bioinform. 18:205–214.

Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science 352:1009–1013.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 44:D862-868.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. Trends Genet. 21:602–607.

Lian S, Liu T, Jing S, Yuan H, Zhang Z, Cheng L. 2018. Intrachromosomal colocalization strengthens co-expression, co-modification and evolutionary conservation of neighboring genes. BMC Genomics 19:455.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. Trends Genet. TIG 25:152–155.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc. Natl. Acad. Sci. 107:9270–9274.

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. Nat. Genet. 31:200–204.

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res. 17:1254–1265.

Ohno S. 1970. Evolution by gene duplication. Springer-Verlag Berlin Heidelberg

Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. Hereditas 59:169–187.

Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH. 2008. Functional organization of the transcriptome in human brain. Nat. Neurosci. 11:1271–1282.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197.

Parikshak NN, Gandal MJ, Geschwind DH. 2015. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nat. Rev. Genet. 16:441–458.

Pierson E, Consortium the Gte, Koller D, Battle A, Mostafavi S. 2015. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. PLOS Comput. Biol. 11:e1004220.

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet.

3:827–837.

Qian W, Liao B-Y, Chang AY-F, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet. TIG 26:425–430.

Roux J, Liu J, Robinson-Rechavi M. 2017. Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates. Mol. Biol. Evol. 34:2773–2791.

Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, Hériché J-K, Hu Y, Kristiansen K, Li R, et al. 2008. TreeFam: 2008 Update. Nucleic Acids Res. 36:D735-740.

Satake M, Kawata M, McLysaght A, Makino T. 2012. Evolution of Vertebrate Tissues Driven by Differential Modes of Gene Duplication. DNA Res. 19:305–316.

Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H. 2012. On the Expansion of "Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates. Cell Rep. 2:1387–1398.

Singh PP, Affeldt S, Malaguti G, Isambert H. 2014. Human dominant disease genes are enriched in paralogs originating from whole genome duplication. PLoS Comput. Biol. 10:e1003754–e1003754.

Stephens SG. 1951. Possible Significance of Duplication in Evolution. In: Advances in Genetics. Vol. 4. Elsevier. p. 247–265.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. 2010. Diversity of Human Copy Number Variation and Multicopy Genes. Science 330:641–646.

Teshima KM, Innan H. 2008. Neofunctionalization of Duplicated Genes Under the Pressure of Gene Conversion. Genetics 178:1385–1398.

Tinti M, Johnson C, Toth R, Ferrier DEK, MacKintosh C. 2012. Evolution of signal multiplexing by 14-3-3-binding 2R-ohnologue protein families in the vertebrates. Open Biol. 2:120103.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327–335.

Xie T, Yang Q-Y, Wang X-T, McLysaght A, Zhang H-Y. 2016. Spatial Colocalization of Human Ohnolog Pairs Acts to Maintain Dosage-Balance. Mol. Biol. Evol. 33:2368–2375.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659.

Zhang B, Horvath S. 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. Stat. Appl. Genet. Mol. Biol. 4:Article17.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Table 1. Enrichments in tissue-specific genes for the tested and reference gene groups

Reference group[a]	Tested group for tissue-specificity[a]	Percentage of tissue-specific genes in the tested group (%)	Chi-squared test P-value[b]	Odds ratio[c]
Protein coding genes	Paralogous genes	19.2	2.045E-18*	1.48
Paralogous genes[d]	WGD genes	15.7	1.061E-18*	0.64
	SSD genes	22.6	9.022E-11*	1.39
	ySSD genes	28.6	6.341E-18*	1.82
SSD genes	ySSD genes	28.6	3.483E-09*	1.62
	oSSD genes	15.6	2.729E-13*	0.52
WGD + wSSD genes	WGD genes	15.7	5.185E-12*	0.59

[a] Abbreviations for gene duplication categories : WGD (Whole-Genome Duplication), SSD (Small-Scale Duplication), ySSD (younger SSD occuring after WGD events), oSSD (older SSD occuring before WGD events) and wSSD (WGD-old SSD occuring around WGD events).

[b] Application of Chi-squared tests (or of Fisher's exact test when the Chi-squared test could not be applied) with a corrected p-value threshold = 7.14E-03 (Bonferroni correction for 7 statistical tests).

[c] The odds ratio (>1 or <1) indicates the group (tested or non-tested respectively) in which there is an enrichment.

[d] The paralog reference group includes the genes belonging to WGD, SSD and WGD-SSD categories and the paralogs without annotation.

Table 2. Multivariate linear regression to model the relationship between the tissue-specificity of genes (corresponding to the Tau score of each gene) and explanatory variables (the maximal gene expression value, the duplication status, the phyletic age of the duplication and the duplication type) depending on the reference group.

(a)

Reference group: Protein coding genes	**t value**	**P-value[b]**
(Intercept)	139.39	0.00E00*
Max Expression[c]	-71.84	0.00E00*
Duplication status[d]	18.69	3.78E-77*
Residual standard error: 0.1791 on 16424 degrees of freedom Multiple R-squared: 0.2511, Adjusted R-squared: 0.251 F-statistic: 2754 on 2 and 16424 DF, p-value: < 2.2E-16		

(b)

Reference group: Paralogous genes[a]	**t value**	**P-value[b]**
(Intercept)	86.068	0.00E00*
Max Expression[c]	-48.557	0.00E00*
Phyletic age	-5.506	3.77E-08*
Duplication type[e]	-0.104	0.917
Residual standard error: 0.1819 on 8815 degrees of freedom		

Multiple R-squared: 0.2239, Adjusted R-squared: 0.2236 F-statistic: 847.7 on 3 and 8815 DF, p-value: < 2.2E-16

(c)

Reference group: WGD + wSSD genes[a]	**t value**	**P-value[b]**
(Intercept)	99.811	0.00E00*
Max Expression[c]	-37.964	1.81E-284*
wSSD status[f]	3.676	2.39E-04*
Residual standard error: 0.1816 on 6363 degrees of freedom Multiple R-squared: 0.1917, Adjusted R-squared: 0.1914 F-statistic: 754.4 on 2 and 6363 DF, p-value: < 2.2E-16		

[a] Paralogs without duplication type annotation and annotated as both WGD and SSD (WGD-SSD) were not considered.

[b] p-value threshold = 0.05.

[c] An average expression value per gene is first calculated for each CNS tissue across the patients, and then each gene is associated with the maximum tissue-averaged values.

[d] The duplication status corresponds to singleton or paralog.

[e] The duplication type corresponds to SSD or WGD.

[f] The wSSD status corresponds to not wSSD or wSSD.

Table 3. Enrichments in genes from homogeneously expressed families for the tested and reference gene groups

Reference group[a]	**Tested group for homogeneous family expression[a]**	**Percentage of homogeneous family genes in the tested group (%)**	**Chi-squared test P-value[b]**	**Odds ratio[c]**
Paralogous genes[d]	SSD genes	3.3	2.777E-04*	1.59
	ySSD genes	5.2	5.758E-10*	2.49
	Tissue-specific genes	4.7	5.374E-12*	2.45
Paralogous genes[d]	Human-specific paralogous genes	50	2.457E-05*	39.65
SSD genes		50	6.741E-05*	30.73
ySSD genes		50	3.868E-04*	19.58
Paralogous genes[d]	Tissue-specific families[e]	45	1.691E-69*	42.94

[a] Abbreviations for gene duplication categories : WGD (Whole-Genome Duplication), SSD (Small-Scale Duplication) and ySSD (younger SSD occuring after WGD events).

[b] Application of Chi-squared tests (or of Fisher's exact test when the Chi-squared test could not be applied) with a corrected p-value threshold = 7.14E-03 (Bonferroni correction for 7 statistical tests).

[c] The odds ratio (>1 or <1) indicates the group (tested or non-tested respectively) in which there is an enrichment.

[d] The paralog reference group includes the genes belonging to WGD, SSD and WGD-SSD categories and the paralogs without annotation.

[e] Genes included into tissue-specific families. Only genes specific to the major tissue are considered.

Fig. 1. Unsupervised hierarchical clustering of genes expressed in human central nervous system regions

Hierarchical clustering of genes expressed in the CNS regions was performed based on gene pairwise distance in terms of correlation across gene expression values. The three gene groups considered are: protein-coding genes, singletons and paralogous genes. Each CNS region is represented by a different color. The tissues belonging to the same anatomically defined CNS region are represented in the same color: blue for the cerebellum region (cerebellum and cerebellar hemisphere tissues), green for the cortex region (cortex, frontal cortex and anterior cingulate cortex tissues), purple for the basal ganglia region (putamen, nucleus accumbens and caudate tissues), and red for the amygdala-hippocampus region (amygdala and hippocampus tissues). The remaining tissues are considered as independent CNS regions: pink for the hypothalamus region, yellow for the spinal cord region and black for the substantia nigra.

Fig. 2. Tissue-specific expression of paralogous genes across human CNS regions.

(a) Density plot of original Tau scores (blue line) calculated from the expression values of 16227 protein coding genes, and permutated Tau scores (purple line) calculated from 1000 x 16427 permutations. The tissue-specifity threshold of 0.525 (red dotted line) is defined, from permutated scores using the Benjamini-Hotchberg corrected P-value of 0.01. **(b)** Unsupervised hierarchical clustering of tissue-specific genes expressed across CNS regions. The heatmap illustrates the mean gene expression calculated over sample cohort for each CNS region.

Fig. 1.

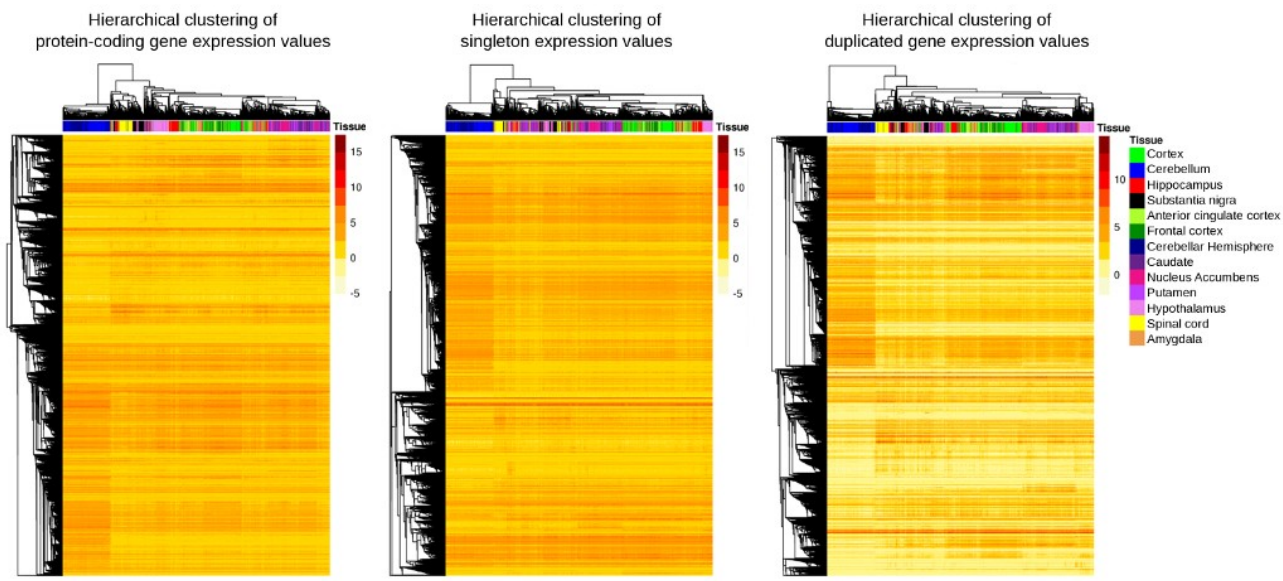
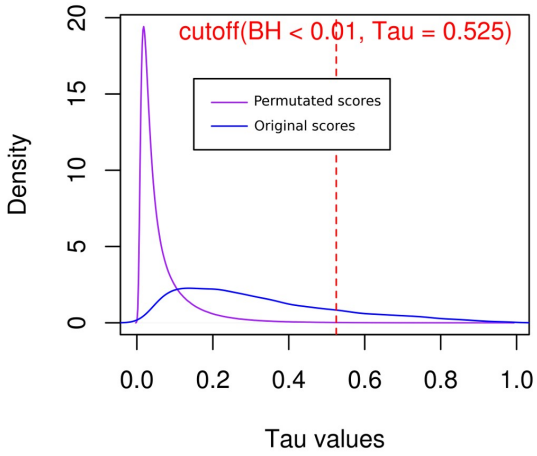
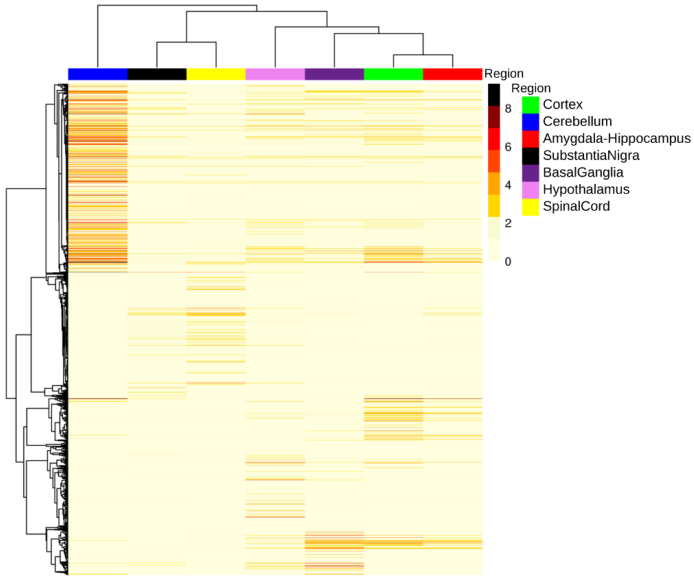


Fig. 2.

(a)


cutoff(BH < 0.01, Tau = 0.525)
Permutated scores
Original scores
Density
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
Tau values

(b)


Region
Region
8
6
4
2
0
Cortex
Cerebellum
Amygdala-Hippocampus
SubstantiaNigra
BasalGanglia
Hypothalamus
SpinalCord