RESEARCH

TITLE

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

AUTHOR NAMES

Solène Brohard-Julien,[*,1,2,3] Vincent Frouin,[2] Vincent Meyer,[1] Smahane Chalabi,[1] Jean-François Deleuze,[1,4,5] Edith Le Floch[*,1,7], and Christophe Battail[*,1,6,7]

INSTITUTIONS

[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]Centre d'Etude du Polymorphisme Humain, Fondation Jean Dausset, Paris, France

[5]Centre de Référence, d'Innovation, d'expertise et de transfert (CREFIX), Evry, France

[6]Univ. Grenoble Alpes, CEA, INSERM, BCI U1036, 38000 Grenoble, France

[7] These authors contributed equally to this work.

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

RUNNING TITLE

Duplicate gene expression in human brain regions

KEY WORDS

paralog, small-scale duplication, tissue-specific expression, human central nervous system, co-expression network

ABSTRACT

Gene duplication has generated new biological functions during evolution that have contributed to the increase in tissue complexity. Several comparative genomics and large-scale transcriptional studies performed across different organs have observed that paralogs and particularly small-scale duplications (SSD) tend to be more tissue-specifically expressed than other gene categories. However, the major involvement of whole-genome duplications (WGD) was also suggested in the emergence of tissue-specific expression features in the brain. Our work complements these previous studies by exploring intra-organ expression properties of paralogs through multiple territories of the human central nervous system (CNS) using transcriptome data generated by the Genotype-Tissue Expression (GTEx) consortium. Interestingly, we show that paralogs, and especially those originating from young SSDs (ySSD), are significantly implicated in tissue-specific expression between CNS territories. Our analysis of co-expression of gene families across human CNS tissues allows also the detection of the tissue-specific ySSD duplicates expressed in the same tissue. Moreover, we uncover the distinct effect of the young duplication age, in addition to the SSD type effect, on the tissue-specific expression of ySSDs within the CNS. Overall, our study suggests the major involvement of ySSDs in the differentiation of human CNS territories and shows the added value of exploring tissue-specific expression at both the inter and intra-organ levels.

INTRODUCTION

Comparative genomics and large-scale transcriptional studies have highlighted the major contribution of gene duplication to tissue differentiation and phenotypic diversity (Ohno 1970; S. Chen et al. 2013). The fact that some paralogs are retained in genomes through evolution seems to be initially favored by dosage balance (Zhang 2003) and their long-term preservation is then made possible by the following two processes: the neo-functionalization, which consists in the gain of a new function by one duplicate potentially associated with a different spatial expression (Stephens 1951; Force et al. 1999; Teshima and Innan 2008; Innan and Kondrashov 2010), or the sub-functionalization which consists in the partition of the ancestral function or spatial expression between duplicates (Prince and Pickett 2002;

Assis and Bachtrog 2015). The divergence of spatial expression between paralogs can be approached by the study of gene tissue-specificity, which indicates whether a gene has a broad or narrow expression pattern across a collection of tissues (Zhang 2003; Freilich et al. 2006; Lan and Pritchard 2016). The comparison of transcriptomes between different mouse organs has shown that the brain was the one that expresses the highest proportion of tissue-specific paralogs in relation to the total number of genes expressed in the brain, while it does not express the highest proportion of tissue-specific singletons (Freilich et al. 2006). The brain is therefore a model perfectly suited to the detailed exploration of the transcriptional properties of the duplicated genes.

Among the 60% of human genes considered as paralogs (S. Chen et al. 2013), some come from whole-genome duplications (WGD) in early vertebrate lineage approximately 500 million years ago (McLysaght et al. 2002; Nakatani et al. 2007), the others come from small scale duplications (SSD) that have occurred throughout the evolution (Hakes et al. 2007). A comparison in mammals, notably in human, of the brain transcriptome with those of other organs has shown that WGDs tend to be enriched in brain-specific genes compared to SSDs (Satake et al. 2012; Guschanski et al. 2017; Roux et al. 2017). This supports the theory that genome duplications have allowed vertebrates to develop more complex cellular organizations, such as the different brain tissues (Holland 2009; Chen et al. 2011).

In complement of the role of the WGDs in the tissue complexity, some theories support the idea that young duplicated genes tend to be preferentially expressed in evolutionarily young tissues (Domazet-Lošo and Tautz 2010). Moreover, a higher proportion of primate-specific paralogs were found to be up-regulated in the developing human brain compared to the adult brain, whereas this expression pattern was not found for older duplications (Zhang et al. 2011). Regarding recent duplications, that emerged in the human lineage, studies have suggested their contribution to human-specific adaptive traits, such as the gain of brain complexity (Sudmant et al. 2010; Dennis and Eichler 2016; Dennis et al. 2017; Guschanski et al. 2017).

While the expression properties of paralogs between different organs, including the brain, have been well studied, we have little knowledge of the expression characteristics of duplicated genes between different regions of the same organ. Large-scale transcriptional profiling of neuroanatomic regions

(Melé et al. 2015) allows us now to further investigate paralog expression between the different territories of the human central nervous system (CNS) according to their evolutionary properties.

This present study explores in detail the expression patterns of paralogs between the different territories of the human CNS, using the GTEx resource, according to their evolutionary characteristics and gene families. We started assessing whether duplicated genes were associated with differences in expression between CNS tissues and we investigated their tissue-specificity. Secondly, we studied the evolutionary characteristics of tissue-specific paralogs such as their age and the type of duplication event. We then analyzed the organization of paralogs in families using co-expression to define co-expressed gene families and studied their tissue-specificity and evolutionary characteristics.

A better comprehension of the biology of paralogs could also support our understanding of diseases, since disease-associated genes have been found to be over-represented in paralogs compared to singletons (Makino and McLysaght 2010; Dickerson and Robertson 2012; W.-H. Chen et al. 2013) and particularly in WGDs and old SSDs (Singh et al. 2014; Acharya and Ghosh 2016). Thus, we finally explored the association of paralogs with human brain diseases.

RESULTS

1/ Association of paralog expression with CNS differentiation

We considered in our study all human protein coding genes and the information collected on duplication events in order to split the gene population into paralogs and singletons (S. Chen et al. 2013) (Methods). 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 (Melé et al. 2015). Gene expression data for hundreds of individuals from 13 normal brain-related tissues (Methods) were obtained from the GTEx consortium. 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 (Melé 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 (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, from the next result sections, we will pool together some of the 13 initial tissues that showed similar expression profiles in order to define a shorter list of 7 CNS regions (Methods) that will be used for the tissue-specificity analysis.

The relevance of our experimental classification was evaluated according to the expected belonging of samples to the 13 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.

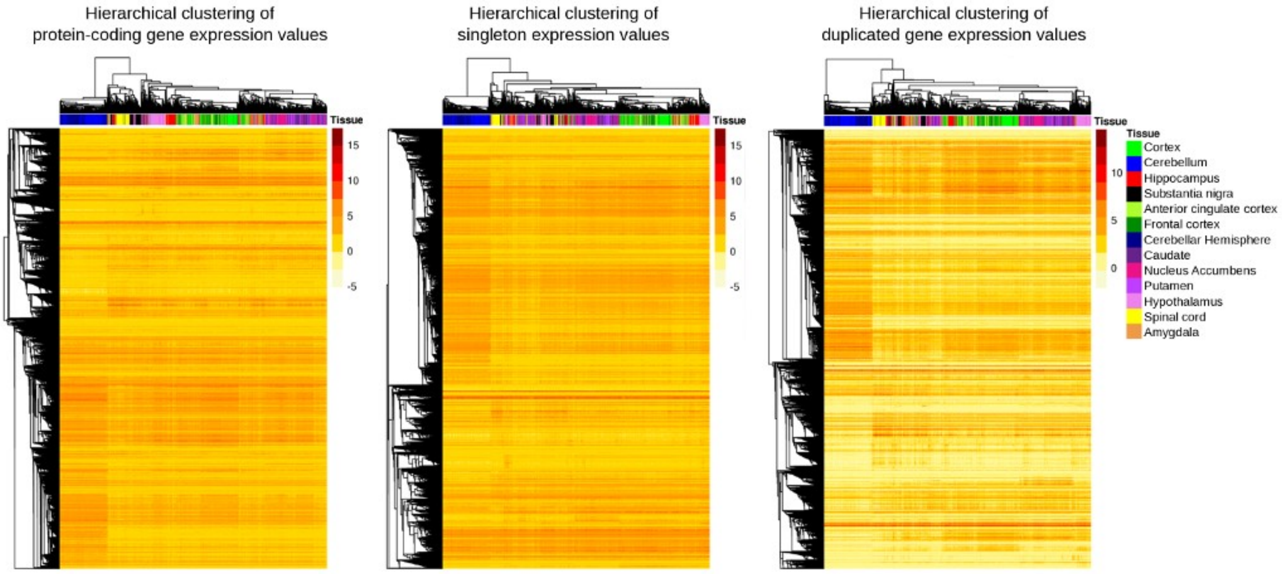

Figure 1. Unsupervised hierarchical clustering of genes expressed in human central nervous system regions. Hierarchical clustering of genes expressed in the CNS regions was performed based on gene pairwise distance in terms of correlation across gene expression values. The three gene groups considered 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.

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 (Methods). We obtained a list of significantly differentially expressed genes (DEGs) for each pair of tissues (Supplemental 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 (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 across 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). This approach led to a Tau threshold of 0.525. We found that 17% (2,829) of protein-coding genes expressed in the CNS regions were tissue-specific (Supplemental 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). We confirmed this association between paralogs and tissue-specificity in addition to their expression level, by using a multivariate linear model, inspired from the analyses of Guschanski et al. 2017, that predicts the Tau score of a gene from its maximal expression over the CNS regions and its duplication status (Supplemental Materials Result S1 and Table S16A).

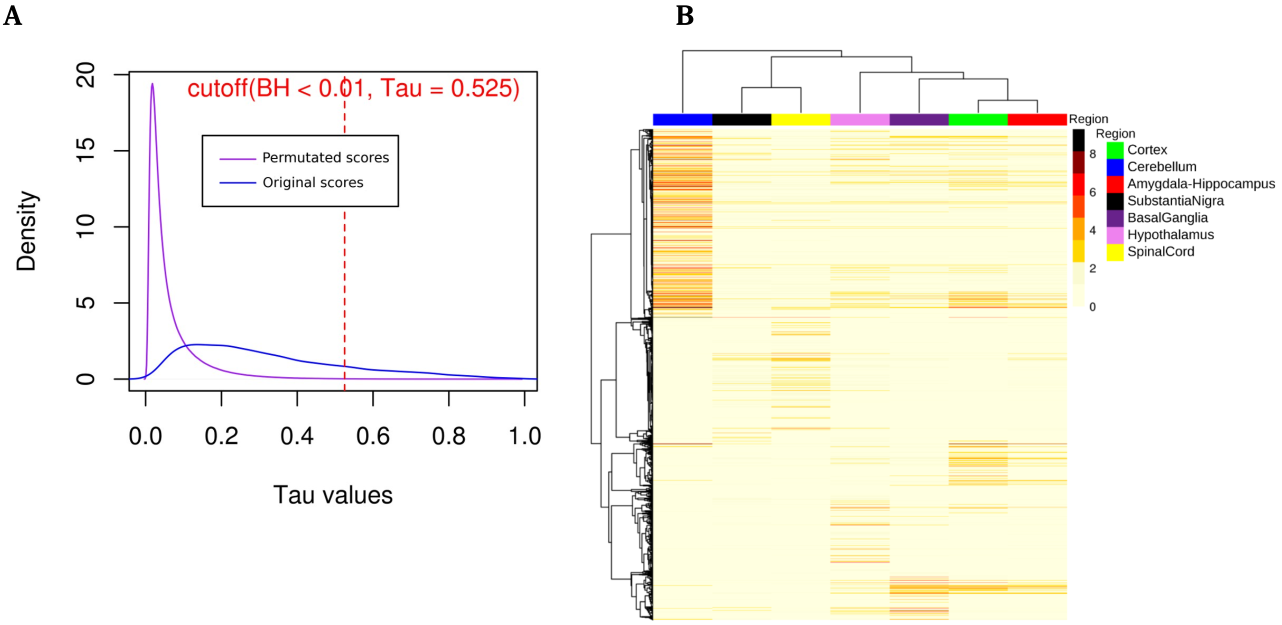

Figure 2. Tissue-specific expression of paralogous genes across human CNS regions. **(A)** Density plot of original Tau scores (blue line) calculated from the expression values of 16227 protein coding genes, and permutated Tau scores (purple line) calculated from 1000 x 16427 permutations. The tissue-specificity 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.

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.

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 (Supplemental 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 (Supplemental Materials Table S6) compared to an almost constant proportion of expressed genes across these regions (Supplemental 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 (Supplemental 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. Precisely, we 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 an 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) (Methods) (Singh et al. 2014). Using our collection of paralogs with tissue-specific expression between CNS regions, we performed statistical tests to determine if they were enriched in particular duplication events (WGD or SSD) or dates of SSDs (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 were 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 paralogs duplicated around the WGD events (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 effect of duplication type, we also tested the association between duplication age categories 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 SSD paralogs only (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 paralogs (15.6% of oSSDs versus 26.2% of the remaining SSDs, p-value = 2.729E-13) (Table 1). We confirmed the contribution of both duplication age and duplication type to the tissue-specificity of paralogs, independently of the effect of their maximal expression level, using multivariate linear models (Supplemental Materials Result S1 and Table S16C, D). 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.

To refine 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 (Methods) (Dennis et al. 2017) and found no significant associations (Supplemental Materials Table S19). However, the statistical test leading to this result may be underpowered because of the small number of genes and of the abundance estimation uncertainty of recent paralogs with high sequence identity (Dougherty et al. 2018). 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 (Supplemental 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.

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

We previously found that paralogs, and especially SSDs and ySSDs, were involved in territorial expression between the different CNS regions, notably through tissue-specificity. In this section, we tried to determine if the paralogs within gene families tended to share the same tissue-specificity across CNS regions. 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 the 13 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 (Supplemental Materials Fig. S2; 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) (Supplemental 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 (Supplemental Materials Result S4).

To first check the relationship between co-expression and shared tissue-specificity, we analyzed 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 (Supplemental Materials Table S15). Therefore, gene modules identified from correlation-based co-expression networks also capture shared tissue-specificity.

This co-expression network analysis allowed us to classify the gene families into two categories, homogeneous and heterogeneous gene families, based on their patterns of expression across CNS tissues (Methods). A homogeneous gene 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 not co-expressed paralogs being removed from all tests on homogeneous family genes in the rest of the article) and thus 3,376 heterogeneous families (10,035 paralogs) (Supplemental 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} (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 (Supplemental 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 (Supplemental 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 2). 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. 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 2).

Table 2. 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
ySSD genes	Human-specific paralogous 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 = 1E-02 (Bonferroni correction for 5 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.

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 2). 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 2; Supplemental Materials Result S7). Similarly, SSD and ySSD genes were significantly enriched in genes coming from tissue-specific families (Supplemental Materials Table S17). Finally, we also analyzed 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 (Supplemental 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 (Supplemental 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) (Supplemental Materials Table S21); this supports the idea that paralog co-expression is favored by proximity along the genome. Moreover, we confirmed that the genomic proximity of duplicates was associated with recent SSDs and that the younger the SSD pair, the more the duplicate were found in tandem in the genome (Supplemental Materials Result S5). The tandem duplication explains 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 CNS 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. 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 (W.-H. 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) (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) (Supplemental Materials Table S18). We further focused on paralog categories and observed that, among paralogs, 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 slightly depleted in tissue-specific genes and were neither enriched in genes coming from homogeneous families or in human-specific paralogs (Supplemental Materials Table S18). In summary, brain disease genes are enriched in paralogs but not in WGDs in particular 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 between the different human CNS territories based on post-mortem human tissues analyzed by the GTEx consortium. Previous studies based on gene expression analysis between organs 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). Among paralogs, SSDs (Satake et al. 2012) and in particular ySSDs (Kryuchkova-Mostacci and Robinson-Rechavi 2016) seem to be more often tissue-specific than other paralogs when comparing tissues from different organs. However, 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 (Satake et al. 2012; Guschanski et al. 2017; Roux et al. 2017). In our study between the tissues that composed the human CNS, we observed that paralogs, especially ySSDs were more tissue-specific than other genes. 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 between brain regions is not only associated with the young age of duplication but also with the type of duplication (i.e. with SSD duplications). Our results, although apparently contradictory, do not question the known involvement of WGDs in brain-specific expression. Indeed, the fact that an 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. Thus, this broad expression of WGDs within brain regions facilitates the detection of their brain-specific expression when comparing several organs, while the analysis of gene expression between organs may not promote the detection of some ySSDs specific to human brain.

A previous study performed using gene expression profiles across mammalian organs established that most of tissue-specificity variance was explained by the expression level, in addition to the duplication status, with no significant contribution of the evolutionary time (Guschanski et al. 2017). Using multivariate linear models, we confirmed the major contribution of expression level and that of duplication status to tissue-specificity in CNS territories. The association with duplication status was more significant when we considered the maximal expression, which gives a better interpretation of gene abundance when studying the tissue-specificity than the average expression. Moreover, among paralogs, we found that the SSD duplication type explained also part of the tissue-specificity variance. Regarding the evolutionary time, low phyletic ages were also significantly associated with high tissue-specificity; a property potentially restricted to CNS tissues. Despite this global effect of the duplication age, we observed that tissue-specific expression did not seem to be associated with human-specific duplications, but rather with less recent ySSDs.

We then studied the gene family level of organization using gene co-expression network analysis of paralogs across CNS tissues. We showed that modules of co-expressed genes were able to identify clusters of paralogs with the same tissue-specificity. 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). 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 for other paralog pairs (Kryuchkova-Mostacci and Robinson-Rechavi 2016). We observed that it 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.

From the analysis of gene expression across human and mouse organs, Lan and Pritchard 2016 proposed a model for the retention of SSD duplicates appearing in mammals. In this model, pairs of young paralogs are often highly co-expressed probably because tandem duplicates are co-regulated by shared regulatory regions. In addition, this model is consistent with the dosage-sharing hypothesis in which down regulation of the duplicates, to match expression of the ancestral gene, is the first step enabling the initial survival of young duplicates (Lan and Pritchard 2016). Our analyses of ySSDs expression features between CNS territories seem to be concordant with this model, indeed ySSDs tend to be organized within small families of co-expressed genes and also weakly expressed in concordance with the sharing of the gene ancestral expression. Furthermore, our results in the CNS tissues seem to confirm that, after the initially high co-expression of SSD paralogs just after their duplication, they become more tissue-specific and less co-expressed in part through chromosomal rearrangement., suggesting a long term survival by sub-/neofunctionalization (Lan and Pritchard 2016). In the case of ySSDs tissue-specific in the same tissue, one of these duplicates might not preserve its coding potential in the long term and would lead to a pseudogene. This does not systematically imply its inactivation, indeed some transcribed pseudogenes associated with low abundance and high tissue-specificity may carry a regulatory function on their parental genes (Guo et al. 2014; Hezroni et al. 2017).

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 mutations accumulate preferentially in paralogs compared to singletons (Dickerson and Robertson 2012; W.-H. 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) potentially linked to their essentiality (Makino et al. 2009; Acharya and Ghosh 2016; Roux et al. 2017). 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, using our list of brain disease genes, we observed no enrichment in WGD or SSD duplications types.

In conclusion, our intra-organ exploration of paralogs suggests the major implication of young SSDs in tissue-specific expression between the different human CNS territories. It will be relevant to explore the expression patterns of these young SSDs between anatomic regions of other complex organs to determine whether or not they are solely associated with the nervous system.

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 (W.-H. 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 (W.-H. 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. 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; W.-H. 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. SSD with duplication date around the WGD events). 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; W.-H. Chen et al. 2013) (Supplemental Materials Fig. S3). The list of our paralogous gene pairs and gene families is given in the Supplemental Materials Table S1. The evolutionary annotation of paralogous genes is indicated in the Supplemental Materials Table S2. The list of singleton genes is given in the Supplemental 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 the 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) (Melé et al. 2015). The CNS tissue associated with each GTEx patient sample used in our study is indicated in the Supplemental 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 shorter 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 (Supplemental 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:

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

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 (Supplemental 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.

Brain disease genes

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 Marc 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.

DISCLOSURE DECLARATION

The authors declare that they have no competing interests.

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.

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.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. Nat. Rev. Genet. 14:645–660.

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.

Chen Y, Ding Y, Zhang Z, Wang W, Chen J-Y, Ueno N, Mao B. 2011. Evolution of vertebrate central nervous system is accompanied by novel expression changes of duplicate genes. J. Genet. Genomics 38:577–584.

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.

Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468:815–818.

Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE. 2018. Transcriptional fates of human-specific segmental duplications in brain. Genome Res. 28:1566–1576.

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.

Guo X, Lin M, Rockowitz S, Lachman HM, Zheng D. 2014. Characterization of human pseudogene-derived non-coding RNAs for functional potential. PloS One 9:e93972.

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.

Hezroni H, Ben-Tov Perry R, Meir Z, Housman G, Lubelsky Y, Ulitsky I. 2017. A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. Genome Biol. 18.

Holland LZ. 2009. Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. Nat. Rev. Neurosci. 10:736–746.

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.

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.

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.

Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. 2015. The human transcriptome across tissues and individuals. Science 348:660–665.

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

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.

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.

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

Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated Recruitment of New Brain Development Genes into the Human Genome. PLOS Biol. 9:e1001179.