Whole genome duplication drove cell type evolution in the vertebrate brain

Sebastian Shimeld
sebastian.shimeld@biology.ox.ac.uk
University of Oxford https://orcid.org/0000-0003-0195-7536
Yuanzhen Zhu
University of Oxford
Jiankai Wei
University of Oxford
Katia Jindrich
University of Oxford
Qiye Li
BGI Research https://orcid.org/0000-0002-5993-0312
Gunter Wagner
Yale University https://orcid.org/0000-0002-3097-002X
Peter Holland
University of Oxford https://orcid.org/0000-0003-1533-9376

Biological Sciences - Article

Keywords:

Posted Date: July 4th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6965966/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

The brains of vertebrates are more complex and have a wider diversity of cell types than those of their closest relatives. Whole-genome duplications (WGDs) occurred during early vertebrate evolution[1], but it remains unclear whether the resultant duplicate genes (ohnologues) facilitated cell type evolution. Using brain single-cell transcriptomes from four vertebrates – human, mouse, lizard, and lamprey – we find major cell type families are conserved with shared core transcription factors. If WGD was more important for cell type evolution than other types of gene duplication then we predict that cell type markers will be likely to be ohnologues, and that ohnologue pairs will be used in cell type-specific patterns. We show both predictions hold, demonstrating ohnologues play more prominent roles in cell type evolution than genes duplicated by other routes. By examining expression of paralogues across cell types and species, we show that expression changes have been mainly driven by dosage selection and subfunctionalization. We then show these processes boosted cellular diversity at different anatomical and cell type scales. Our findings demonstrate systematic and long-lasting effects that potentiated vertebrate brain cell type evolution for hundreds of millions of years following WGD.

Introduction

Susumu Ohno[2] hypothesised that vertebrates underwent whole genome duplications (WGDs) sometime in their ancestry, which generated additional genes important for vertebrate evolution. Evidence for this then came from gene mapping in human and mouse[3], followed by gene family comparisons between vertebrates, amphioxus and tunicates[4]; together these data resolved into the 2R hypothesis, which stated that two WGDs occurred in early jawed vertebrate evolution. Recent studies show the first WGD predated the separation of cyclostome and gnathostome lineages, while subsequent WGDs occurred independently in each lineage[1,5]. WGD is not the only way that genes can duplicate, however, and their impact must be distinguished from additional genetic material generated by the frequent and extensive duplications that occur by other routes, which we here call small-scale duplication (SSD)[6]. Most duplicated genes are lost following duplication, but retained genes may undergo complementary loss of function (subfunctionalization) and/or evolve new functions (neofunctionalization)[7]. Retained genes are also frequently co-opted into evolving gene regulatory networks (GRNs)[8,9], and this is proposed to drive new uses in the development and specification of tissues, organs, and cell types[9,10].

The origin and evolution of cell types has gained much attention in the past 20 years[11–15], with recent deployment of single-cell and single-nucleus RNA-sequencing (sc- and snRNA-seq) allowing systematic description of cell types based on their transcriptomes. An evolutionary definition of cell types has also been proposed defined by common descent regardless of form and function[12]. New cell types can evolve by duplication and divergence (the sister cell type model) which often come from anatomically and developmentally distinct regions, and cell type is an inherently hierarchical concept[12,16]. In many cases, as with gene paralogues, individual cell types are species/clade-specific[17]. These issues highlight the importance of investigating cell type evolution at both different hierarchical levels and across different regions of the body.

Vertebrates possess well-developed brains that enable quick and coordinated responses to environmental stimuli and facilitated vertebrate adaptation to diverse ecological niches. Compared to their closest invertebrate relatives - tunicates and amphioxus - vertebrate brains are highly regionalised and complex with many novel cell types. Previous studies[13,18–20] have demonstrated the similarities and differences of neural cell types within and between vertebrate species. However, ancestral repertoires of neural cell types and their core transcription factor (TF) programmes in vertebrates remain poorly understood. The roles of WGD and SSD paralogues in brain cell type evolution are still obscure.

In this study, we analysed four vertebrate whole brain single-cell transcriptomes and inferred the ancestral repertoires of neural cell type families. We identified core TF programmes for 11 conserved major cell type families in all vertebrates, and 15 restricted to jawed vertebrates. We systemically distinguish the roles of ohnologues and SSD paralogues in cell type evolution, with differences mainly due to differential retention after WGDs of TF and other genes involved in developmental regulation, and those encoding important effectors. We found dramatic shifts in expression patterns within paralogue families and argue subfunctionalization is more predominant than neofunctionalization, consistent with previous results[21–25]. We also found cell type non-specific dosage selection following duplication. Importantly, our conclusions apply across spatial and evolutionary scales, from neural cell families in major brain divisions to cell types in lineage-specific cerebellar nuclei. These findings suggest ohnologues act as ‘parts stores’, promoting cell type evolution during both early vertebrate evolution soon after WGD and over the subsequent hundreds of millions of years.

Results

Major cell type families are conserved in vertebrate brain

To compare vertebrate brain cell types, we surveyed scRNA and snRNA data from four species (human, mouse, lizard, and lamprey)[13,26–28]. No comparable whole brain single-cell data from fish are currently available. We included developmental data from one outgroup, the cephalochordate amphioxus. Although tunicates are more closely related to vertebrates, they are rapidly evolving and secondarily simplified, with few neurons at the developmental stage comparable to vertebrate brains[29–31]. Vertebrate data were filtered to retain only brain tissues at juvenile or adult stages, and low-quality cells were removed. To help balance the number of cells for cross-species integration and accommodate different proportions of neurons and glia, we randomly down-sampled the human and lizard atlases to 10^5 neurons and 10^5 non-neural cells while retaining the full brain atlases for mouse (67,937 neurons and 60,395 non-neuronal cells) and lamprey (18,166 neurons and 41,472 non-neuronal cells). Since atlases for different species were produced by different methods, we reanalysed each species with Self-Assembling-Manifold (SAM)[32], then performed iterative clustering to generate finer clusters of cells (Methods). This resulted in 241, 167, 202, and 141 clusters for human, mouse, lizard, and lamprey, respectively. We attributed clusters to cell types based on reference annotation, expression of markers (Supp. Table 1, Extended Data Fig. 1, and Methods)[33,34], and SAMap[35] mapping (see below). On average, 94% of clusters were reliably assigned to cell types in four species (Supp. Table 2).

Cell types can emerge through duplication and divergence within cell evolutionary lineages[12,16], conceptually similar to gene duplications forming lineage-specific paralogues[17]. This makes cross-species comparison unsuitable for assessing one-to-one cell type homology at high resolution in evolutionarily distant species[17]. The conservation of brain regionalisation and its molecular and developmental basis across vertebrates[36], instead led us to focus on cell type families at the same hierarchical layer. Cell type families can be defined as a set of cell types using the same regulatory programmes driving differentiation and identity[37] (for example, as defined by Core Regulatory Complexes (CoRCs)[12], Character Identity Networks (ChINs)[38], and terminal selectors[33]). To execute this, cell type family specific TFs were predicted and conserved TFs were used to define cell type families in vertebrates (Methods, Fig. 1c and Supp. Table 3). Homeodomain TFs were the most represented TF type in cell type specific TFs and the only enriched class in all four species (Hypergeometric test, adjusted $P < 0.01$; Extended Data Fig. 2a and Methods), supporting previous observations on their roles as terminal selectors in vertebrates and invertebrates[33,34,39].

To conduct cross-species comparisons, we separated neurons and non-neurons to generate two atlases for each species (Fig. 1b) and then performed SAMap[35], a specialized approach for stitching multiple species into a common space while accounting for in- and out-paralogues. In most cases, clusters were mapped to clusters within the same cell type families (Extended Data Fig. 2b,c), with around 76% and 91% of bidirectional linkages connecting clusters within the same neuronal and non-neuronal families, respectively. These cell type families were also represented by a conserved TF programme (Fig. 1c and Supp. Table 3). However, some discrepancies were noted. Lamprey erythrocytes mapped to jawed vertebrate oligodendrocytes with low mapping scores and a low number of supporting gene pairs (median = 0.19; 157 vs 906 gene-pairs on average for homologous astrocytes, and Methods). Lamprey rhombencephalon GABAergic neurons shared several TFs (*Bhlhe22, Lbx1, Lhx1/5, Neurodo1/2, En2, Hoxb3 and Tfap2a/b*) with jawed vertebrate counterparts. However, mapping relationships between lamprey rhombencephalon and jawed vertebrate rhombencephalon cells were unstable, with low mapping scores and associations with several other neuronal families. No lamprey rhombencephalon clusters expressed the excitatory neuron marker *Slc17a6/7*, and markers for the cerebellum (Purkinje cell marker: *Car8*; Cerebellar granule cell markers: *Math1* and *Barhl1*)[40] were not expressed in any clusters of lamprey hindbrain or midbrain (Extended Data Fig. 2d).

In summary, we found most major cell type families are shared across vertebrates except for oligodendrocytes and cerebellar cell types. These outcomes are consistent with several observations: the absence of myelination in lamprey brains[41], the lack of a well-defined cerebellum in lampreys[42,43], and findings from single-cell analyses comparing lamprey and mouse brains[28]. In addition, we identified a set of conserved TFs that can be used to define major cell type families across vertebrates, including lamprey, suggesting these cell type families evolved before the divergence of cyclostome and gnathostomes and that their core regulatory programmes are at least partially conserved.

WGD was more important than SSD in cell type diversity and evolution

To address how and to what extent ohnologues and SSD paralogues contribute to cell type diversity and evolution, we identified ohnologues, SSD paralogues, and differential expressed genes (DEGs or 'markers') across cell type families (Methods). We then asked whether ohnologues or SSD paralogues were DEGs at the cell type family level. A single-cell dataset from amphioxus at late neurula stage[45], filtered to retain only CNS cells, was used as a 'control' for analysing relationships between extensive SSD paralogues and DEGs. The Odds Ratios (ORs) in Fisher's exact test between SSD paralogues and DEGs were lower than 1 for most cell type families in all five species (Fig. 2a, OR < 1 between combined DEGs and SSD paralogues for all species, $P < 0.001$), suggesting SSD paralogues in chordates are not likely to be cell type markers. In contrast, the ORs between ohnologues and DEGs were above 1 for most cell type families in the four vertebrates (Fig. 2a, OR > 1, $P < 0.001$, between combined DEGs and ohnologues for all species). This shows ohnologues have a significant tendency to be markers. In line with this, the percentage of SSD paralogues in markers was generally lower than the background (except in mouse) while that for ohnologues was significantly higher than the background (one-sample t-test, $P < 0.0001$, Fig. 2b). This finding was reinforced by similar results in other tissues and at different levels of cell types (Extended Data Fig. 3a,b and see below). Additionally, TFs and putative target genes in cell-type-specific regulons detected by pySCENIC[46] also showed the same patterns with ohnologues and SSD paralogues (Extended Data Fig. 3c-f and Methods).

To understand why ohnologues were more associated with markers than SSD paralogues, we performed GO enrichment analysis (Methods) and found ohnologues and SSD paralogues were enriched in distinct categories of biological process (Fig. 2c and Extended Data Fig. 3g). Specifically, ohnologues were enriched in development, cell fate commitment, signalling, and neurotransmitter transport, while SSD paralogues were enriched for immune response and sensory perception in all species, matching with previous reports[6,47,48]. Consistently, we found that TFs, cofactors, transporters, etc were preferentially retained following WGDs (Fig. 2d, Extended Data Fig. 3h, and Methods). This suggests the positive association between ohnologues and markers is partly due to preferential retention of TFs, genes in developmental regulation, and some effectors.

Being associated with markers does not, in itself, prove that ohnologues contributed to the generation of new cell types. However, if this pattern reflects deployment of ohnologues to increasingly specialised cell types in vertebrate evolution, we expect *pairs* of ohnologue genes to be used in *different* cell type families or cell types (Fig. 2e). To test this, for each cell type family we calculated the number of paralogue families having a copy (or copies) as markers and the number of paralogues as markers. The ratio between the two was close to 1, significantly higher than expectation (One-sample t-test, $P < 0.001$, Fig. 2f). We found the same pattern at the cell type level (Extended Data Fig. 3i). This shows that if a paralogue is a marker for a specific cell type, other paralogues from the same family are less likely to be the markers for that cell type, regardless of duplication type and cell type granularity. While several studies have revealed an association between ohnologues or paralogues and cell type markers[15,24,25,49], this analysis distinguishes the association of ohnologues and SSD paralogues and clarifies their contribution to the evolution and diversification of cell types.

To compare paralogous genes with those showing conserved cell-type signals, we applied PCA on 'metagenes' or orthologues and variance decomposition to filter out genes with strong 'species signals'[50–53] or batch effects (Methods). In both analyses, batch- and species-specific effects dominated over cell-type signals, with ~70% of genes showing stronger species/batch contributions (Fig. 2g and Extended Data Fig. 4a-c). An additional cortical dataset from

mice[54] was incorporated and results suggested that variance from batch effects was greater than that from species differences (Extended Data Fig. 4d). As expected, metagenes with high contribution to cell type family signals were significantly associated with orthogroups containing ohnologues (Fisher's exact test, $P < 2.2e-16$, OR = 3.2).

Subfunctionalization is prevalent during cell type evolution

Following gene duplication, paralogous genes could become markers/expressed by different cell type families through subfunctionalization (splitting of ancestral roles) or neofunctionalization (evolution of new roles). We sought to determine which of these mechanisms plays the predominant role for both WGD and SSD paralogues. Expression was binarized as 'markers' (1) and 'not markers' (0), and the smallest groups of cell types deploying at least one paralogue as a marker in at least 3 out of 4 of the vertebrate species were considered as the ancestral state for vertebrates (Methods, Fig. 3a). We then calculated the changes in each paralogue family compared to the inferred ancestral state for each species (Methods). In paralogue families, approximately 78%, 21%, and < 1% of the total changes in marker usage can be attributed to subfunctionalization, neofunctionalization, and loss-of-function (loss of being markers for all copies in that cell type), respectively. Across paralogue families, SSD paralogue families, and ohnologue families, these proportions remain relatively consistent. The number of changes and number of changes per gene explained by subfunctionalization were significantly and consistently higher than explained by neofunctionalization for all species and duplication types (Fig. 3b,c). Similar patterns were found from inferring ancestral states and estimating sub-and neo-functionalization for amniotes only (Methods and Extended Data Fig. 5a,b). We also performed the above approaches on expression matrices binarized by Trinarization score[27] which displayed similar trends in vertebrates (Methods and Extended Data Fig. 5c,d). For example, lamprey's single *Slc17a6/7* was broadly expressed in excitatory neurons, while in jawed vertebrates, *Slc17a6* and *Slc17a7* underwent subfunctionalization, with *Slc17a7* predominant in telencephalon, and *Slc17a6* in di-/mesencephalon, and both in rhombencephalon excitatory neurons (Extended Data Fig. 5e). These results indicate that the deployment of paralogues to increasingly specialised cell types in vertebrate evolution after gene duplication was mainly driven by subfunctionalization. These results support the duplication-degeneration-complementation (DDC) model[7] and match previous studies based on bulk transcriptomics and one study based on single-cell transcriptome[22–24].

Given the changes within paralogue families, we next asked whether expression of genes in paralogue families shifts to similar degrees across species. We defined the expression domain based on Trinarization score[27] at homologous cell type family level and then calculated the average expression divergence (dT) among paralogues in each orthogroup for each species (Methods). This revealed most paralogues extensively diverged in both species and some exhibited shifts mainly in one species on the pairwise comparison (Fig. 3d and Extended Data Fig. 6a). For example, the *Tbr1* subfamily of T-box genes (dT = 1 for human, mouse and lizard, and dT = 0.67 for lamprey) duplicated through WGDs at the base of vertebrates, giving rise to *Tbr1*, *Eomes*, and *Tbx21* in jawed vertebrates (Extended Data Fig. 6b and confirmed elsewhere[1]). In our dataset, *Tbr1* and its lamprey homologues are exclusively expressed in glutamatergic neurons of the telencephalon, while *Eomes* is expressed in rhombencephalon glutamatergic neurons in amniotes (Fig. 3e). We observed a decrease of cells expressing *Tbr1* from lamprey to mammals, alongside cortical expansion and olfactory bulb reduction in lizards and mammals, reflecting ohnologues adaptations and neuronal specialization across vertebrates.

Dosage selection is prevalent across cell types

Studies using bulk transcriptomes have revealed cases of gene expression changes following gene duplication[55,56]. This could be explained by the gene balance hypothesis, which proposes that high expression of duplicated genes can be selectively disadvantageous due to stoichiometry effects[57,58]. We tested if this extended to cell type level, potentially contributing to the route to subfunctionalization. This showed that most (>70%) WGD and SSD paralogue families contained at least one copy that significantly differed from others with respect to expression level or percentage of expressing cells (Friedman test, $P < 0.01$, Extended Data Fig. 7a,b). We next limited analysis to gene families with two copies to avoid multiple comparisons, and counted the number of paralogue families having a significantly dominant copy as assessed by expression level and by percentage of expressing cells. Over two-thirds (66-80%) of these paralogue families, whether derived by WGD or SSD, have a significantly dominant copy in a pervasive way, not specific to cell types (Fig. 3f and Methods). Hence, after gene duplication, one of the genes often becomes dominant over multiple cell types.

We then asked whether different species use the same genes as the dominant copy within each gene family. We found humans and mice had the highest similarity compared to other species pairs, reflecting their closer phylogenetic relationship (Fig. 3g). Although the total number of SSD paralogue families was higher than ohnologue families, the number of ohnologue families sharing a dominant copy was consistently higher than SSD families over all species comparisons (Fig. 3g). For example, *Pax6* was highly expressed in astrocytes, ReExc, and several other cell types, whereas *Pax4* showed limited expression (Fig. 3h and Extended Data Fig. 7c). *Pax4* has also been independently lost across multiple vertebrate lineages, including lamprey, hagfish, and bearded dragon (Extended Data Fig. 7d), likely due to its limited expression domain and relaxed selection[59]. This general trend can be explained by either stronger dosage selection on ohnologues and/or if SSDs emerged in a specific lineage and are not shared by species analysed. The high degree of conservation of dominant copies across vertebrates suggests that dosage selection occurred soon after duplication (especially WGD) irrespective of cell type and prior to divergence of lineages studied; this could be a result of selection following an immediate transcriptional response after genome duplication[7,55,58]. This allows genes to be retained sufficiently long before subfunctionalization and/or neofunctionalization[7,60]. Our case studies also have shown that many paralogues (e.g., *Ppp2ca*, *Ppp2cb*; *Ctbp1*, *Ctbp2*; *Atfb6*, *Atfb6b*; *Strada*, *Stradb*) are used differently by different species (Extended Data Fig. 7e).

WGD roles in the regional identity and evolution of neuronal and glial cells

A key question is whether different neural cell type families, such as glutamatergic neurons in different brain divisions in a species, evolved from a common ancestral population (are monophyletic) or arose independently (are polyphyletic), given that their TFs of top regulons were largely not shared (Extended Data Fig. 9a,b). Brain regionalisation occurs early in development and the transcriptomic differences across brain regions correspond with progenitor cell position,

which has lasting effects on cell specification and differentiation[61–63]. Interestingly, a recent study suggest that neurons and astrocytes share some region-specific markers[64] but this is limited to neocortex and thalamus.

To explore how WGD contributed to regionalisation and the broader complexity of neural cell types, we first examined macroglia. Macroglia, including astrocytes, ependymal cells, and oligodendrocytes, make up the majority of glial cells in the brain and, like neurons, originate from neuroectoderm[65]. We first identified subtypes based on known markers (Supp. Table 4), dissection locations, and annotation from the published reference atlases[13,27,54] (Extended Data Fig. 8a,b). Astrocyte was the most diversified macroglia family, with strong regional variance, (Fig. 4a and Extended Data Fig. 8c), compared to oligodendrocytes and ependymal cells (Extended Data Fig. 8d,e)[27,54,66]. An exception was observed in the mouse dataset, which displayed higher diversity in all macroglia subtypes, because of batch effects confounded with biological variance in their experimental design. Ohnologues were significantly associated with DEGs of astrocytes and oligodendrocyte subtypes, respectively (Fig. 4b and Extended Data Fig. 8f-h). However, DEGs within ependymal cells only showed significant association with ohnologues in lizard and lamprey (Extended Data Fig. 8i,j), likely due to their low representation (and therefore low number of markers detected) in the human and mouse data. Regional variance of astrocytes prompted us to perform GO enrichment analysis on DEGs, which revealed their differences in development and functions (Extended Data Fig. 9c).

Using variance decomposition, we systematically identified genes that contribute to regional identity and to astrocyte-neuron variance (Fig. 4c, Supp. Table 5, and Methods). To do this, we compared GABAergic neurons with astrocytes and glutamatergic neurons with astrocytes. Regional genes significantly overlapped across cell types (Fig. 4d and Extended Data Fig. 9d) and relevant orthogroups were significantly shared across species (Fig. 4e), suggesting at least part of the region-specific programmes were shared across neural cell type families and conserved across vertebrates. These genes were significantly enriched in regionalisation, brain development, and cell specification (Extended Data Fig. 9e). Among these, we identified several key genes implicated in brain development and cell specification, including *Foxg1*, *Emx1*, *Fezf2*, and *Prox1* (telencephalon)[67–70]; *Tcf7l2*, and *Six3* (diencephalon)[71,72]; *Otx2*, *En1/2*, and *Pax7* (mesencephalon)[73–75]; *En1/2*, *Pax3*, *Zic1/2*, and *Neurod1/2* (rhombencephalon)[74,76,77]. Notably, the average number of copies (2.5 to 3.4) in these regionalisation orthogroups were significantly larger than the average size of orthogroups for each species (1.3 to 1.5, Wilcoxon signed-rank test, $P < 0.0001$). In addition, genes in regionalisation orthogroups were more strongly associated with ohnologues than SSD paralogues (Fig. 4f; except in lizard with similar ORs). As previously observed, these genes and their ohnologue pairs underwent significant expression shifts following WGDs (Fig. 4g). These findings indicate that some regional programmes for major cell type families likely evolved before the common ancestor of vertebrates, were preferentially retained and underwent expression shifts following WGDs to contribute cell type family diversity across brain divisions.

Ohnologues contribute to cell type evolution long after WGDs

To test whether our findings hold true in relation to cell types that emerged at least 150 million years after WGD[1,44], we leveraged recent scRNA atlases of human, mice, and chicken cerebellar nuclei (CN)[78]. During vertebrate evolution an archetypal CN including its conserved combination of cell subtypes has been shown to have duplicated (Fig. 5a)[78]. Specifically lamprey lacks CN, while cartilaginous fishes and amphibians have one CN pair, reptiles and birds two pairs, and mammals three pairs[79,80]. We focused on excitatory neurons within the CN since they are regionally variant and largely confined to cytoarchitecturally defined subnuclei[78,81]. We performed the same approach as above to explore relationships between ohnologues/SSD paralogues and DEGs, finding the same patterns as were found at the cell type family and cluster level (Fig. 5b and Extended Data Fig. 10a). Consistent with this, 10/20 experimentally validated markers of CN excitatory neurons[78] were ohnologues (Supp. Table 6). DEGs were enriched in axonogenesis, axon guidance, migration, and synaptic organization, reflecting their functional diversity across subnuclei and between neuron classes (Fig. 5c and Extended Data Fig. 10b,c).

To understand how ohnologues might be involved in the duplication and divergence of excitatory neurons in CN, we used hierarchical clustering to build a cell type dendrogram, binarized expression, and inferred candidate genes involved in the key branching events (Methods). Previously identified class A or B excitatory neurons[78] were generally clustered together in the dendrogram for all three species (Fig. 5d and Extended Data Fig. 10d,e). We identified several important TF genes, including *Lmx1a/b*, *Pax6*, *Tbr1*, *Lhx9*, and *Prox1*. These genes are involved in the chronological sequence of CN glutamatergic neuron development[82,83] and they have ohnologue pairs with distinct expression patterns across cell subtypes (Fig. 5e and Supp. Table 7). *Lmx1a*, *Pax6*, and *Tbr1* were primarily expressed in the medial nucleus, while *Lhx9* was mainly expressed in lateral and interposed nucleus (Fig. 5e), matching with previous reports[82,83]. We also found that many ohnologues encoding axon guidance molecules (Robo/Slit, Unc5/Netrin, and Unc6/Netrin) that control axonal trajectories in the nervous system were differentially expressed (Fig. 5f and Extended Data Fig. 10f,g), reflecting the difference of afferent projection to CN and efferent projections from CN[82,84,85]. Clusters of Interposed X (IntX, a spatially isolated region in chicken) are considered to have no directly homologous nuclei in mice[78] and these were clustered with medial nuclei within chicken (Extended Data Fig. 10e). We found several TFs which may be related to IntX specification (Extended Data Fig. 10h) and many were ohnologues. To summarize, these findings highlight that ohnologues derived from WGD over 500 million years ago were still involved in the duplication and divergence of potential sister cell types within CN long after WGD[78,86].

Discussion

Cell types are fundamental biological units and show both deep conservation and frequent innovation in animal evolution. While a small number of studies have linked gene duplication to specific phenotypic changes (for example, Tbx4/5 in limb development[87]; OPN1SW/OPN1LW in cones/rods diversification[88,89]; r-/c-opsin, Gqα/Giα in the phototransduction of rhabdomeric and ciliary photoreceptors[90]), the extent to which paralogues contribute to cell type evolution and diversification has not been systematically studied.

In this study, we leveraged single-cell transcriptomics to compare vertebrate brain cell types across multiple evolutionary scales, both within and across species. Our results show most major neural cell type families predate the cyclostome-gnathostome split. We show WGD contributed to (and was not simply associated with) cell type evolution and that this was to a significantly greater degree than SSD. This occurred primarily through preferential retention of TFs, other developmental genes, and key effectors such as neurotransmitter transporters. Subfunctionalization and dosage selection emerged as the dominant outcomes for both WGD- and SSD-derived paralogues. We demonstrate WGD and its consequences enhanced cellular diversity across cell types in different brain regions and at finer anatomical and phylogenetic levels, through influencing regional, developmental, and functional programmes.

These results demonstrate that WGD events early in vertebrate evolution had a pivotal role in vertebrate neural cell type evolution and underpins vertebrate brain complexity. We also propose that this impact persisted for hundreds of millions of years after WGD, with paralogues still underpinning cell type diversity in much more recent evolutionary changes in amniote cerebellar nuclei. We propose that WGD paralogues have potentiated vertebrate neural evolution, and by extension likely also the evolution of cell type diversity in other tissues as we show the same pattern in cell types of human eye and lung (Extended Data Fig. 3 a,b).

Our analyses of expression patterns show subfunctionalization is the predominant pattern of expression evolution in paralogues. This aligns with several studies[22–24] but contrasts with one report[21], although the methods deployed in that study have been questioned by others[23]. At either expression level or percentage of cells expressing, we observed widespread dosage selection in a manner shared by multiple cell types. This explains the cellular basis for previous findings from bulk transcriptome analysis, which noted dominant paralogue expression at tissue level[55]. Although DDC and the gene balance hypothesis offer explanations, alternative proposals also exist[91], some of which partially overlap with these models. Comparison among retained paralogues by either duplication mechanism revealed similar patterns in both expression specificity and expression levels. Novel cell types (new CoRC programmes) and cell type with novel functions (new downstream effector patterns) are two different things[92]. Our results (Fig. 2c and Extended Data Fig. 3g) show the importance of ohnologue contribution to both cell type identity and functionality.

It could be argued that subfunctionalization is simply using duplicated genes to do the job that one gene did before duplication, and therefore duplication may have a limited role in innovation, for example, as proposed for Tfap2, SoxE, Twist1/2, and Gata1/2/3 in the neural crest[1]. On the contrary, the emergence of a regulatory motif driving Tbx4/5 expression in the lateral plate mesoderm facilitated the evolutionary origin of limbs, and then subsequent WGD and functional divergence of Tbx5 and Tbx4 separated forelimbs and hindlimbs[87,93]. In addition, subfunctionalization helps to preserve duplication for neofunctionalization[60]. However, our analyses show this is too simplistic a view. We show that from large scale (vertebrate brain divisions) to fine scale (cerebellar nuclei) ohnologues are involved in cell type evolution. This shows WGD was not only impactful for early vertebrate evolution but acted as a genetic reservoir for long-term cell type evolution. We propose that this will be true for other vertebrate tissues and organs, a prediction that can be tested as additional single-cell datasets across vertebrate phylogeny are developed. Our definition of conserved cell type families and their core TFs will also be a reference framework for comparison to cell types in cephalochordates, tunicates and other invertebrates, helping to reveal the origin of vertebrate brain cell type families.

Methods

scRNA and snRNA atlas collection, filtering, and preprocessing

Cell atlases were retrieved from references[13,26–28,45]. Low-quality cells in the human atlas were further filtered based on nCount (UMI) < 400. Low-quality cells in the other atlases were already filtered. To focus on neural cells in brain, vertebrate datasets were filtered to retain only brain tissues at juvenile or adult stages. To help balance the number of cells for cross-species integration and accommodate different proportions of neurons and glia, we randomly down-sampled the human and lizard atlases to 10^5 neurons and 10^5 non-neurons while retaining the full brain atlases for mouse (67,937 neurons and 60,395 non-neuronal cells) and lamprey (18,166 neurons and 41,472 non-neuronal cells). Only protein-coding genes were retained for downstream analyses.

Since the original atlases were generated using different pipelines, we applied a standardized preprocessing approach to ensure consistency. We performed self-assembling manifold (SAM) analysis on each individual atlas by directly invoking the SAMAP function from the SAMap package[35]. Specifically, UMI counts from each cell were firstly normalised to give the median total count per cell, then log2-transformed followed by SAM function with following parameters (preprocessing="StandardScaler", npcs=100, weight_PCs=False, k=20, n_genes=3000, weight_mode='rms'). The anndata objects were then converted to Seurat format for downstream clustering.

Iterative clustering and annotation

To find good quality and high resolution cell clusters in the SAM pre-processed atlases, we performed hierarchical and iterative clustering using scrattch.hicat and scrattch.bigcat packages[94,95] from the Allen Institute. Raw counts (UMI) were firstly normalised using the cpm function provided in the above packages, followed by log2 transformation with a pseudo-count added to prevent $\log_2(0)$. Cells were initially classified into broad groups and hierarchically clustered based on the expression of highly variable genes, PCA, and Jaccard–Louvain clustering. Clustering was performed iteratively within each group using the iter_clust function, continuing until no further subclusters satisfied predefined thresholds for the number of differentially expressed genes (DEGs) or minimum cluster size. As our analysis did not aim to resolve extremely fine-scale cell types, we applied more relaxed parameters than those typically used with this method. DEG thresholds were defined via the de_param settings: padj.th = 0.05, q1.th = 0.4, q2.th = NULL, q.diff.th = 0.5, de.score.th = 100, min.cells = 100, and min.genes = 6. Dimensionality reduction and clustering parameters were specified as follows: dim.method = "pca", max.dim = 80, method = "louvain". Minimum cluster sizes were set via split.size as 800, 500, 500, and 500 for human, mouse, lizard, and lamprey datasets, respectively.

Clusters were then checked and merged at the end of iteration to make sure they were separable with scrattch.bigcat::merge_cl. The relevant scripts have been uploaded to GitHub (https://github.com/DiracZhu1998/WGD2celltype_evolution). We next confirmed and refined the annotation of individual atlases by examining the expression of canonical markers (Supp. Table 1 and Extended Data Fig. 1), reference annotation in our clustering, and their main dissection locations (Supp. Table 2).

Identifying gene relationships: orthologues, paralogues, ohnologues, and SSD paralogues

To identify gene relationships, we first collected genome assemblies and gene annotation files for the species listed in Supplementary Table 8. For each protein-coding gene, only the transcript with the longest coding sequences (CDS) was retained. CDSs were extracted from genomes based on gene annotation files and translated into protein with in-house scripts. We then performed phylogenetic orthology inference with OrthoFinder (v2.5.5)[96,97]. The species tree inferred from orthogroups matched with references (data not shown). Orthologues were identified based on OrthoFinder output, applying a reciprocal best hit (RBH) criterion. (In-)paralogues were defined as duplicated genes in the same orthogroup for each species. Ohnologues were identified based on Ohnologs-v2.0[98] (details in https://github.com/SinghLabUCSF/Ohnologs-v2.0) with updated genome and annotations (additional information and species used for ohnologue detection were listed in Supplementary Table 9). Due to limited sampling of jawless vertebrates and the lack of separate ohnologue inference for jawed and jawless lineages, ohnologue identification in lamprey remains challenging. Nevertheless, recent studies[5,99] suggest that the second round of whole genome duplication (WGD) in jawed vertebrates likely involved interspecific hybridization, resulting in asymmetric gene loss. Specifically, genes from the 'alpha' parental lineage are ~4 times more likely to be retained than those from the 'beta' lineage (based on results in chicken; see: https://raw.githubusercontent.com/fmarletaz/hagfish/refs/heads/main/Paralogons/Vert_Evt_OGrrA.txt). Although the second WGD in jawless vertebrates remains less well understood, we considered it appropriate to infer ohnologues across all vertebrates collectively at this stage, given the extensive retention of genes from the first WGD in the alpha lineage in gnathostomes. To assess the robustness of our ohnologue predictions, we compared our results to the Ohnologs v2 database (http://ohnologs.curie.fr), finding that 80% of human and 66% of mouse ohnologues in our dataset were also present in the database. Small-scale duplication (SSD) paralogues were defined as paralogues rather than ohnologues.

Atlas integration and cross-species mapping

Homologous gene relationships for initial weighting gene-gene graph with cross-species edges in SAMap were generated by blast on protein-coding genes using SAMap map_gene.sh. We then performed cross-species mapping using the SAMap run function with 5 iterations, edge weight calculated and updated by Pearson correlation (hom_edge_mode = "pearson"), and 20 cross-species edges per cell (crossK = 20). Mutual nearest neighbourhoods were independently calculated between each pair of species (pairwise=True). The alignment scores between cell types across species were calculated using get_mapping_scores from SAMap. We next used the GenePairFinder function to identify gene pairs (genes between species) that positively contributed to cross-species correlation between cell types and were differentially expressed in respective atlases.

Identification of cell-type-specific transcription factors and conserved sets for cell type families

To identify transcription factor (TF) coding genes for each species, we employed DeepTFactor[100], a deep learning-based tool optimized for TF prediction. Then, cell-type specific TFs were identified for each major cell type family using NS-Forest v4.0[101], a method designed to identify minimum combinations of necessary and sufficient marker genes for distinguishing different cell types. This employs a random forest algorithm on pre-selected genes by binary scoring, a measurement of binary expression (specificity) for a gene. For our analysis, we utilized the binary score to rank TF specificity and extracted the top 30 TFs with the highest scores as cell-type specific TFs, using the nsforesting.NSForest function and parameters (gene_selection = "BinaryFirst_high", n_top_genes = 30, n_binary_genes = 30, n_trees = 1500).

We next assigned cell-type specific TFs to individual orthogroups and defined an orthogroup as a conserved TF orthogroup for cell type families if at least three (out of four) species contained these TFs. This approach identified 81 orthogroups (Supp. Table 3-1), which we manually reviewed for expression patterns across species and summarized in Supp. Table 3-2 with supporting references. The orthogroups generated by OrthoFinder were also uploaded to GitHub (https://github.com/DiracZhu1998/WGD2celltype_evolution).

Classification of TFs and TF enrichment analysis

TF families were downloaded from AnimalTFDB v4.0[102] (http://guolab.wchscu.cn/AnimalTFDB4/#/Download). To classify TF families in species not represented in the database, we assigned TF family classifications at the orthogroup level using OrthoFinder output. If any human gene within an orthogroup was annotated with a specific TF family, we classified the entire orthogroup under that family. The high overlap (>90%, not shown) in classifications based on model organisms (human, mouse, and zebrafish) validated the robustness of this approach. The enrichment of TF class was assessed by Hypergeometric test using the stats::phyper function for each TF class in each species. *P*-values were further adjusted using p.adjust(method = "fdr") from the stats package.

Identification of marker genes at cell type family and cluster level

Marker genes were identified for each species using the FindAllMarkers function of Seurat (v5.0.0)[103] with the Wilcoxon Rank Sum test (min.pct = 0.01, logfc.threshold = 0.58, test.use = 'wilcox', only.pos = TRUE) at both the cell type family level and cluster level. For related downstream analysis, only marker genes with FDR < 0.01 were used. Since a few studies[104,105] previously questioned the quality of Seurat 'wilcox' output, we also identified markers using FindAllMarkers with ROC analysis (test.use = "roc", only.pos = TRUE), leading to the same conclusions (data not shown).

Gene regulatory network analysis

We performed gene regulatory network (GRN) analysis and identified regulons by pySCENIC[46,106]. To reduce noise introduced by imbalance in the number of cells in each major cell type family, we first randomly subset 2,000 cells for each major cell type family. To reduce noise of lowly expressed genes, we filtered genes with less than 0.5% cells expressing and filtered genes with low total UMI (equivalent to 1 UMI detected in 1% cells).

The grn command in pySCENIC was employed to infer gene-gene co-expression relationships between TFs and their potential target genes with grnboost2 algorithm. This returned an adjacency edge list with TF, its potential target gene, and an associated importance score. The adjacency edge list was then used as input for the ctx command to identify regulons, each consisting of a TF and its target genes enriched for the TF's binding motifs. Human and mouse TF lists were downloaded from the link (https://resources.aertslab.org/cistarget/tf_lists/). The ctx command utilized a motif annotation database and ranking databases, both of which were downloaded from Aerts Lab's cistarget resources (motif ranking datasets: https://resources.aertslab.org/cistarget/databases/old/homo_sapiens/hg38/refseq_r80/mc9nr/gene_based;https://resources.aertslab.org/cistarget/databa and motif annotation files: https://resources.aertslab.org/cistarget/motif2tf/). Next, the aucell command was used to compute regulon activity scores for each major cell type family and regulon specificity score (RSS) was calculated by regulon_specificity_scores function. The top regulons for each cell type were selected based on RSS.

Gene ontology (GO) annotations and enrichment analyses

Due to the lack of recent updates to the GO annotation of lizard (*Pogona vitticeps)*, lamprey (*Petromyzon marinus)*,and amphioxus (*Branchiostoma floridae)*, we decided to re-annotate the GO annotations for these three species. GO annotations for the protein-coding genes of model organisms (*Danio rerio*, *Mus musculus*, and *Homo sapiens*) were downloaded from Ensembl through BioMart. GO terms were associated with protein-coding genes from *Pogona vitticeps*, *Petromyzon marinus*, and *Branchiostoma floridae* according to their one-to-one orthologues in *Homo sapiens*, *Mus musculus*, and *Danio rerio* in an order of priority (human > mouse > zebrafish). The lizard, lamprey, and amphioxus genes that could not be annotated by the above method were then BLAST searched to the UniProtKB database[107] (release-2024_03) using BLAST (2.9.0+)[108] with parameters (-evalue 1e-8). The best hit for each query was selected based on bit score and its corresponding GO terms (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/goa_uniprot_all.gaf.gz) assigned to the respective query. In total, we annotated nearly all protein-coding genes for lizard and over 70% of lamprey and amphioxus. This is higher than GO annotation in Ensembl for another amphioxus species, *Branchiostoma lanceolatum*, where more than half of the protein-coding genes were not annotated.

The datasets of GO annotations for lizard, lamprey, and amphioxus were built using makeOrgPackage function from the AnnotationForge package[109]. The dataset packages for human and mouse were retrieved from Bioconductor (v3.20) at org.Hs.eg.db[110] and org.Mm.eg.db[111], respectively. GO enrichment analysis was performed with clusterProfiler[112] enrichGO with Benjamini-Hochberg (BH) adjustment and cut-off = 0.05. Only protein-coding genes in our datasets were used as background genes in GO enrichment analysis. The redundancy of enriched terms was filtered by simplify() with the parameters (cutoff=0.7, by="p.adjust", select_fun=min).

Classification of protein class and over-representation analysis

To investigate protein class in ohnologues and SSD paralogues, we used "Functional classification viewed in graphic charts" with bar plot in PANTHER database[113] (v19.0: https://www.pantherdb.org). Over-representation analysis was conducted using the 'Statistical Overrepresentation Test' in PANTHER website, with protein-coding genes in our datasets as the background. Fisher's exact test was applied with false discovery rate (FDR) correction to assess statistical significance. Due to the absence of corresponding data for lizard, lamprey, and amphioxus in PANTHER, this analysis was limited to human and mouse.

Variance decomposition

To assess the contribution of cell type family and (species signals plus batch effect), we used a linear mixed model (LMM) on the normalised pseudobulk expression for each 'metagene' and orthologue; to evaluate variance properly, this approach can only use genes shared across species. The one-to-one-to-one orthologue relationships were retrieved from OrthoFinder output. The 'metagenes' were calculated by the sum of UMI of copies in each orthogroup for each species. The pseudobulk expression was calculated by the sum of gene counts (UMI) for each orthologue/metagene in individual cell type families. To balance cell number differences, 2,000 cells were randomly selected for each cell-type family in each species. We then used DESeq2[114] to normalise library size and performed LMM for each gene with the lme4 package[115]. The restricted maximum likelihood (REML) estimators for the random effects of cell type family, species/batch and residual variance were normalised by their sum to give the variance components (see Fig. 2g). Genes that contributed more than half of the total variance to cell type family were considered as genes highly contributing to cell type signal shared in vertebrates.

Subfunctionalization and neofunctionalization

Gene relationships were assessed based on the above OrthoFinder output. To exclude high copy-number SSDs, only orthogroups with less than five gene copies for each species were retained. To infer ancestral states without considering gene losses, we retained orthogroups with at least one copy for each species. This allows us to do cross-species comparison directly at orthogroup level since at least one gene for each species is present for each orthogroup. An orthogroup was classified as an ohnologue orthogroup if it contained at least one pair of ohnologues in any of the four species, resulting in 2,201 ohnologue orthogroup. The same approach was applied to SSDs, and 3,696 SSD paralogue orthogroup were identified. Some orthogroup (1,655) were considered as both ohnologue orthogroup and SSD paralogue orthogroup due to lineage-specific SSDs.

To predict ancestral states, we next binarized expression matrices using two separate approaches: based on whether genes were classified as markers, and based on expressed or not determined by the Trinarization score. For the second approach, a gene was considered expressed if it was estimated to be present in at least 10% of the cells, with a posterior error probability of no more than 5%. Details of Trinarization score are described here[27]. We inferred

ancestral states for vertebrate and amniote lineages across homologous cell type families. For vertebrate ancestral states, a gene family was considered expressed (state = 1) in a given cell type family if at least three out of four species utilized one or more copies from that paralogue family in that cell type family. The same criterion was applied for predicting amniote ancestral states, where expression (state = 1) was assigned if at least two out of three species being considered. The extent of subfunctionalization and neofunctionalization within gene families was quantified by comparing the binarized expression patterns of individual genes to the inferred ancestral states. Specifically, the difference between the binarized expression of a gene and orthogroup ancestral state was computed, in which a value of -1 indicated subfunctionalization (unless all copies in that species are -1 which suggest loss-of-function, see Fig. 3a) and a value of +1 denoted neofunctionalization.

Expression divergence (dT) among paralogues

Gene relationships were based on the above OrthoFinder output. To exclude high copy-number SSDs, only orthogroups with less than five copies for each species were retained. To do the pairwise comparison in shared orthogroups, orthogroups with at least one copy for any of four species were further retained. Paralogue orthogroups were then defined as orthogroups that included at least one pair of paralogue genes for a species. For a combination of paralogues in orthogroup, we calculated expression divergence (dT) based on a simple formula[116] for each species separately. Specifically, dT was first calculated for each pair of paralogues by the fractional difference between the number of cell families expressing either paralogue (N_{either}) and the number of cell families expressing both paralogues (N_{both}), relative to N_{either}. dT was next averaged within paralogue orthogroup (when there was more than one pair of paralogues) for each species.

Cell type non-specific dominant expression

To compare gene expression level between paralogues for each species, we first calculated the average normalised expression levels for each gene using the Seurat::AverageExpression function[103], and the proportion of cells expressing specific genes (pct. exp.) with an expression count greater than 0. These calculations were performed at both the cell type family and cluster levels. Next, we use the igraph package[117] to construct ohnologue and SSD paralogue families based on previously identified ohnologue pairs and SSD paralogue pairs, respectively. We tested the expression levels and pct. exp. values using the friedman_test function from the rstatix package[118], as the data did not follow a normal distribution. For species pairwise comparisons in individual ohnologue and SSD paralogue families, we applied the rstatix::wilcox_test function with Bonferroni adjusted p-value to identify the highest-expressed (dominant) copy within each family and search for whether their orthologues are the dominant copy in another species. One-to-one orthologue relationships underpinning this were derived from above OrthoFinder results.

Analyses in cerebellar nuclei

We downloaded human, mouse, and chicken cerebellar nuclei (CN) datasets from Kebschull's work[78]. These datasets were further filtered to retain only protein-coding genes and excitatory neurons, which show higher regional variants than inhibitory neurons in CN. We detected DEGs as described above, using FindAllMarkers with parameters (wilcox, only.pos = TRUE) and only DEGs with log2FC > 0.58 and p_val_adj < 0.01 were retained. Scaled average expression was calculated by Seurat AverageExpression[103] and then normalized by dividing each gene's expression by its mean among different cell types. The transcriptomic dendrogram was calculated based of scaled average expression of DEGs using pvclust with the following parameters: Spearman correlation-based distance 1 – cor() and average linkage with 1000 bootstrap replicates. Expression profiles were binarized using the Trinarization score, a gene was considered expressed if it was estimated to be present in at least 20% of the cells, with a posterior error probability of no more than 5%. The binarized data were then used to infer ancestral states based on dendrograms using Maximum Parsimony (MP). Specifically, we utilized the phangorn package[119], converted the binarized expression into phyDat format, and applied the ancestral.pars function with the accelerated transform (ACCTRAN) approach to estimate ancestral states and return probability. Genes in each ancestral node were classified as "expressed" if the probability exceeded 0.5, and "not expressed" otherwise. Finally, we identified gene expression "gain and loss events" along branching points in the tree to identify candidate genes that might be involved in the cell type duplication and divergence.

Declarations

Acknowledgments

We are grateful to Yichen Dai and Guang Li for providing amphioxus single-cell dataset prior to publication of their paper. We are also grateful to Guang Li and Ferdinand Marlétaz for discussion about ohnologue detection and importance of WGD, and Jonathan Bard for constructive suggestions. This work was funded by the National Natural Science Foundation of China (Grant No. 32370461), the Science & Technology Innovation Project of Laoshan Laboratory (LSKJ202203001), the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (No. 895927), and China Oxford Scholarship Fund.

Author Contributions

Y.Z. and S.M.S. conceived the study. Y.Z. and K.J. standardized the dataset. Y.Z. and J.W. performed quality control, integration, clustering, and annotation. Q.L. contributed through method development and implementation. Y.Z. conducted cross-species analyses, paralogue analyses and downstream case studies. G.P.W. and P.W.H.H contributed to evaluation and interpretation of cell type evolution and gene duplication analyses. Y.Z. and S.M.S wrote the manuscript with input from all authors. Y.Z, J.W., K.J., Q.L., G.P.W., P.W.H.H., and S.M.S revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The reference genome, gene models, functional annotations of protein-coding genes, full marker gene list of each cell cluster, and important intermediate files are deposited in the figshare repository under the link https://doi.org/10.6084/m9.figshare.29327111 (publicly available upon acceptance of the manuscript). All in-house scripts are uploaded in the GitHub repository under the link https://github.com/DiracZhu1998/WGD2celltype_evolution/tree/main (publicly available upon acceptance of the manuscript).

References

1. Marlétaz, F. *et al.* The hagfish genome and the evolution of vertebrates. *Nature* **627**, 811–820 (2024).
2. Ohno, S. *Evolution by Gene Duplication*. (Springer Berlin Heidelberg, Berlin, Heidelberg, 1970). doi:10.1007/978-3-642-86659-3.
3. Lundin, L. G. Evolution of the Vertebrate Genome as Reflected in Paralogous Chromosomal Regions in Man and the House Mouse. *Genomics* **16**, 1–19 (1993).
4. Holland, P. W. H., Garcia-Fernàndez, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Development* **1994**, 125–133 (1994).
5. Yu, D. *et al.* Hagfish genome elucidates vertebrate whole-genome duplication events and their evolutionary consequences. *Nat Ecol Evol* (2024) doi:10.1038/s41559-023-02299-z.
6. Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
7. Force, A. *et al.* Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics* **151**, 1531–1545 (1999).
8. Teichmann, S. A. & Babu, M. M. Gene regulatory network growth by duplication. *Nat Genet* **36**, 492–496 (2004).
9. True, J. R. & Carroll, S. B. Gene Co-Option in Physiological and Morphological Evolution. *Annu. Rev. Cell Dev. Biol.* **18**, 53–80 (2002).
10. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**, 25–36 (2008).
11. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nat Rev Genet* **9**, 868–882 (2008).
12. Arendt, D. *et al.* The origin and evolution of cell types. *Nat Rev Genet* **17**, 744–757 (2016).
13. Hain, D. *et al.* Molecular diversity and evolution of neuron types in the amniote brain. *Science* **377**, eabp8202 (2022).
14. Wang, J. *et al.* Tracing cell-type evolution by cross-species comparison of cell atlases. *Cell Reports* **34**, 108803 (2021).
15. Wu, B. *et al.* Single-cell analysis of the amphioxus hepatic caecum and vertebrate liver reveals genetic mechanisms of vertebrate liver evolution. *Nat Ecol Evol* (2024) doi:10.1038/s41559-024-02510-9.
16. Liang, C., the FANTOM Consortium, Forrest, A. R. R. & Wagner, G. P. The statistical geometry of transcriptome divergence in cell-type evolution and cancer. *Nat Commun* **6**, 6066 (2015).
17. Arendt, D., Bertucci, P. Y., Achim, K. & Musser, J. M. Evolution of neuronal types and families. *Current Opinion in Neurobiology* **56**, 144–152 (2019).
18. Tosches, M. A. *et al.* Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).
19. Zaremba, B. *et al.* Developmental origins and evolution of pallial cell types and structures in birds. *Science* **387**, eadp5182 (2025).
20. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).
21. Lien, S. *et al.* The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
22. Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* **48**, 427–437 (2016).
23. Sandve, S. R., Rohlfs, R. V. & Hvidsten, T. R. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat Genet* **50**, 908–909 (2018).
24. Shafer, M. E. R., Sawh, A. N. & Schier, A. F. Gene family evolution underlies cell-type diversification in the hypothalamus of teleosts. *Nat Ecol Evol* **6**, 63–76 (2021).
25. Guillotin, B. *et al.* A pan-grass transcriptome reveals patterns of cellular divergence in crops. *Nature* **617**, 785–791 (2023).
26. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).
27. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).
28. Lamanna, F. *et al.* A lamprey neural cell type atlas illuminates the origins of the vertebrate brain. *Nat Ecol Evol* (2023) doi:10.1038/s41559-023-02170-1.
29. Holland, L. Z. Tunicates. *Current Biology* **26**, R146–R152 (2016).
30. Berná, L. & Alvarez-Valin, F. Evolutionary genomics of fast evolving tunicates. *Genome Biol Evol* **6**, 1724–1738 (2014).
31. Ryan, K., Lu, Z. & Meinertzhagen, I. A. The CNS connectome of a tadpole larva of Ciona intestinalis (L.) highlights sidedness in the brain of a chordate sibling. *eLife* **5**, e16962 (2016).
32. Tarashansky, A. J., Xue, Y., Li, P., Quake, S. R. & Wang, B. Self-assembling manifolds in single-cell RNA sequencing data. *eLife* **8**, e48994 (2019).
33. Hobert, O. Terminal Selectors of Neuronal Identity. in *Current Topics in Developmental Biology* vol. 116 455–475 (Elsevier, 2016).
34. Hobert, O. & Kratsios, P. Neuronal identity control by terminal selectors in worms, flies, and chordates. *Current Opinion in Neurobiology* **56**, 97–105 (2019).

35. Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *eLife* **10**, e66747 (2021).
36. Nieuwenhuys, R. *Towards a New Neuromorphology.* (Springer International Publishing AG, Cham, 2016).
37. Wagner, G. P. Devo-Evo of Cell Types. in *Evolutionary Developmental Biology* (eds. Nuno De La Rosa, L. & Müller, G.) 1–18 (Springer International Publishing, Cham, 2019). doi:10.1007/978-3-319-33038-9_153-1.
38. Wagner, G. P. The developmental genetics of homology. *Nat Rev Genet* **8**, 473–479 (2007).
39. Serrano-Saiz, E. *et al.* Modular Control of Glutamatergic Neuronal Identity in C. elegans by Distinct Homeodomain Proteins. *Cell* **155**, 659–673 (2013).
40. Luo, W. *et al.* Single-cell spatial transcriptomic analysis reveals common and divergent features of developing postnatal granule cerebellar cells and medulloblastoma. *BMC Biol* **19**, 135 (2021).
41. Bullock, T. H., Moore, J. K. & Fields, R. D. Evolution of myelin sheaths: Both lamprey and hagfish lack myelin. *Neuroscience Letters* **48**, 145–148 (1984).
42. Lannoo, M. J. & Hawkes, R. A search for primitive Purkinje cells: zebrin II expression in sea lampreys (Petromyzon marinus). *Neuroscience Letters* **237**, 53–55 (1997).
43. Sugahara, F., Murakami, Y., Pascual-Anaya, J. & Kuratani, S. Reconstructing the ancestral vertebrate brain. *Dev Growth Differ* **59**, 163–174 (2017).
44. Gemmell, N. J. *et al.* The tuatara genome reveals ancient features of amniote evolution. *Nature* **584**, 403–409 (2020).
45. Dai, Y. *et al.* Evolutionary origin of the chordate nervous system revealed by amphioxus developmental trajectories. *Nat Ecol Evol* (2024) doi:10.1038/s41559-024-02469-7.
46. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083–1086 (2017).
47. Singh, P. P., Arora, J. & Isambert, H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput Biol* **11**, e1004394 (2015).
48. Vance, Z. & McLysaght, A. Ohnologs and SSD Paralogs Differ in Genomic and Expression Features Related to Dosage Constraints. *Genome Biology and Evolution* **15**, evad174 (2023).
49. Li, Y. *et al.* Origin and stepwise evolution of vertebrate lungs. *Nat Ecol Evol* (2025) doi:10.1038/s41559-025-02642-6.
50. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
51. Lin, S. *et al.* Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17224–17229 (2014).
52. Musser, J. M. & Wagner, G. P. Character trees from transcriptome data: Origin and individuation of morphological characters and the so-called "species signal". *J Exp Zool Pt B* **324**, 588–604 (2015).
53. Liang, C., Musser, J. M., Cloutier, A., Prum, R. O. & Wagner, G. P. Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes. *Genome Biology and Evolution* **10**, 538–552 (2018).
54. Yao, Z. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023).
55. Gillard, G. B. *et al.* Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol* **22**, 103 (2021).
56. Marlétaz, F. *et al.* Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
57. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14746–14753 (2012).
58. Song, M. J., Potter, B. I., Doyle, J. J. & Coate, J. E. Gene Balance Predicts Transcriptional Responses Immediately Following Ploidy Change in *Arabidopsis thaliana*. *The Plant Cell* **32**, 1434–1448 (2020).
59. Feiner, N., Meyer, A. & Kuraku, S. Evolution of the vertebrate Pax4/6 class of genes with focus on its novel member, the Pax10 gene. *Genome Biol Evol* **6**, 1635–1651 (2014).
60. Rastogi, S. & Liberles, D. A. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* **5**, 28 (2005).
61. Metzis, V. *et al.* Nervous System Regionalization Entails Axial Allocation before Neural Differentiation. *Cell* **175**, 1105-1118.e17 (2018).
62. Jessell, T. M. Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nat Rev Genet* **1**, 20–29 (2000).
63. Mayer, C. *et al.* Developmental diversification of cortical inhibitory interneurons. *Nature* **555**, 457–462 (2018).
64. Herrero-Navarro, Á. *et al.* Astrocytes and neurons share region-specific transcriptional signatures that confer regional identity to neuronal reprogramming. *Sci. Adv.* **7**, eabe8978 (2021).
65. Rowitch, D. H. & Kriegstein, A. R. Developmental genetics of vertebrate glial–cell specification. *Nature* **468**, 214–222 (2010).
66. Rodrigo Albors, A. *et al.* An ependymal cell census identifies heterogeneous and ongoing cell maturation in the adult mouse spinal cord that changes dynamically on injury. *Developmental Cell* **58**, 239-255.e10 (2023).
67. Martynoga, B., Morrison, H., Price, D. J. & Mason, J. O. Foxg1 is required for specification of ventral telencephalon and region-specific regulation of dorsal telencephalic precursor proliferation and apoptosis. *Developmental Biology* **283**, 113–127 (2005).
68. Shimizu, T. *et al.* Zinc finger genes *Fezf1* and *Fezf2* control neuronal differentiation by repressing *Hes5* expression in the forebrain. *Development* **137**, 1875–1885 (2010).
69. Yoshida, M. *et al. Emx1* and *Emx2* functions in development of dorsal telencephalon. *Development* **124**, 101–111 (1997).
70. Karalay, Ö. *et al.* Prospero-related homeobox 1 gene (Prox1) is regulated by canonical Wnt signaling and has a stage-specific role in adult hippocampal neurogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5807–5812 (2011).

71. Lipiec, M. A. *et al.* TCF7L2 regulates postmitotic differentiation programs and excitability patterns in the thalamus. *Development* dev.190181 (2020) doi:10.1242/dev.190181.
72. Lavado, A., Lagutin, O. V. & Oliver, G. *Six3* inactivation causes progressive caudalization and aberrant patterning of the mammalian diencephalon. *Development* **135**, 441–450 (2008).
73. Vernay, B. *et al.* Otx2 Regulates Subtype Specification and Neurogenesis in the Midbrain. *J. Neurosci.* **25**, 4856–4867 (2005).
74. Sgaier, S. K. *et al.* Genetic subdivision of the tectum and cerebellum into functionally related regions based on differential sensitivity to engrailed proteins. *Development* **134**, 2325–2335 (2007).
75. Thompson, J. A., Zembrzycki, A., Mansouri, A. & Ziman, M. Pax7 is requisite for maintenance of a subpopulation of superior collicular neurons and shows a diverging expression pattern to Pax3 during superior collicular development. *BMC Dev Biol* **8**, 62 (2008).
76. Boudjadi, S., Chatterjee, B., Sun, W., Vemu, P. & Barr, F. G. The expression and function of PAX3 in development and disease. *Gene* **666**, 145–157 (2018).
77. Aruga, J. & Millen, K. J. ZIC1 Function in Normal Cerebellar Development and Human Developmental Pathology. in *Zic family* (ed. Aruga, J.) vol. 1046 249–268 (Springer Singapore, Singapore, 2018).
78. Kebschull, J. M. *et al.* Cerebellar nuclei evolved by repeatedly duplicating a conserved cell-type set. *Science* **370**, eabd5059 (2020).
79. Yopak, K. E., Pakan, J. M. P. & Wylie, D. The Cerebellum of Nonmammalian Vertebrates. in *Evolution of Nervous Systems* 373–385 (Elsevier, 2017). doi:10.1016/B978-0-12-804042-3.00015-4.
80. Green, M. J. & Wingate, R. J. Developmental origins of diversity in cerebellar output nuclei. *Neural Dev* **9**, 1 (2014).
81. Paxinos, G. & Franklin, K. B. J. *Paxinos and Franklin's The Mouse Brain in Stereotaxic Coordinates*. (Academic Press, an imprint of Elsevier, London, 2019).
82. Development of Cerebellar Nuclei. in *Handbook of the Cerebellum and Cerebellar Disorders* 179–205 (Springer Netherlands, Dordrecht, 2013). doi:10.1007/978-94-007-1333-8_10.
83. Fink, A. J. *et al.* Development of the Deep Cerebellar Nuclei: Transcription Factors and Cell Migration from the Rhombic Lip. *J. Neurosci.* **26**, 3066–3076 (2006).
84. Tamada, A. *et al.* Crucial roles of Robo proteins in midline crossing of cerebellofugal axons and lack of their up-regulation after midline crossing. *Neural Dev* **3**, 29 (2008).
85. Kim, D. & Ackerman, S. L. The UNC5C Netrin Receptor Regulates Dorsal Guidance of Mouse Hindbrain Axons. *J. Neurosci.* **31**, 2167–2179 (2011).
86. Teune, T. M., Van Der Burg, J., Van Der Moer, J., Voogd, J. & Ruigrok, T. J. H. Topography of cerebellar nuclear projections to the brain stem in the rat. in *Progress in Brain Research* vol. 124 141–172 (Elsevier, 2000).
87. Minguillon, C., Gibson-Brown, J. J. & Logan, M. P. *Tbx4/5* gene duplication and the origin of vertebrate paired appendages. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21726–21730 (2009).
88. Lamb, T. D. Evolution of phototransduction, vertebrate photoreceptors and retina. *Progress in Retinal and Eye Research* **36**, 52–119 (2013).
89. Tommasini, D., Yoshimatsu, T., Puthussery, T., Baden, T. & Shekhar, K. Comparative transcriptomic insights into the evolution of vertebrate photoreceptor types. *Current Biology* **35**, 2228-2239.e4 (2025).
90. Arendt, D. Evolution of eyes and photoreceptor cell types.
91. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**, 97–108 (2010).
92. Wagner, G. P. *Homology, Genes, and Evolutionary Innovation*. (Princeton University Press, 2014). doi:10.23943/princeton/9780691156460.001.0001.
93. Ouimette, J.-F., Jolin, M. L., L'honoré, A., Gifuni, A. & Drouin, J. Divergent transcriptional activities determine limb identity. *Nat Commun* **1**, 35 (2010).
94. Yao, Z. *et al.* AllenInstitute/scrattch.hicat: doi_release. Zenodo https://doi.org/10.5281/ZENODO.11405898 (2024).
95. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* **19**, 335–346 (2016).
96. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
97. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157 (2015).
98. Singh, P. P. & Isambert, H. OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. *Nucleic Acids Research* gkz909 (2019) doi:10.1093/nar/gkz909.
99. Simakov, O. *et al.* Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol* **4**, 820–830 (2020).
100. Kim, G. B., Gao, Y., Palsson, B. O. & Lee, S. Y. DeepTFactor: A deep learning-based tool for the prediction of transcription factors. *Proc Natl Acad Sci U S A* **118**, e2021171118 (2021).
101. Liu, A. *et al.* Discovery of optimal cell type classification marker genes from single cell RNA sequencing data. *BMC Methods* **1**, 15 (2024).
102. Shen, W.-K. *et al.* AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res* **51**, D39–D45 (2023).
103. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).
104. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692 (2021).
105. Junttila, S., Smolander, J. & Elo, L. L. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. *Briefings in Bioinformatics* **23**, bbac286 (2022).
106. Van de Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat Protoc* **15**, 2247–2276 (2020).

107. The UniProt Consortium *et al.* UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **53**, D609–D617 (2025).
108. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
109. Marc Carlson, H. P. AnnotationForge. Bioconductor https://doi.org/10.18129/B9.BIOC.ANNOTATIONFORGE (2017).
110. Carlson, M. org.Hs.eg.db. Bioconductor https://doi.org/10.18129/B9.BIOC.ORG.HS.EG.DB (2017).
111. Carlson, M. org.Mm.eg.db. Bioconductor https://doi.org/10.18129/B9.BIOC.ORG.MM.EG.DB (2017).
112. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).
113. Thomas, P. D. *et al.* PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res* **31**, 334–341 (2003).
114. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
115. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Soft.* **67**, (2015).
116. Farre, D. & Alba, M. M. Heterogeneous Patterns of Gene-Expression Diversification in Mammalian Gene Duplicates. *Molecular Biology and Evolution* **27**, 325–335 (2010).
117. Csardi, G. & Nepusz, T. The igraph software package for complex network research.
118. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. 0.7.2 https://doi.org/10.32614/CRAN.package.rstatix (2019).
119. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics***27**, 592–593 (2011).

Figures

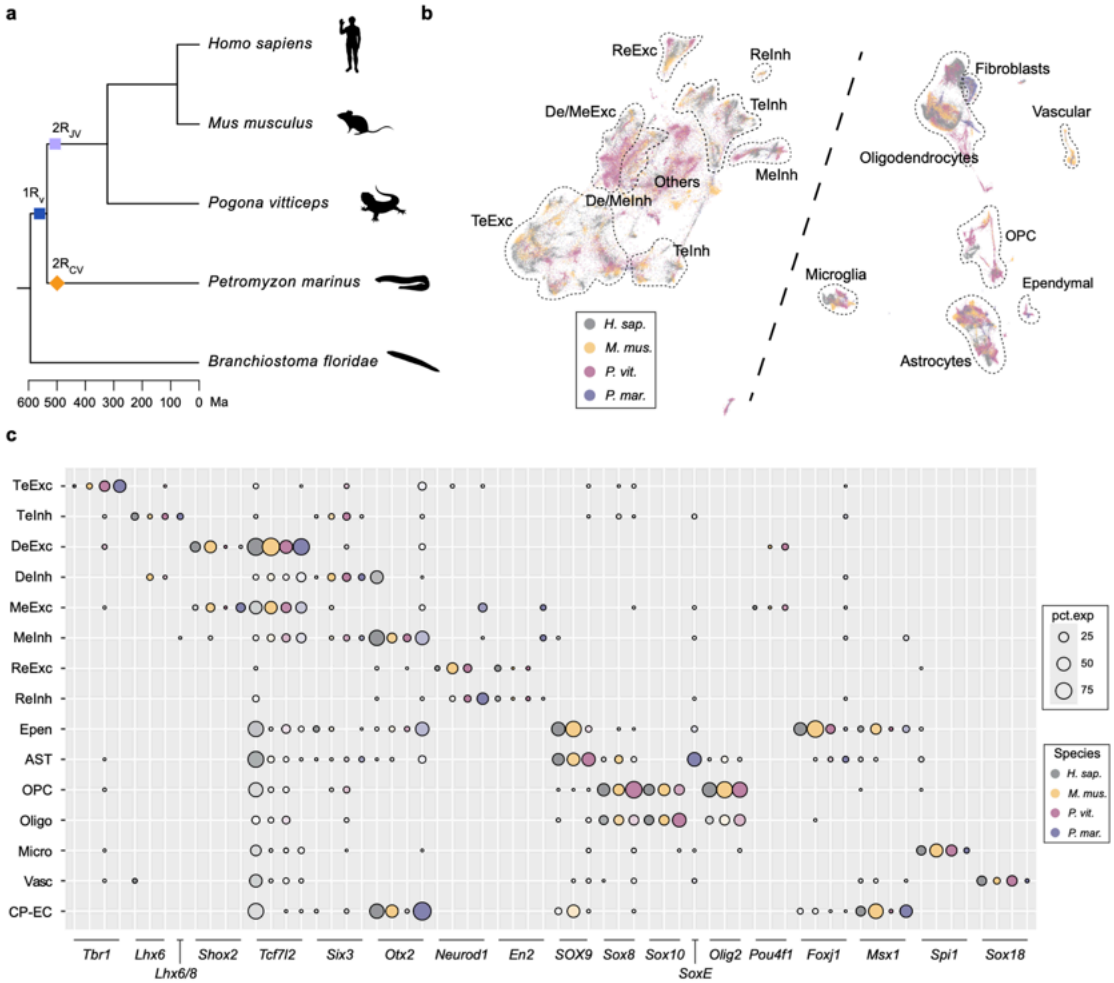

Figure 1

Vertebrate brain atlases and core TF programmes defining major cell type families. a, Phylogenetic tree showing the approximate timing (million years ago, Ma) for vertebrate shared auto-tetraploidization ($1R_V$), a jawed-vertebrate-specific auto-tetraploidization ($2R_{JV}$), and a cyclostome-specific hexaploidization ($2R_{CV}$) based on recent studies[1,5,44]. **b**, UMAP visualization of the integrated four species neuronal (left) and non-neuronal (right) atlases. Each dot represents a single nucleus or cell. To ensure balanced representation across datasets, only 20,000 randomly sampled cells or nuclei are shown per species for both neuronal and non-neuronal integrated atlases. **c**, Dot plot showing conserved TFs which define major cell type families in vertebrates. Dot size represents the percentages of cells within each cell type family expressing that gene. The colour gradient for each dot is scaled for each gene within the species. For TF

families with multiple copies in lamprey, only the copy with the highest expression is displayed. Cell type families are: TeExc, Telencephalon glutamatergic neurons; TeInh, Telencephalon GABAergic neurons; DeExc, Diencephalon glutamatergic neurons; DeInh, Diencephalon GABAergic neurons; MeExc, Mesencephalon glutamatergic neurons; MeInh, Mesencephalon GABAergic neurons; ReExc, Rhombencephalon glutamatergic neurons; ReInh, Rhombencephalon GABAergic neurons; Epen, Ependymal cells; AST, Astrocytes; OPC, Oligodendrocyte precursor cells; Oligo, Oligodendrocytes; Micro, Microglia; Vasc, Vascular cells; CP-EC, Choroid plexus epithelial cells.

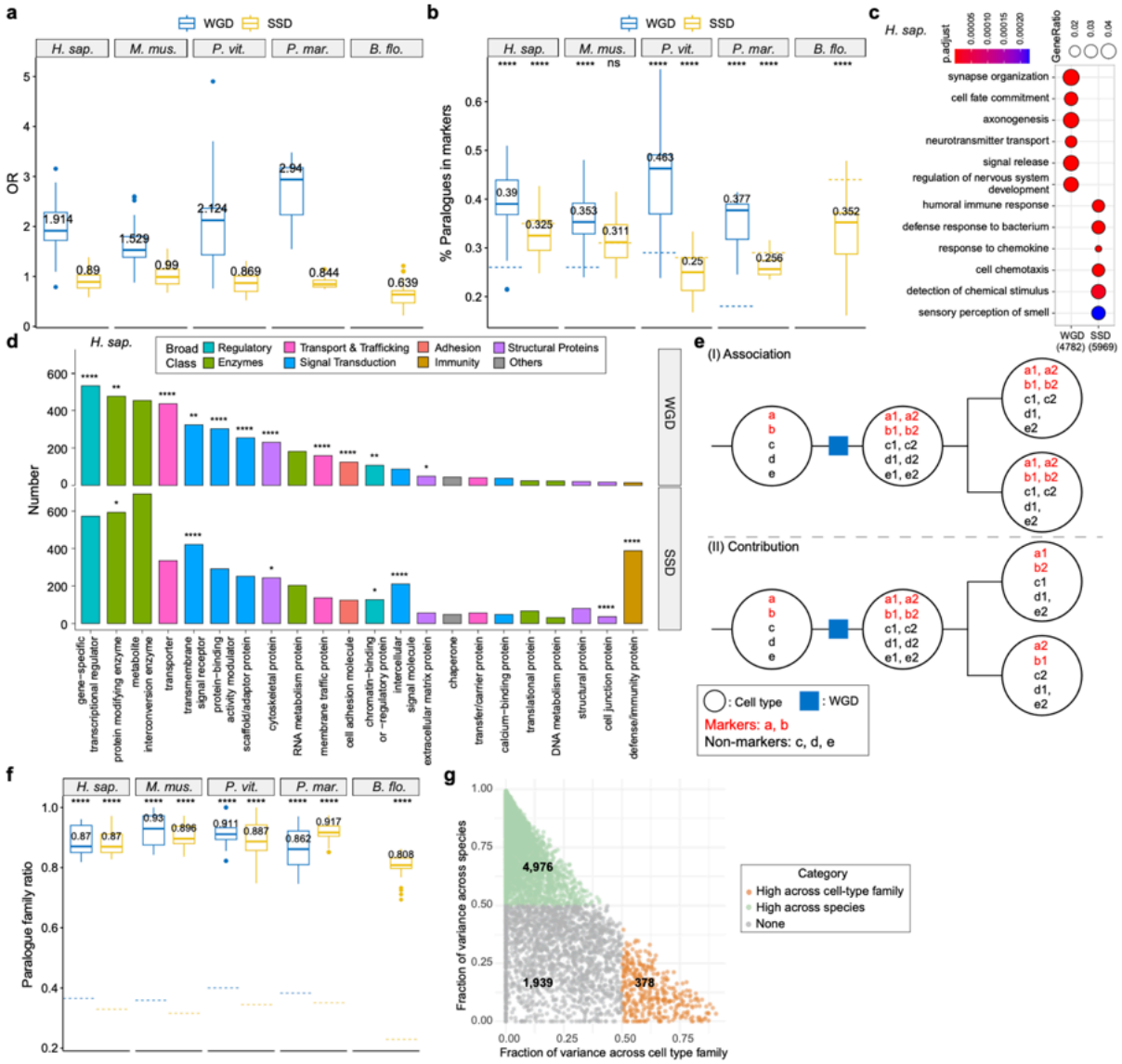

Figure 2

Ohnologues contributed more to cell type evolution than SSD paralogues. a, ORs calculated from Fisher's exact test on WGD (ohnologues) and SSD paralogues with cell type DEGs. Fisher's exact test was used to measure association between WGD/SSD paralogue and DEGs. **b,** The ratio of WGD (blue) and SSD (yellow) paralogues in cell type DEGs. The ratio of WGD and SSD paralogues in expressed protein-coding genes (background) is represented by blue and yellow dotted lines for each species, respectively. The significance of differences between the ratio of paralogues as markers and their total percentage in all protein-coding genes was assessed using a one-sample t-test. ns (not significant); * ($P < 0.05$); ** ($P < 0.01$); *** ($P < 0.001$); **** ($P < 0.0001$). The number of markers in different cell types ranges from 189 to 1447 (lamprey), 60 to 882 (lizard), 41 to 1375 (mouse), and 111 to 2418 (human). **c,** Selected top enriched GO terms of WGD and SSD paralogues in human. The colour denotes FDR range and size of circle represents gene ratio. The numberin bracket are the number of retained WGD/SSD paralogue genes. **d,**Bar plot showing the number of different protein classes in WGD and SSD paralogues of human. Colour represents broad classifications. FDR values of overrepresented classes are shown as same as (**b**) whereas FDR in under-represented classes and non-significant classes are not shown. **e,** Illustration showing the differences between ohnologue associated with markers (I) and ohnologue contribute to cell type evolution (II). **f,** The ratio of the number of WGD families (blue) and SSD families (yellow) that include markers to the number of paralogues that are markers for each cell type family. If only one copy in each gene family were used as a marker, the ratio would be 1. The number shown for each box represents the median value. The background ratio denotes the number of paralogue families to number of paralogues and were represented by a blue (WGD) and a yellow (SSD) dotted line to be used as background for each species. The *P*-value is displayed as in (**b**). **g,** Variance decomposition of transcriptome estimates the relative contribution of cell-type family and species/batch to the observed variance in gene expression for metagenes. Green dots indicate genes with higher contribution (> 50%) to species or batch effects, and genes with orange colour have higher (>50%) cell type family

contributions. Grey dots represent genes that do not highly contribute to either. The bold numbers indicate the number of metagenes matching the three categories.

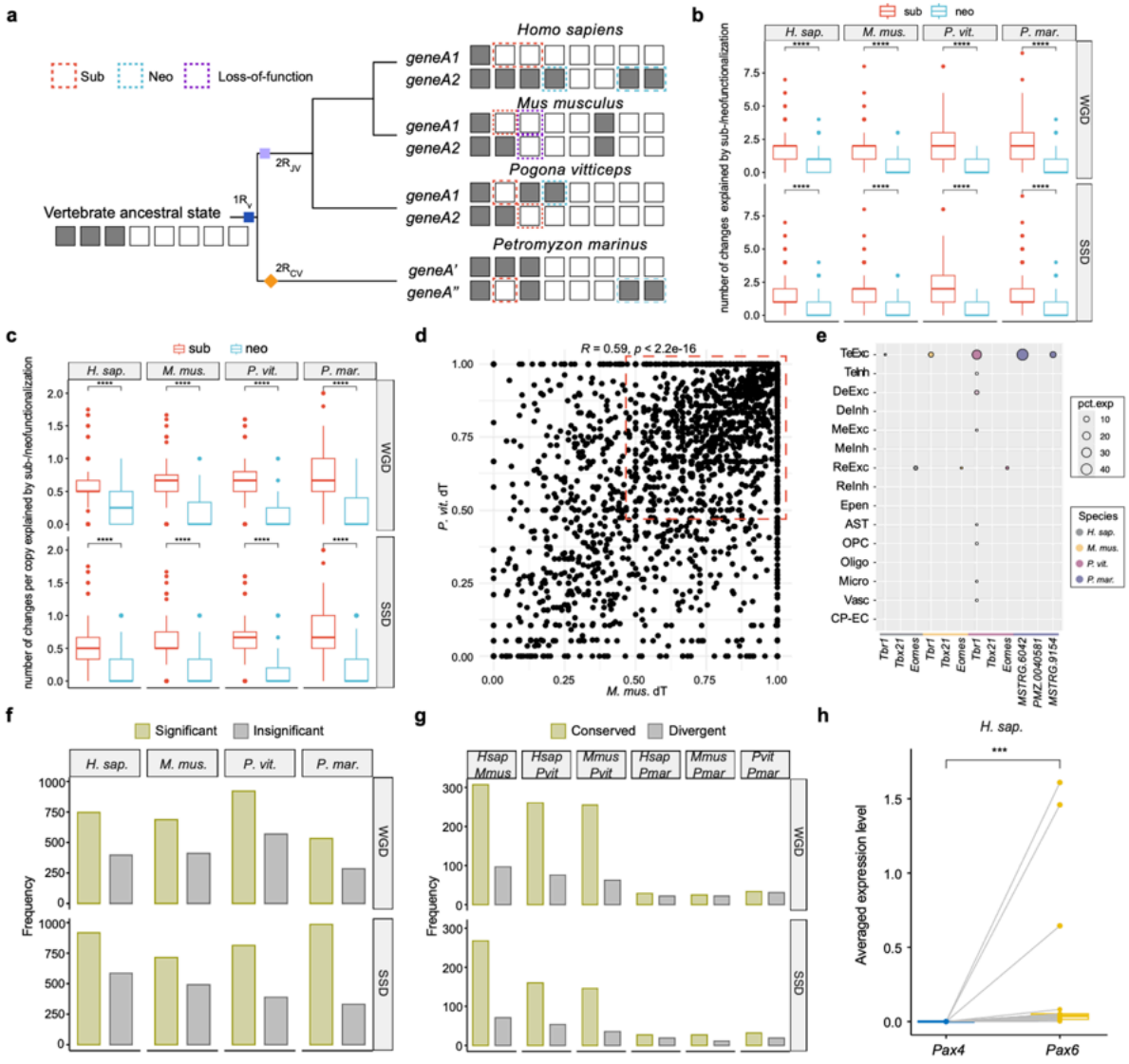

Figure 3

Divergence of paralogues drive cell type evolution. a,Illustration of ancestral state inferring and the cases of subfunctionalization, neofunctionalization, and loss-of-function. **b,** The number of changes explained by either subfunctionalization or neofunctionalization for each WGD/SSD paralogue family. Statistics comparison by paired Wilcoxon signed-rank test. ns (not significant); * ($P < 0.05$); ** ($P < 0.01$); *** ($P < 0.001$); **** ($P < 0.0001$). **c,** The number of changes per copy explained by either subfunctionalization or neofunctionalization for each WGD/SSD paralogue family. The p-value was calculated and displayed as in (**b**). **d,** Expression divergence of orthogroups for mice and lizards. dT = 1 represents copies were not expressed in the same cell type at all. The red dotted box highlights paralogue families with high expression divergence at both species. **e,** Dot plot showing expression pattern of *Tbr1* subfamily. Legend details as in Fig. 1c. **f,**The number of WGD and SSD paralogue families (with only two copies) that have and do not have a significantly dominant copy. **g,** The number of WGD and SSD paralogue family using the same genes or not as the dominant copy in species pairwise comparisons. Since the second round of WGD happened independently in cyclostome and gnathostome lineage, in many cases, we cannot find 1-to-1 orthologues between lamprey and jawed vertebrates. **h,** Pseudobulk expression of *Pax6-Pax4* family for human. Pseudobulk expression was calculated by the average of SAMalg normalised expression for each cell type family. Each dot represents a cell type family and the grey line between genes denotes the comparison of paralogues at the same cell type family. Both Friedman test and paired Wilcoxon signed-rank test show significance for all species ($P < 0.001$ for all three amniotes; $P = 0.00166$ for lamprey).

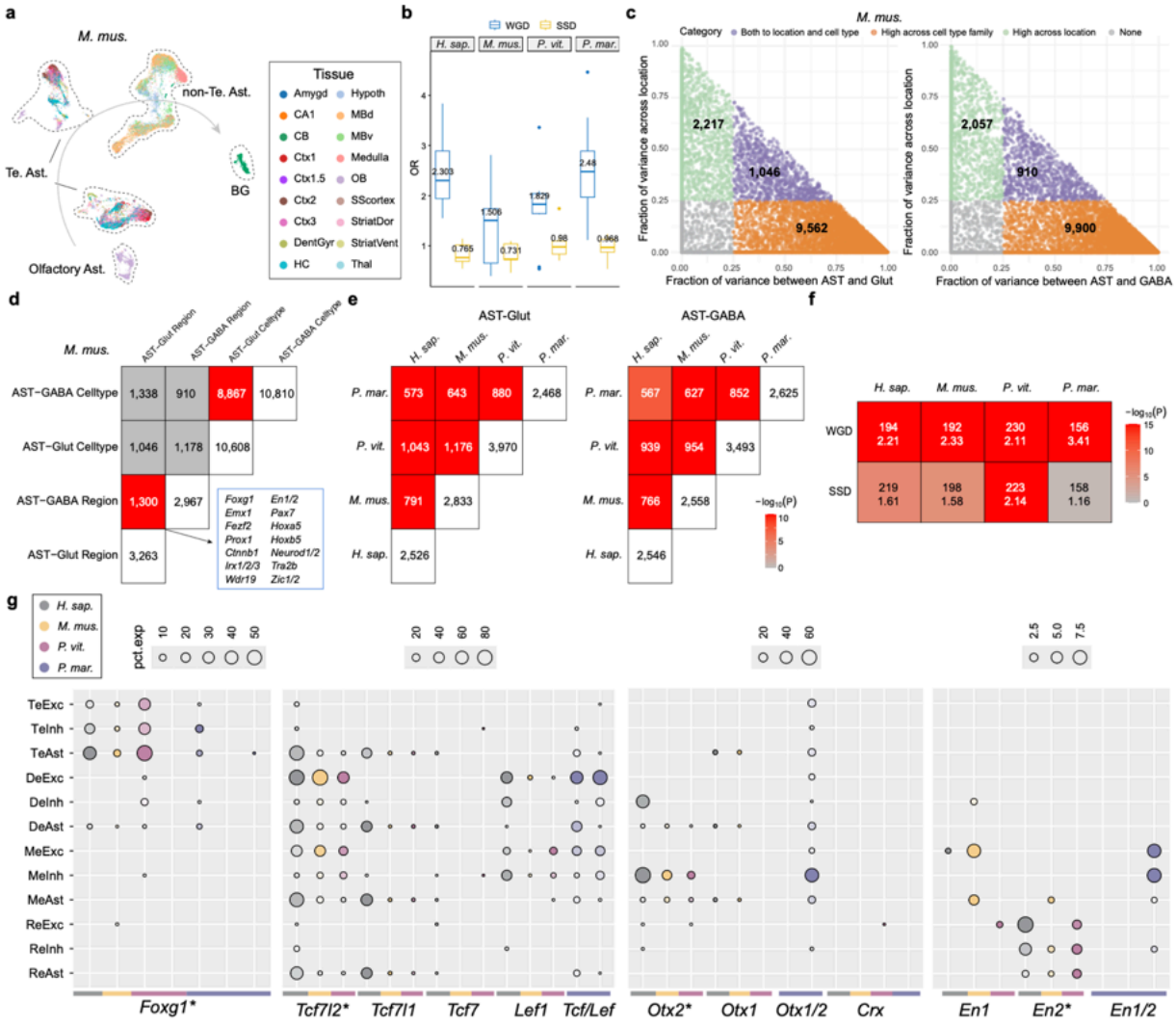

Figure 4

Evolution of regional programmes contributing to cell diversity. a, UMAP showing projection of mouse astrocytes with the anterior to posterior location indicated. BG: Bergmann glia; Te: Telencephalon. **b,** Odds ratio (OR) calculated from Fisher's exact test between WGD/SSD paralogues and DEGs of astrocyte subtypes. Each dot represents a subtype of astrocytes. **c,** Variance decomposition of transcriptome estimates the relative contribution of cell type family and brain divisions (Te/De/Me/Re) to the observed variance in gene expression of mouse. Green dots indicate genes with relatively stronger contribution (> 25%) to location, and genes with red dots have strong (>25%) cell-type variance contributions. Purple dots represent genes that contribute to both variables. Grey dots represent genes that contribute weakly to both variables. The bold numbers on dots indicate the number of metagenes matching the different categories. **d,** The comparison matrix of gene sets identified in (**c**). The colour is red if Fishers' exact test shows significance with OR > 1. The gradient red colour represents -$\log_{10}(P)$. The listed genes are some overlapped genes between two sets of regional identity genes. **e,** The comparison matrix of AST-GABAergic neurons and AST-Glutamatergic neurons regional orthogroups across species. For each species, a regional orthogroup was defined as such if any regional identity genes of that species were in that orthogroup. Colour intensity is as in (**d**). **f,** The comparison matrix showing relationships between paralogues (WGD and SSD paralogues) and genes in conserved regional orthogroups for each species. Genes in conserved regional orthogroups were identified if regional genes in (**c**) in the vertebrate conserved regional orthogroup. The upper number for each comparison denotes the number of overlaps while the lower number represents the odds ratio. **g,** Dot plots showing the expression of selected important genes (labelled with *) involved in regional identity and their ohnologues in neural families from different brain divisions. Dot size represents the percentages of cells within each cluster expressing that gene. The gradient colours from white to species colour were scaled for each gene on individual species. Species identity is also shown at the bottom of the plots as coloured line.

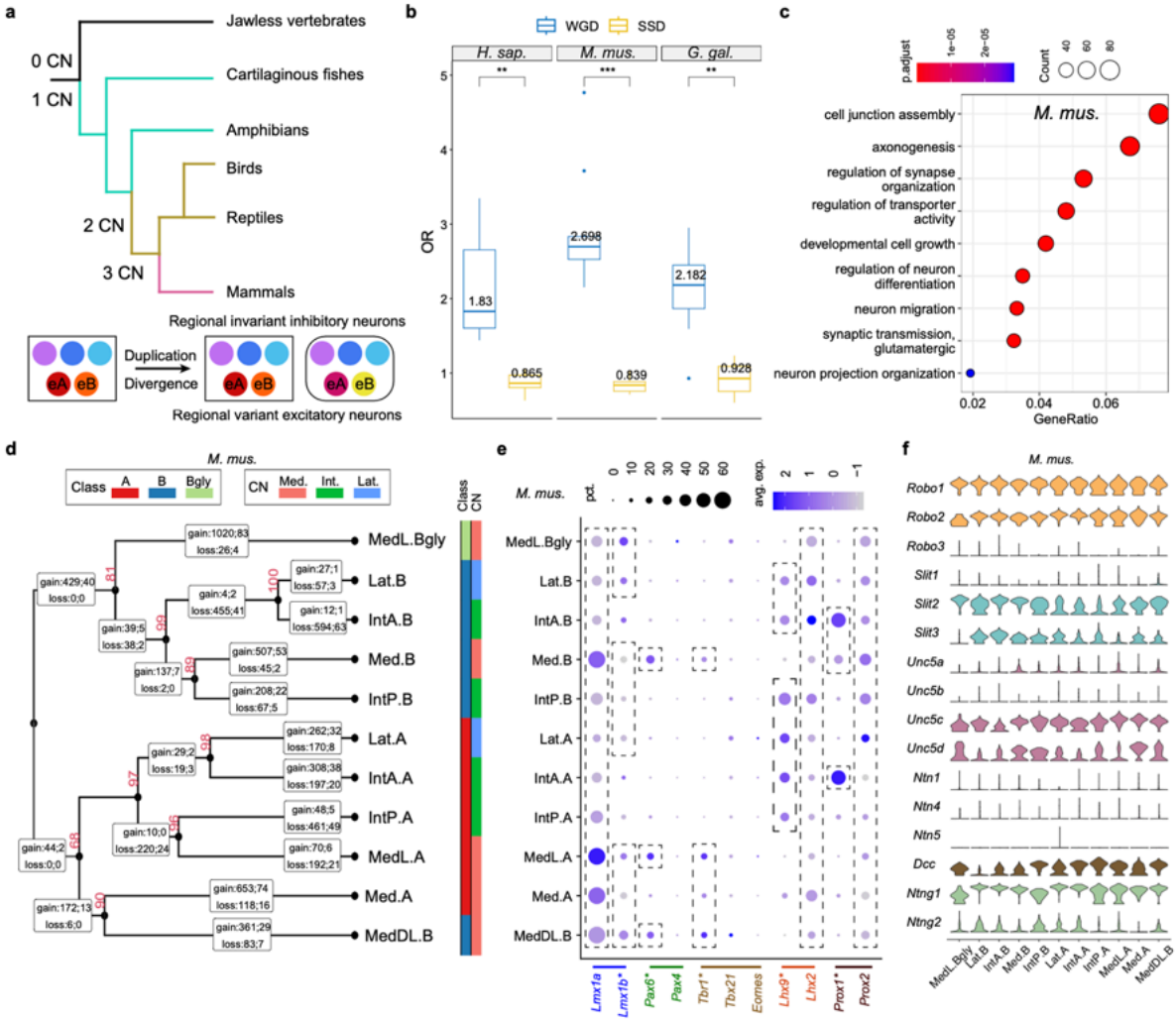

Figure 5

Ohnologues in Cerebellar Nucleus evolution. a, Illustration (modified from[78]) showing the number of CN pairs in vertebrates and the duplication and divergence of excitatory neurons within them. **b,** Odds ratio (OR) calculated from Fisher's exact test on WGD and SSD paralogues with cell type DEGs, respectively. **c,** GO enrichment of DEGs of excitatory neurons in CN of mice. Colour represents adjusted p-value and dot size denotes gene ratio. **d,** Dendrogram of excitatory neurons in CN and the number of predicted genes with gain- and loss-of-function (regarding expression domain). The left colour bar represents cell type classes and right bar represents location of CN. The red number near the node is Approximately Unbiased (AU) p-values, computed by pvclust with 1,000 times multiscale bootstrap resampling. Rectangles in the branch contain the number of genes gained and lost in expression domain. The number following the semicolon represents the number of TF genes involved. **e,** Dot plots showing expression pattern of important TF genes (labelled with *) and their ohnologues. *Lmx1a* expressed in all subtypes but differentially expressed across subtypes. Colour represents scaled average expression for each gene. Dot size represents the percentage of cells expressing that gene in a certain cell type, percentage above than 60% were converted to 60%. **f,** Violin plot of genes involved in axon guidance system. Genes from the same orthogroup are the same colour.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1marker.xlsx
- SupplementaryTable2Refinedannotation.xlsx
- SupplementaryTable3conservedTFs.xlsx
- SupplementaryTable4macrogliamarker.xlsx
- SupplementaryTable5LMM.xlsx
- SupplementaryTable6CNmarkersandohnologues.xlsx
- SupplementaryTable7CNgainlossgenes.xlsx
- SupplementaryTable8datasetsusedforgenetree.xlsx

- SupplementaryTable9paralogextractedfromEnsembl.xlsx
- ExtendedDataFigures21June25.docx