Title: Systematic detection of divergent brain protein-coding genes in human evolution
and their roles in cognition

Short title: Divergent brain protein-coding genes in human evolution

Guillaume Dumas\textsuperscript{a}, Simon Malesys\textsuperscript{a} and Thomas Bourgeron\textsuperscript{a}

\textsuperscript{a} Human Genetics and Cognitive Functions, Institut Pasteur, UMR3571 CNRS, Université de Paris, Paris,
(75015) France

\textit{Corresponding author:}
Guillaume Dumas
Human Genetics and Cognitive Functions
Institut Pasteur
75015 Paris, France
Phone: +33 6 28 25 56 65
\texttt{guillaume.dumas@pasteur.fr}
Abstract

The human brain differs from that of other primates, but the genetic basis of these differences remains unclear. We investigated the evolutionary pressures acting on almost all human protein-coding genes (N=11,667; 1:1 orthologs in primates) on the basis of their divergence from those of early hominins, such as Neanderthals, and non-human primates. We confirm that genes encoding brain-related proteins are among the most strongly conserved protein-coding genes in the human genome. Combining our evolutionary pressure metrics for the protein-coding genome with recent datasets, we found that this conservation applied to genes functionally associated with the synapse and expressed in brain structures such as the prefrontal cortex and the cerebellum. Conversely, several of the protein-coding genes that diverge most in hominins relative to other primates are associated with brain-associated diseases, such as micro/macrocephaly, dyslexia, and autism. We also showed that cerebellum granule neurons express a set of divergent protein-coding genes that may have contributed to the emergence of fine motor skills and social cognition in humans. This resource is available from http://neanderthal.pasteur.fr and can be used to estimate evolutionary constraints acting on a set of genes and to explore their relative contributions to human traits.
Introduction

Modern humans (*Homo sapiens*) can perform complex cognitive tasks well and communicate with their peers [1]. Anatomic differences between the brains of humans and other primates are well documented (e.g. cortex size, prefrontal white matter thickness, lateralization), but the way in which the human brain evolved remains a matter of debate [2]. A recent study of endocranial casts of *Homo sapiens* fossils indicates that, brain size in early *Homo sapiens*, 300,000 years ago, was already within the range of that in present-day humans [3]. However, brain shape, evolved more gradually within the *Homo sapiens* lineage, reaching its current form between about 100,000 and 35,000 years ago. It has also been suggested that the enlargement of the prefrontal cortex relative to the motor cortex in humans is mirrored in the cerebellum by an enlargement of the regions of the cerebellum connected to the prefrontal cortex [4]. These anatomic processes of tandem evolution in the brain paralleled the emergence of motor and cognitive abilities, such as bipedalism, planning, language, and social awareness, which are particularly well developed in humans.

Genetic differences in primates undoubtedly contributed to these brain and cognitive differences, but the genes or variants involved remain largely unknown. Indeed, demonstrating that a genetic variant is adaptive requires strong evidence at both the genetic and functional levels. Only few genes have been shown to be human-specific. They include *SRGAP2C* [5], *ARHGAP11B* [6] and *NOTCH2NL* [7], which emerged through recent gene duplication in the *Homo* lineage [8]. Remarkably, the expression of these human specific genes in the mouse brain expand cortical neurogenesis [6,7,9,10]. Several genes involved in brain function have been shown to display accelerated coding region evolution in humans. For example, *FOXP2* has been associated with verbal apraxia and *ASPM* with microcephaly [11,12]. Functional studies have also shown that mice carrying a “humanized” version of *FOXP2* display qualitative changes in ultrasonic vocalization [13]. However, these reports
targeting only specific genes sometimes provide contradictory results [14]. Other studies have reported sequence conservation to be stronger in the protein-coding genes of the brain than in those of other tissues [15–17], suggesting that the main substrate of evolution in the brain is regulatory changes in gene expression [18–20] and splicing [21]. In addition, several recent studies have recently explored the genes subjected to the highest degrees of constraint during primate evolution or in human populations, to improve estimations of the pathogenicity of variants identified in patients with genetic disorders [22,23]. By contrast, few studies have systematically detected genes that have diverged during primate evolution [24,25].

We describe here an exhaustive screening of all protein-coding genes for conservation and divergence from the common primate ancestor, making use of rich datasets of brain single-cell transcriptomics, proteomics and imaging to investigate the relationships between these genes and brain structure, function, and diseases.
Figure 1 Evolution of protein-coding genes across tissues and biological functions. (a) Analysis pipeline for the extraction of $\omega_{\text{GC12}}$, a corrected and normalized measurement of evolution of protein-coding genes that behaves like a Z-score and takes into account the GC content of codons. (b) Hierarchical clustering, on the basis of $\omega_{\text{GC12}}$, across all protein-coding genes (1:1 orthologs in hominins with medium coverage; See Supplementary Table 1). (c) Gene ontology (GO) enrichments for the red and blue clusters in panel b (See Supplementary Table 2 for all GO terms). Horizontal lines indicate the 95% confidence intervals. (d) Funnel plot summarizing the evolution of protein-coding genes specifically expressed in different tissues of the human body (Supplementary Table 3). The dashed horizontal line indicates the threshold for significance after Bonferroni correction. Stars indicate the set of genes for which statistical significance was achieved in multiple comparisons after correction, with a bootstrap taking GC12 content and coding sequence length into account. HS: *Homo sapiens*; 6-EPO ancestor: the reconstructed ancestral genome of primates based on alignments of human, chimpanzee, gorilla, orangutan, rhesus macaque, and marmoset genomes.
**Results**

**Strong conservation of brain protein-coding genes**

We first compared the sequences of modern humans, archaic humans, and other primates to those of their common primate ancestor (inferred from the Compara 6-way primate Enredo, Pecan, Ortheus multiple alignments [26]), to extract a measurement of evolution for 11,667 of the 1:1 orthologs across primates, selected from the 17,808 protein-coding genes in the modern human genome (Fig. 1a, see also Supplementary Fig. 1 and 2; 27). This resource is available online from [http://neanderthal.pasteur.fr](http://neanderthal.pasteur.fr). Our measurement is derived from one of the most widely used and reliable measurements of evolutionary pressure on protein-coding regions, the dN/dS ratio [28], also called $\omega$. This measurement compares the rates of non-synonymous and synonymous mutations of coding sequences. If there are more non-synonymous mutations than expected, there is divergence, if fewer, there is conservation. We first estimated dN and dS for all 1:1 orthologous genes, because the evolutionary constraints on duplicated genes are relaxed [29] (note: only the Y chromosome was excluded from these analyses). We then adjusted the dN/dS ratio for biases induced by variations of mutations rate with the GC content of codons. Finally, we renormalized the values obtained for each taxon across the whole genome. The final $\omega_{GC12}$ obtained took the form of Z-score corrected for GC content that quantified the unbiased divergence of genes relative to the ancestral primate genome [27].

Using the $\omega_{GC12}$ for all protein-coding genes in *Homo sapiens*, Denisovans, Neanderthals, and *Pan troglodytes*, we identified two distinct clusters in hominins (Fig. 1b and Supplementary Table 1): one containing divergent protein-coding genes, enriched in olfactory genes (OR=1.48, $p=8.4e-9$), and one with conserved protein-coding genes, enriched in brain-related biological functions (Fig. 1c and Supplementary Table 2). This second cluster
revealed a particularly strong conservation of genes encoding proteins involved in nervous system development (OR=1.2, \( p=2.4e-9 \)) and synaptic transmission (OR=1.35, \( p=1.7e-8 \)).

We investigated the possible enrichment of specific tissues in conserved and divergent proteins by analyzing RNAseq (Illumina Bodymap2 and GTEx), microarray and proteomics datasets (Methods). For expression data, we evaluated the specificity of genes by normalizing their profile across tissues (Supplementary Fig. 3). The results confirmed a higher degree of conservation for protein-coding genes expressed in the brain (Wilcoxon rank correlation (rc)=-0.1, \( p=4.1e-12 \), bootstrap corrected for gene length and GC content) than for those expressed elsewhere in the body, with the greatest divergence observed for genes expressed in the testis (Wilcoxon rc=0.3, \( p=7.8e-11 \), bootstrap corrected for gene length and GC content; Fig. 1d, see also Supplementary Fig. 4 and 5). This conservation of brain protein-coding genes was replicated with two other datasets (MicroArray: Wilcoxon OR=-0.18, \( p=1.8e-12 \); mass spectrometry: Wilcoxon rc=-0.21, \( p=1.55e-9 \); bootstrap corrected for gene length and GC content).

**Conservation of protein-coding genes relating to nervous system substructure and neuronal functions**

We then used microarray [30] and RNAseq [31] data to investigate the evolutionary pressures acting on different regions of the central nervous system. Three central nervous system substructures appeared to have evolved under the highest level of purifying selection at the protein sequence level (\( \omega_{GC12}<2 \), i.e. highly conserved): (i) the cerebellum (Wilcoxon rc=-0.29, \( p=5.5e-6 \), Bonferroni corrected) and the cerebellar peduncle (Wilcoxon rc=-0.11, \( p=3.2e-4 \), bootstrap corrected for gene length and GC content), (ii) the amygdala (Wilcoxon rc=-0.11, \( p=4.1e-6 \), bootstrap corrected for gene length and GC content), and, more
surprisingly, (iii) the prefrontal cortex (Wilcoxon $rc=-0.1$, $p=5.7e^{-10}$, bootstrap corrected for gene length and GC content; Fig. 2a, see also Supplementary Table 3). Indeed, it has been suggested that the prefrontal cortex is one of the most divergent brain structure in human evolution [32], this diversity being associated with high-level cognitive function [33]. Only one brain structure was more divergent than expected: the superior cervical ganglion (Wilcoxon $rc=0.22$, $p=1e^{-6}$, bootstrap corrected for gene length and GC content). This structure provides sympathetic innervation to many organs and is associated with the archaic functions of fight-or-flight response. The divergent genes expressed in the superior cervical ganglion include \textit{CARF}, which was found to be specifically divergent in the genus \textit{Homo}. This gene encodes a calcium-responsive transcription factor that regulates the neuronal activity-dependent expression of \textit{BDNF} [34] and a set of singing-induced genes in the song nuclei of the zebra finch, a songbird capable of vocal learning [35]. This gene had a raw $dN/dS$ of 2.44 (7 non-synonymous vs 1 synonymous mutations in \textit{Homo sapiens} compared to the common primate ancestor) and was found to be one of the most divergent protein-coding genes expressed in the human brain.

We then investigated the possible enrichment of conserved and divergent genes in brain-specific gene ontology terms. All pathways displayed high overall levels of conservation, but genes encoding proteins involved in glutamatergic and GABAergic neurotransmission were generally more conserved (Wilcoxon $rc=-0.25$; $p=9.8e^{-6}$, Bonferroni corrected) than those encoding proteins involved in dopamine and peptide neurotransmission and intracellular trafficking (Fig. 2b, see also Supplementary Fig. 6 and Supplementary Table 3). The recently released ontology of the synapse provided by the SynGO consortium (http://syngoportal.org) was incorporated into this analysis, not only confirming the globally strong conservation of the synapse, but also revealing its close relationship to trans-synaptic signaling processes (Wilcoxon $rc=-0.21$, $p=4.5e^{-5}$, Bonferroni corrected) and to postsynaptic
Figure 2 Evolution of brain-related protein-coding genes. (a) Funnel plot summarizing the evolution of protein-coding genes specifically expressed in brain substructures; the dashed horizontal line indicates the threshold for significance after Bonferroni correction. Stars indicate sets of genes for which statistical significance was achieved for multiple comparisons with bootstrap correction; (b) Matrix summarizing the effect size of the difference in protein-coding gene divergence between synaptic functions; colored cells indicate Mann-Whitney comparisons with a nominal $p$-value <0.05. Black dots indicate comparisons satisfying the Bonferroni threshold for statistical significance. (c,
d) SynGO sunburst plots showing nested statistically conserved (blue) biological
processes and cellular components of the synapse.

Divergent protein-coding genes and their correlation with brain expression and
function

We focused on the genes situated at the extremes of the $\omega_{GC12}$ distribution (>2SD;
Fig. 3a; Supplementary Table 4) and those fixed in the modern Homo sapiens population
(neutrality index<1), to ensure that we analyzed the most-divergent protein-coding genes.
Only 126 of these 352 highly divergent protein-coding genes were brain-related
(impoverishment for brain genes, Fisher’s exact test OR=0.66, $p=1e-4$), listed as synaptic
genes [36,37], specifically expressed in the brain (+2SD for specific expression) or related to
a brain disease (extracted systematically from Online Mendelian Inheritance in Man -
OMIM: https://www.omim.org and Human Phenotype Ontology - HPO:
https://hpo.jax.org/app/). For comparison, we also extracted the 427 most strongly conserved
protein-coding genes, 290 of which were related to the brain categories listed above
(enrichment for brain genes, Fisher’s exact test OR=1.26, $p=0.0032$).
**Figure 3 Brain protein-coding genes and human diseases.** (a) Distribution of $\omega_{GC12}$ and Venn diagrams describing the most conserved and divergent protein-coding genes specifically expressed in the brain, related to the synapse, or brain diseases (Supplementary Table 4). (b) Odds ratios for protein-coding gene sets related to brain diseases (Fisher’s exact test; Asterisks indicate $p$-values significant after Bonferroni correction; horizontal lines indicate 95% confidence intervals) (c) Protein-protein interaction (PPI) network for the most conserved and divergent protein-coding genes associated with brain diseases. The *RSPH1* gene has accumulated variants with a high combined annotation-dependent depletion (CADD) score, which estimates the deleteriousness of a genetic variant.
Using these 427 highly conserved and 352 highly divergent genes, we first used the Brainspan data available from the specific expression analysis (SEA) to confirm that the population of genes expressed in the cerebellum and the cortex was enriched in conserved genes (Supplementary Figure 7). Despite this conservation, based on the adult Allen Brain atlas, we identified a cluster of brain subregions (within the hypothalamus, cerebral nuclei, and cerebellum) more specifically expressing highly divergent genes (Supplementary Figure 8). Analyses of the prenatal human brain laser microdissection microarray dataset [38] also revealed an excess of divergent protein-coding genes expressed in the medial ganglionic eminence (MGE; OR=2.78[1.05, 7.34], p=0.039; Supplementary Table 5) which is implicated in production of GABAergic interneurons and their migration to neocortex during development [39].

Figure 4 Evolution of protein-coding genes expressed in different cerebellum cell types.
(a) Funnel plot summarizing the evolution of protein-coding genes specifically expressed in different cell types within the cerebellum (Supplementary Table 6). (b) Venn diagram summarizing the divergent protein-coding genes of Homo sapiens (HS), Neanderthals (NH), Denisovans (DS), and Pan troglodytes (PT) specifically expressed in Cluster 47, so-called “ciliated cells” [40].
In single-cell transcriptomic studies of the mouse cerebellum [40,41], we found that cells expressing cilia marker genes, such as \textit{DYNLRB2} and \textit{MEIG1}, were the principal cells with higher levels of expression of the most divergent protein-coding genes (after stringent Bonferroni and bootstrap correction for gene length and GC content, Fig. 4a). Those “ciliated cells” were not anatomically identified in the cerebellum [40], but their associated cillum markers were found to be expressed at the site of the cerebellar granule cells [42]. These cells may, therefore, be a subtype of granule neurons involved in cerebellar function. The most divergent proteins in these ciliated cells code for the tubulin tyrosine ligase like 6 (TTLL6), the DNA topoisomerase III alpha (TOP3A), the dynein cytoplasmic 2 light intermediate chain 1 (DYNC2LI1) and the lebercilin (LCA5) localized to the axoneme of ciliated cells. Given that most protein coding divergence occurs in testes and that the flagella of sperm and cilia of other cells are structurally related, is it possible that the enrichment of ciliated cells among the most divergent genes could be another feature of testis rather than brain divergence. However, only \textit{TTLL6} is highly expressed in testes, suggesting a neural relevance for \textit{DYNC2LI1}, \textit{LCA5}, and \textit{TOP3A}. Interestingly, some of these protein coding genes are also involved in human brain-related ciliopathies such as Joubert syndrome [43] and microcephaly (see below). A similar single-cell transcriptomic analysis of the human cerebral cortex [41] revealed no such strong divergent pattern in any cell type (Supplementary Figure 9).

Finally, we assessed the potential association with brain functions, by extracting 19,244 brain imaging results from 315 fMRI-BOLD studies (T and Z score maps; see Supplementary Table 7 for the complete list) from NeuroVault [44] and comparing the spatial patterns observed with the patterns of gene expression in the Allen Brain atlas [45,46]. The correlation between brain activity and divergent gene expression was stronger in subcortical structures than in the cortex (Wilcoxon \textit{rc}=0.14, \textit{p}=2.5e-248). The brain activity maps that
correlate with the expression pattern of the divergent genes (see Supplementary Table 8 for details) were enriched in social tasks (empathy, emotion recognition, theory of mind, language; Fisher’s exact test $p=2.9\times 10^{-20}$, OR=1.72, CI$_{95\%}$=[1.53, 1.93]; see Supplementary Figure 10 for illustration).

**Divergent protein-coding genes and their relationship to brain disorders**

Our systematic analysis revealed that highly constrained protein-coding genes were more associated with brain diseases or traits than divergent protein-coding genes, particularly for microcephaly ($p=0.002$, OR=0.37, CI$_{95\%}$=[0.16, 0.69], Bonferroni-corrected), intellectual disability ($p=7.9\times 10^{-05}$, OR=0.30 CI$_{95\%}$=[0.16, 0.57], Bonferroni-corrected) and autism ($p=0.0005$, OR=0.26 CI$_{95\%}$=[0.11, 0.59], Bonferroni-corrected) and for diseases associated with myelin (Fisher’s exact test $p=0.005$, OR=0.09, CI$_{95\%}$=[0.01, 0.72], uncorrected) and encephalopathy (Fisher’s exact test $p=0.045$, OR=0.22, CI$_{95\%}$=[0.05, 1.0], uncorrected; Figure 3b). The highly conserved protein-coding genes associated with brain diseases included those encoding tubulins (TUBA1A, TUBB3, TUBB4A), dynamin (DNM1), chromatin remodeling proteins (SMARCA4) and signaling molecules, such as AKT1, DVL1, NOTCH1 and its ligand DLL1, which were associated with neurodevelopmental disorders of different types (Supplementary Table 4). We also identified 31 highly divergent protein-coding genes associated (based on OMIM and HPO data) with several human diseases or conditions, such as micro/macrocephaly, autism or dyslexia.

A comparison of humans and chimpanzees with our common primate ancestor revealed several protein-coding genes associated with micro/macrocephaly with different patterns of evolution in humans and chimpanzees (Fig. 5). Some genes displayed a divergence specifically in the hominin lineage (AHI1, ASXL1, CSPP1, DAG1, FAM111A, GRIP1, NHEJ1, QDPR, RNF135, RNF168, SLX4, TCTN1, and TMEM70) or in the chimpanzee (ARHGAP31, ATRIP, CPT2, CTC1, HDAC6, HEXB, KIF2A, MKKS, MRPS22,
RFT1, TBX6, and WWOX). The PPP1R15B phosphatase gene associated with microcephaly diverged from the common primate ancestor in both taxa. None of the genes related to micro/macrocephaly was divergent only in Homo sapiens (Fig. 5).

Figure 5. Evolution of the protein-coding genes associated with micro- or macrocephaly in humans. Comparison of $\omega_{GC12}$ across taxa for the microcephaly- and macrocephaly-associated genes. Venn diagrams for the conserved (a) and divergent (c) protein-coding genes for Homo sapiens (HS), Neanderthals (NH), Denisovans (DS), and Pan troglodytes (PT). (b) Violin plots of $\omega_{GC}$ for protein-coding genes associated with microcephaly (purple), macrocephaly (green) or both (gray). Scatter plots of $\omega_{GC}$ for the same genes, comparing Homo sapiens with either Neanderthals (d) or Pan troglodytes (e).
We also identified divergent protein-coding genes associated with communication disorders (Fig. 3c), such as autism (CNTNAP4, AHI1, FAN1, SNTG2 and GRIP1) and dyslexia (KIAA0319). Interestingly, these genes diverged from the common primate ancestor only in the hominin lineage, and were strongly conserved in all other taxa (Fig. 6). They all have roles relating to neuronal connectivity (neuronal migration and synaptogenesis) and, within the human brain, were more specifically expressed in the cerebellum, except for GRIP1, which was expressed almost exclusively in the cortex.

Figure 6. Examples of brain disorder-associated protein-coding genes displaying specific divergence in hominins during primate evolution. (a) Representation of 16 genes with dN/dS>1 in Homo sapiens and archaic hominins but dN/dS<1 for other primates. (b) Representation of hominin-specific variants of the AHI1 gene, showing the correspondence with the protein; note how two variants lie within the WP40 functional domains. Red stars indicate variants (CADD>5) relative to the ancestor present in Homo sapiens, Neanderthals, and Denisovans, but not in Pan troglodytes.
The genes associated with autism include *CNTNAP4*, a member of the neurexin protein family involved in correct neurotransmission in the dopaminergic and GABAergic systems [47]. SNTG2 encodes a cytoplasmic peripheral membrane protein that binds to NLGN3 and NLGN4X, two proteins associated with autism [48], and several copy-number variants affecting *SNTG2* have been identified in patients with autism [49]. GRIP1 (glutamate receptor-interacting protein 1) is also associated with microcephaly and encodes a synaptic scaffolding protein that interacts with glutamate receptors. Variants of this gene have repeatedly been associated with autism [50].

We also identified the dyslexia susceptibility gene *KIAA0319*, encoding a protein involved in axon growth inhibition [51,52], as one of the most divergent brain protein-coding genes in humans relative to the common primate ancestor (raw dN/dS=3.9; 9 non-synonymous vs 1 synonymous mutations in *Homo sapiens* compared to the common primate ancestor). The role of *KIAA0319* in dyslexia remains a matter of debate, but its rapid evolution in the hominoid lineage warrants further genetic and functional studies.

Finally, several genes display very high levels of divergence in *Homo sapiens*, but their functions or association with disease remain unknown. For example, the zinc finger protein ZNF491 (raw dN/dS=4.7; 14 non-synonymous vs 1 synonymous mutations in *Homo sapiens* compared to the common primate ancestor) is specifically expressed in the cerebellum and is structurally similar to a chromatin remodeling factor, but its biological role remains to be determined. Another example is the *CCP110* gene, encoding a centrosomal protein resembling ASPM, but not associated with a disease. Its function suggests that this divergent protein-coding gene would be a compelling candidate for involvement in microcephaly in humans. A complete list of the most conserved and divergent protein-coding genes is available in Supplementary Table 4 and on the companion website.
Discussion

Divergent protein-coding genes and brain size in primates

Several protein-coding genes are thought to have played a major role in the increase in brain size in humans. Some of these genes, such as \textit{ARHGAP11B}, \textit{SRGAP2C} and \textit{NOTCH2NL} \cite{7}, are specific to humans, having recently been duplicated \cite{53}. Other studies have suggested that a high degree of divergence in genes involved in micro/macrocephaly may have contributed to the substantial change in brain size during primate evolution \cite{24,54}. Several of these genes, such as \textit{ASPM} \cite{55} and \textit{MCPH1} \cite{56}, seem to have evolved more rapidly in humans. However, the adaptive nature of the evolution of these genes has been called into question \cite{57} and neither of these two genes were on the list of highly divergent protein-coding genes in our analysis (their raw dN/dS value are below 0.8).

Conversely, our systematic detection approach identified the most divergent protein-coding genes in humans for micro/macrocephaly, the top 10 such genes being \textit{FAM111A}, \textit{AHI1}, \textit{CSPP1}, \textit{TCTN1}, \textit{DAG1}, \textit{TMEM70}, \textit{ASXL1}, \textit{RNF168}, \textit{NHEJ1}, \textit{GRIP1}. This list of divergent protein-coding genes associated with micro/macrocephaly in humans can be used to select the best candidate human-specific gene/variants for further genetic and functional analyses, to improve estimates of their contribution to the emergence of anatomic difference between humans and other primates.

Some of these genes may have contributed to differences in brain size and to differences in other morphological features, such as skeleton development. For example, the divergent protein-coding genes \textit{FAM111A} (raw dN/dS=2.99; 7 non-synonymous vs 1 synonymous mutations in \textit{Homo sapiens} compared to the common primate ancestor) and \textit{ASXL1} (raw dN/dS=1.83; 12 non-synonymous vs 3 synonymous mutations in \textit{Homo sapiens} compared to the common primate ancestor) are associated with macrocephaly and...
microcephaly, respectively. Patients with dominant mutations of \textit{FAM111A} are diagnosed with Kenny-Caffey syndrome (KCS). They display impaired skeletal development, with small dense bones, short stature, primary hypoparathyroidism with hypocalcemia and a prominent forehead [58]. The function of FAM111A remains largely unknown, but this protein seems to be crucial to a pathway governing parathyroid hormone production, calcium homeostasis, and skeletal development and growth. By contrast, patients with dominant mutations of \textit{ASXL1} are diagnosed with Bohring-Opitz syndrome, a malformation syndrome characterized by severe intrauterine growth retardation, intellectual disability, trigonocephaly, hirsutism, and flexion of the elbows and wrists with deviation of the wrists and metacarpophalangeal joints [59]. \textit{ASXL1} encodes a chromatin protein required to maintain both the activation and silencing of homeotic genes.

Remarkably, three protein-coding genes (\textit{AHII}, \textit{CSPP1} and \textit{TCTN1}) in the top 5 of the most divergent protein-coding genes, with raw dN/dS>2, are required for both cortical and cerebellar development in humans. They are also associated with Joubert syndrome, a recessive disease characterized by an agenesis of the cerebellar vermis and difficulties coordinating movements. AHII is a positive modulator of classical WNT/ciliary signaling. CSPP1 is involved in cell cycle-dependent microtubule organization and TCTN1 is a regulator of Hedgehog during development.

\textit{AHII} was previously identified as a gene subject to positive selection during evolution of the human lineage [60,61], but, to our knowledge, neither \textit{CSPP1} nor \textit{TCTN1} has previously been described as a diverging during primate evolution. It has been suggested that the accelerated evolution of \textit{AHII} required for ciliogenesis and axonal growth may have played a role in the development of unique motor capabilities, such as bipedalism, in humans [54]. Our findings provide further support for the accelerated evolution of a set of genes associated with ciliogenesis. Indeed, we found that three additional genes involved in Joubert
syndrome, CSPP1, TTLL6, and TCTN1, were among the protein-coding genes that have diverged most during human evolution, and our single-cell analysis revealed that ciliated cells (a subtype of granule neurons) were the main category of cerebellar cells expressing divergent genes.

The possible link between a change in the genetic makeup of the cerebellum and the evolution of human cognition

The emergence of a large cortex was undoubtedly an important step for human cognition, but other parts of the brain, such as the cerebellum, may also have made major contributions to both motricity and cognition. In this study, we showed that the protein-coding genes expressed in the cerebellum were among the most conserved in humans. However, we also identified a set of divergent protein-coding genes with relatively strong expression in the cerebellum and/or for which mutations affected cerebellar function. As discussed above, several genes associated with Joubert syndrome, including AHII, CSPP1, TTLL6, and TCTN1, have diverged in humans and are important for cerebellar development. Furthermore, the most divergent protein-coding genes expressed in the brain include CNTNAP4, FANI, SNTG2, and KIAA0319, which also display high levels of expression in the cerebellum and have been associated with communication disorders, such as autism and dyslexia.

In humans, the cerebellum is associated with higher cognitive functions, such as visuo-spatial skills, the planning of complex movements, procedural learning, attention switching, and sensory discrimination [62]. It plays a key role in temporal processing [63] and in the anticipation and control of behavior, through both implicit and explicit mechanisms [62]. A change in the genetic makeup of the cerebellum would therefore be
expected to have been of great advantage for the emergence of the specific features of human
cognition.

Despite this possible link between the cerebellum and the emergence of human
cognition, much less attention has been paid to this part of the brain than to the cortex, on
which most of the functional studies investigating the role of human-specific genes/variants
have focused. For example, \( \text{SRGAP2C} \) expression is almost exclusively restricted to the
cerebellum in humans, but the ectopic expression of this gene has been studied in mouse
cortex [5,10], in which it triggers human-like neuronal characteristics, such as an increase in
dendritic spine length and density. We therefore suggest that an exploration of human
genes/variants specifically associated with the development and functioning of the
cerebellum might shed new light on the evolution of human cognition.

Limitations

The present results have potential limits in their interpretations. Sources of error in the
alignments (e.g. false orthologous, segmental duplications, errors in ancestral sequence
reconstruction) are still possible and can result in inflated \( \text{dN/dS} \). Moreover, methods to
estimate the proteins evolution are expected to give downwardly biased estimates [64].

However, our GC12 normalization have already proved to correct for most of those biases in
systematic analyses [27] and our raw \( \text{dN/dS} \) values highly correlate with other independent
studies on primates [65]. Moreover, for the enrichment analyses, we used bootstrapping
techniques to better control for potential biases induced by differences in GC content and
gene length, especially for genes implicated in brain disorders [66]. Finally, our data are
openly available on the companion website and allow to check at the variant level which
amino acids changed.
Perspectives

Our systematic analysis of protein sequence diversity confirmed that protein-coding genes relating to brain function are among the most highly conserved in the human genome. The set of divergent protein-coding genes identified here may have played specific roles in the evolution of human cognition, by modulating brain size, neuronal migration and/or synaptic physiology, but further genetic and functional studies would shed new light on the role of these divergent genes. Beyond the brain, this resource will be also be useful for estimating the evolutionary pressure acting on genes related to other biological pathways, particularly those displaying signs of positive selection during primate evolution, such as the reproductive and immune systems.

Materials and Methods

Genetic sequences

Alignments with the reference genome: We collected sequences and reconstructed sequence alignments with the reference human genome version hg19 (release 19, GRCh37.p13). For the primate common ancestor sequence, we used the Ensemble 6-way Enredo-Pecan-Ortheus (EPO) [26] multiple alignments v71, related to human (hg19), chimpanzee (panTro4), gorilla (gorGor3), orangutan (ponAbe2), rhesus macaque (rheMac3), and marmoset (calJac3). For the two ancestral hominins, Altai and Denisovan, we integrated variants detected by Castellano and colleagues [67] into the standard hg19 sequence (http://cdna.eva.mpg.de/neandertal/, date of access 2014-07-03). Finally, we used the whole-
genome alignment of all the primates used in the 6-EPO from the UCSC website (http://hgdownload.soe.ucsc.edu/downloads.html, access online: August 13th, 2015).

**VCF annotation:** We combined the VCF file from Castellano and colleagues [67] with the VCF files generated from the ancestor and primate sequence alignments. The global VCF was annotated with ANNOVAR [68] (version of June 2015), using the following databases: refGene, cytoBand, genomicSuperDups, esp6500siv2_all, 1000g2014oct_all, 1000g2014oct_afr, 1000g2014oct_eas, 1000g2014oct_eur, avsnp142, lj26_all, gerp++elem, popfreq_max, exac03_all, exac03_afr, exac03_amr, exac03_eas, exac03_fin, exac03_nfe, exac03_oth, exac03_sas. We also used the Clinvar database (https://ncbi.nlm.nih.gov/clinvar/, date of access 2016-02-03).

**ω\_GC12 calculation**

Once all the alignments had been collected, we extracted the consensus coding sequences (CCDS) of all protein-coding genes referenced in Ensembl BioMart Grc37, according to the HGNC (date of access 05/05/2015) and NCBI Consensus CDS protein set (date of access 2015-08-10). We calculated the number of non-synonymous mutations N, the number of synonymous mutations S, the ratio of the number of nonsynonymous mutations per non-synonymous site dN, the number of synonymous mutations per synonymous site dS, and their ratio dN/dS —also called \( \omega \)—between all taxa and the ancestor, using the yn00 algorithm implemented in PamL software [69]. We avoided infinite and null results, by calculating a corrected version of dN/dS. If S was null, we set its value to one to avoid having zero as the numerator. The obtained values were validated through the replication of a recent systematic estimation of dN/dS between Homo Sapiens and two great apes [65] (Pan troglodytes and Pongo abelii; Pearson’s r>0.8, p<0.0001; see Fig. S2). Finally, we obtained our \( \omega_{GC12} \) value by correcting for the GC12 content of the genes with a generalized linear model and by calculating a Z-score for each taxon [27]. GC content has been associated with biases in
mutation rates, particularly in primates [70] and humans [71]. We retained only the 11667 genes with 1:1 orthologs in primates (extracted for GRCh37.p13 with Ensemble Biomart, access online: February 27th, 2017).

Gene sets

We used different gene sets, starting at the tissue level and then focusing on the brain and key pathways. For body tissues, we used Illumina Body Map 2.0 RNA-Seq data, corresponding to 16 human tissue types: adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells (for more information: https://personal.broadinstitute.org/mgarber/bodymap_schroth.pdf; data preprocessed with Cufflinks, accessed May 5, 2015 at http://cureffi.org ). We also used the microarray dataset of Su and colleagues [30] (Human U133A/GNF1H Gene Atlas, accessed May 4, 2015 at http://biogps.org). Finally, we also replicated our results with recent RNAseq data from the GTEx Consortium [31] (https://www.gtexportal.org/home/).

For the brain, we used the dataset of Su and colleagues and the Human Protein Atlas data (accessed November 7, 2017 at https://www.proteinatlas.org). For analysis of the biological pathways associated with the brain, we used KEGG (accessed February 25, 2015, at http://www.genome.jp/kegg/), synaptic genes curated by the group of Danielle Posthuma at Vrije Universiteit (accessed September 1, 2014, at https://ctg.cnrc.nl/software/genesets), and mass spectrometry data from Loh and colleagues [72]. Finally, for the diseases associated with the brain, we combined gene sets generated from Human Phenotype Ontology (accessed April 5, 2016, at http://human-phenotype-ontology.github.io) and OMIM (accessed April 5, 2016, at https://omim.org), and curated lists: the 65 risk genes proposed by Sanders and colleagues [73] (TADA), the candidate genes for autism spectrum disorders from SFARI (accessed July 17, 2015 at https://gene.sfari.org), the Developmental Brain Disorder or DBD (accessed July 12, 2016 at https://geisingeradmi.org/care-innovation/studies/dbd-genes/), and
Cancer Census (accessed November 24, 2016 at cancer.sanger.ac.uk/census) data. Note that the combination of HPO & OMIM is the most exhaustive, making it possible to avoid missing potential candidate genes, but this combination does not identify specific associations.

SynGO was generously provided by Matthijs Verhage (access date: January 11, 2019). This ontology is a consistent, evidence-based annotation of synaptic gene products developed by the SynGO consortium (2015-2017) in collaboration with the GO-consortium. It extends the existing Gene Ontology (GO) of the synapse and follows the same dichotomy between biological processes (BP) and cellular components (CC).

For single-cell transcriptomics datasets, we identified the genes specifically highly expressed in each cell type, following the same strategy as used for the other RNAseq datasets. The single-cell data for the developing human cortex were kindly provided by Maximilian Haeussler (available at https://cells.ucsc.edu; access date: October 30, 2018). The single-cell transcriptional atlas data for the developing murine cerebellum [40] were kindly provided by Robert A. Carter (access date: January 29, 2019). For each cell type, we combined expression values cross all available replicates, to guarantee a high signal-to-noise ratio. We then calculated the values for the associated genes in *Homo sapiens* according to the paralogous correspondence between humans and mice (Ensembl Biomart accessed on February 23, 2019).

**Gene nomenclature**

We extracted all the EntrezId of the protein-coding genes for Grc37 from Ensembl Biomart. We used the HGNC database to recover their symbols. For the 46 unmapped genes, we searched the NCBI database manually for the official symbol.
We assessed the possible fixation of variants in the *Homo sapiens* population by first calculating the relative ratio of non-synonymous to synonymous polymorphism (pN/pS) from the 1000 Genomes VCF for all SNPs, for SNPs with a minor allele frequency (MAF) <1% and >5%. SNPs were annotated with ANNOVAR across 1000 Genomes Project (ALL+5 ethnicity groups), ESP6500 (ALL+2 ethnicity groups), ExAC (ALL+7 ethnicity groups), and CG46 (see [http://annovar.openbioinformatics.org/en/latest/user-guide/filter/#popfreqmax-and-popfreqall-annotations for more details](http://annovar.openbioinformatics.org/en/latest/user-guide/filter/#popfreqmax-and-popfreqall-annotations)). We then performed the McDonald–Kreitman test by calculating the neutrality index (NI) as the ratio of raw pN/pS and dN/dS values [74]. We considered the divergent genes to be fixed in the population when NI < 1.

**Protein-protein interaction network**

NeuroVault analyses

We used the NeuroVault website [44] to collect 19,244 brain imaging results from fMRI-BOLD studies (T and Z score maps) and their correlation with the gene expression data [46] of the Allen Brain atlas [45]. The gene expression data of the Allen Brain atlas were normalized and projected into the MNI152 stereotactic space used by NeuroVault, using the spatial coordinates provided by the Allen Brain Institute. An inverse relationship between cortical and subcortical expression dominated the pattern of expression for many genes. We therefore calculated the correlations for the cortex and subcortical structures separately.

Allen Brain data

We downloaded the Allen Brain atlas microarray-based gene data from the Allen Brain website (accessed January 19, 2018 at http://www.brain-map.org). Microarray data were available for six adult brains; the right hemisphere was missing for three donors so we considered only the left hemisphere for our analyses. For each donor, we averaged probes targeting the same gene and falling in the same brain area. We then subjected the data to log normalization and calculated Z-scores: across the 20787 genes for each brain region to obtain expression levels; across the 212 brain areas for each gene to obtain expression specificity. For genes with more than one probe, we averaged the normalized values over all probes available. As a complementary dataset, we also used a mapping of the Allen Brain Atlas onto the 68 brain regions of the Freesurfer atlas [79] (accessed April 4, 2017 at https://figshare.com/articles/A_FreeSurfer_view_of_the_cortical_transcriptome_generated_from_the_Allen_Human_Brain_Atlas/1439749).

Statistics

Enrichment analyses: We first calculated a two-way hierarchical clustering on the normalized dN/dS values ($\omega_{GC}$) across the whole genome (see Fig. 1b; note: 11,667 genes
were included in the analysis to ensure medium-quality coverage for *Homo sapiens*, Neanderthals, Denisovans, and *Pan troglodytes*; see Supplementary table 2). According to 30 clustering indices [80], the best partitioning in terms of evolutionary pressure was into two clusters of genes: constrained (*N*=4825; in HS, mean=-0.88 median=-0.80 SD=0.69) and divergent (*N*=6842; in HS, mean=0.60 median=0.48 sd=0.63. For each cluster, we calculated the enrichment in biological functions in Cytoscape [81] with the BINGO plugin [82]. We used all 12,400 genes as the background. We eliminated redundancy, by first filtering out all the statistically significant Gene Ontology (GO) terms associated with fewer than 10 or more than 1000 genes, and then combining the remaining genes with the EnrichmentMap plugin [83]. We used a *P*-value cutoff of 0.005, an FDR Q-value cutoff of 0.05, and a Jaccard coefficient of 0.5.

For the cell type-specific expression analysis (CSEA; 86), we used the CSEA method with the online tool http://genetics.wustl.edu/jdlab/csea-tool-2/. This method associates gene lists with brain expression profiles across cell types, regions, and time periods.

**Wilcoxon and rank-biserial correlation:** We investigated the extent to which each gene set was significantly more conserved or divergent than expected by chance, by performing Wilcoxon tests on the normalized dN/dS values (ω̄GC) for the genes in the set against zero (the mean value for the genome). We quantified effect size by matched pairs rank-biserial correlation, as described by Kerby [85]. Following non-parametric Wilcoxon signed-rank tests, the rank-biserial correlation was evaluated as the difference between the proportions of negative and positive ranks over the total sum of ranks:

\[
rc = \frac{\sum r_+ - \sum r_-}{\sum r_+ + \sum r_-} = f - u
\]

It corresponds to the difference between the proportion of observations consistent with the hypothesis (f) minus the proportion of observations contradicting the hypothesis (u), thus...
representing an effect size. Like other correlational measures, its value ranges from minus one to plus one, with a value of zero indicating no relationship. In our case, a negative rank-biserial correlation corresponds to a gene set in which more genes have negative \( \omega_{GC} \) values than positive values, revealing a degree of conservation greater than the mean for all genes (i.e. \( \omega_{GC} = 0 \)). Conversely, a positive rank-biserial correlation corresponds to a gene set that is more divergent than expected by chance (i.e. taking randomly the same number of genes across the whole genome; correction for the potential biases for GC content and CDS length are done at the bootstrap level). All statistics relating to Figures 1d, 2a and 2b are summarized in Supplementary table 3.

**Validation by resampling:** We also used bootstrapping to correct for potential bias in the length of the coding sequence or the global specificity of gene expression (Tau, see the methods from Kryuchkova-Mostacci and Robinson-Rechavi in [86]). For each of the 10000 permutations, we randomly selected the same number of genes as for the sample of genes from the total set of genes for which dN/dS was not missing. We corrected for CCDS length and GC content by bootstrap resampling. We estimated significance, to determine whether the null hypothesis could be rejected, by calculating the number of bootstrap draws \( B_i \) falling below and above the observed measurement \( m \). The related empirical \( p \)-value was calculated as follows:

\[
p = 2 \times \min \left( \frac{1 + \sum_i B_i \geq m}{N + 1}, \frac{1 + \sum_i B_i \leq m}{N + 1} \right)
\]

**Data & code availability:** All the data and code supporting the findings of this study are available from our resource website: [http://neanderthal.pasteur.fr](http://neanderthal.pasteur.fr)
Acknowledgments

We thank J-P. Changeux, L. Quintana-Murci, E. Patin, G. Laval, B. Arcangioli, D. DiGregorio, L. Bally-Cuif, A. Chedotal, C. Berthelot, H. Roest Crollius, and V. Warrier for advice and comments, and the members of the Human Genetics and Cognitive Functions laboratory for helpful discussions. We also thank C. Gorgolewski, R. Carter, M. Haeussler, M. Verhage and the SynGO consortium for providing key datasets without which this work would not have been possible. This work was supported by the Institut Pasteur; Centre National de la Recherche Scientifique; Paris Diderot University; the Fondation pour la Recherche Médicale [DBI20141231310]; the Human Brain Project; the Cognacq-Jay Foundation; the Bettencourt-Schueller Foundation; and the Agence Nationale de la Recherche (ANR) [SynPathy]. This research was supported by the Laboratory of Excellence GENMED (Medical Genomics) grant no. ANR-10-LABX-0013, Bio-Psy and by the INCEPTION program ANR-16-CONV-0005, all managed by the ANR part of the Investments for the Future program. The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Author contributions

G.D. and T.B. devised the project and came up with the main conceptual ideas. G.D. developed the methods, performed the analyses, and designed the figures. G.D. and T.B. discussed the results and wrote the manuscript. S.M. developed the companion website.
References


