

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

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

4

5 Guillaume Dumas<sup>a</sup>, Simon Malesys<sup>a</sup> and Thomas Bourgeron<sup>a</sup>

6 <sup>a</sup> Human Genetics and Cognitive Functions, Institut Pasteur, UMR3571 CNRS, Université de Paris, Paris,

7 (75015) France

8

9

10 *Corresponding author:*

11 Guillaume Dumas

12 Human Genetics and Cognitive Functions

13 Institut Pasteur

14 75015 Paris, France

15 Phone: +33 6 28 25 56 65

16 [guillaume.dumas@pasteur.fr](mailto:guillaume.dumas@pasteur.fr)

17 **Abstract**

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

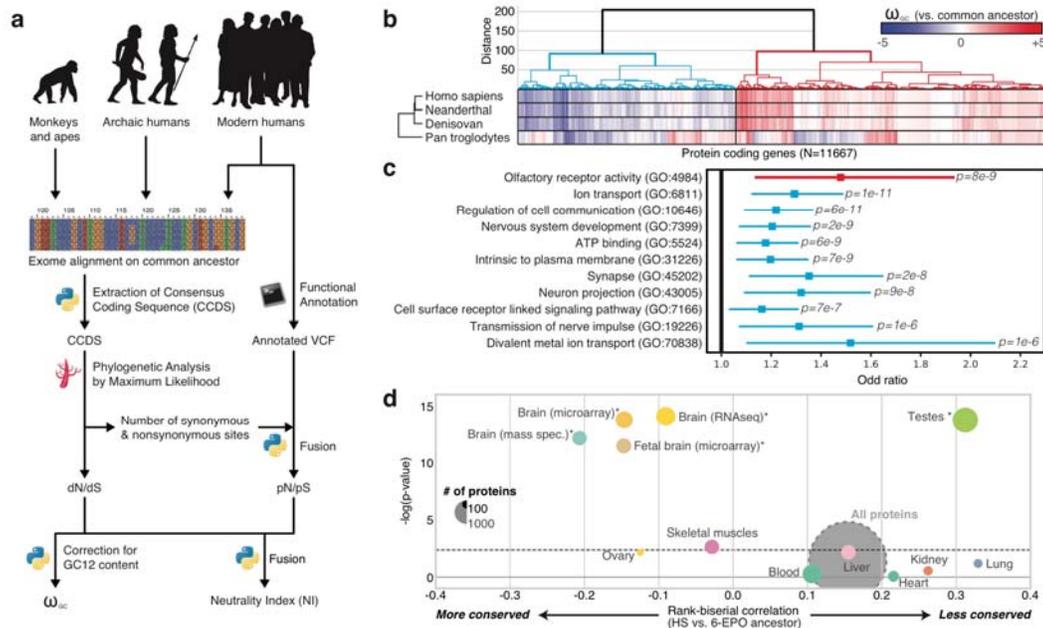
### 33 **Introduction**

34 Modern humans (*Homo sapiens*) can perform complex cognitive tasks well and communicate  
35 with their peers [1]. Anatomic differences between the brains of humans and other primates  
36 are well documented (e.g. cortex size, prefrontal white matter thickness, lateralization), but  
37 the way in which the human brain evolved remains a matter of debate [2]. A recent study of  
38 endocranial casts of *Homo sapiens* fossils indicates that, brain size in early *Homo sapiens*,  
39 300,000 years ago, was already within the range of that in present-day humans [3]. However,  
40 brain shape, evolved more gradually within the *Homo sapiens* lineage, reaching its current  
41 form between about 100,000 and 35,000 years ago. It has also been suggested that the  
42 enlargement of the prefrontal cortex relative to the motor cortex in humans is mirrored in the  
43 cerebellum by an enlargement of the regions of the cerebellum connected to the prefrontal  
44 cortex [4]. These anatomic processes of tandem evolution in the brain paralleled the  
45 emergence of motor and cognitive abilities, such as bipedalism, planning, language, and  
46 social awareness, which are particularly well developed in humans.

47 Genetic differences in primates undoubtedly contributed to these brain and cognitive  
48 differences, but the genes or variants involved remain largely unknown. Indeed,  
49 demonstrating that a genetic variant is adaptive requires strong evidence at both the genetic  
50 and functional levels. Only few genes have been shown to be human-specific. They include  
51 *SRGAP2C* [5], *ARHGAP11B* [6] and *NOTCH2NL* [7], which emerged through recent gene  
52 duplication in the *Homo* lineage [8]. Remarkably, the expression of these human specific  
53 genes in the mouse brain expand cortical neurogenesis [6,7,9,10]. Several genes involved in  
54 brain function have been shown to display accelerated coding region evolution in humans.  
55 For example, *FOXP2* has been associated with verbal apraxia and *ASPM* with microcephaly  
56 [11,12]. Functional studies have also shown that mice carrying a “humanized” version of  
57 *FOXP2* display qualitative changes in ultrasonic vocalization [13]. However, these reports

58 targeting only specific genes sometimes provide contradictory results [14]. Other studies  
59 have reported sequence conservation to be stronger in the protein-coding genes of the brain  
60 than in those of other tissues [15–17], suggesting that the main substrate of evolution in the  
61 brain is regulatory changes in gene expression [18–20] and splicing [21]. In addition, several  
62 recent studies have recently explored the genes subjected to the highest degrees of constraint  
63 during primate evolution or in human populations, to improve estimations of the  
64 pathogenicity of variants identified in patients with genetic disorders [22,23]. By contrast,  
65 few studies have systematically detected genes that have diverged during primate evolution  
66 [24,25].

67 We describe here an exhaustive screening of all protein-coding genes for conservation  
68 and divergence from the common primate ancestor, making use of rich datasets of brain  
69 single-cell transcriptomics, proteomics and imaging to investigate the relationships between  
70 these genes and brain structure, function, and diseases.



71

72 **Figure 1 Evolution of protein-coding genes across tissues and biological functions. (a)**  
 73 **Analysis pipeline for the extraction of  $\omega_{GC12}$ , a corrected and normalized measurement**  
 74 **of evolution of protein-coding genes that behaves like a Z-score and takes into account**  
 75 **the GC content of codons. (b) Hierarchical clustering, on the basis of  $\omega_{GC12}$ , across all**  
 76 **protein-coding genes (1:1 orthologs in hominins with medium coverage; See**  
 77 **Supplementary Table 1). (c) Gene ontology (GO) enrichments for the red and blue**  
 78 **clusters in panel b (See Supplementary Table 2 for all GO terms). Horizontal lines**  
 79 **indicate the 95% confidence intervals. (d) Funnel plot summarizing the evolution of**  
 80 **protein-coding genes specifically expressed in different tissues of the human body**  
 81 **(Supplementary Table 3). The dashed horizontal line indicates the threshold for**  
 82 **significance after Bonferroni correction. Stars indicate the set of genes for which**  
 83 **statistical significance was achieved in multiple comparisons after correction, with a**  
 84 **bootstrap taking GC12 content and coding sequence length into account. HS: *Homo***  
 85 ***sapiens*; 6-EPO ancestor: the reconstructed ancestral genome of primates based on**  
 86 **alignments of human, chimpanzee, gorilla, orangutan, rhesus macaque, and marmoset**  
 87 **genomes.**

## 88 **Results**

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

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

107 Using the  $\omega_{GC12}$  for all protein-coding genes in *Homo sapiens*, Denisovans,  
108 Neanderthals, and *Pan troglodytes*, we identified two distinct clusters in hominins (Fig. 1b  
109 and Supplementary Table 1): one containing divergent protein-coding genes, enriched in  
110 olfactory genes (OR=1.48,  $p=8.4e-9$ ), and one with conserved protein-coding genes, enriched  
111 in brain-related biological functions (Fig. 1c and Supplementary Table 2). This second cluster

112 revealed a particularly strong conservation of genes encoding proteins involved in nervous  
113 system development (OR=1.2,  $p=2.4e-9$ ) and synaptic transmission (OR=1.35,  $p=1.7e-8$ ).

114 We investigated the possible enrichment of specific tissues in conserved and  
115 divergent proteins by analyzing RNAseq (Illumina Bodymap2 and GTEx), microarray and  
116 proteomics datasets (Methods). For expression data, we evaluated the specificity of genes by  
117 normalizing their profile across tissues (Supplementary Fig. 3). The results confirmed a  
118 higher degree of conservation for protein-coding genes expressed in the brain (Wilcoxon rank  
119 correlation ( $rc$ )=-0.1,  $p=4.1e-12$ , bootstrap corrected for gene length and GC content) than for  
120 those expressed elsewhere in the body, with the greatest divergence observed for genes  
121 expressed in the testis (Wilcoxon  $rc=0.3$ ,  $p=7.8e-11$ , bootstrap corrected for gene length and  
122 GC content; Fig. 1d, see also Supplementary Fig. 4 and 5). This conservation of brain  
123 protein-coding genes was replicated with two other datasets (MicroArray: Wilcoxon OR=-  
124 0.18,  $p=1.8e-12$ ; mass spectrometry: Wilcoxon  $rc=-0.21$ ,  $p=1.55e-9$ ; bootstrap corrected for  
125 gene length and GC content).

126

### 127 **Conservation of protein-coding genes relating to nervous system substructure and** 128 **neuronal functions**

129 We then used microarray [30] and RNAseq [31] data to investigate the evolutionary pressures  
130 acting on different regions of the central nervous system. Three central nervous system  
131 substructures appeared to have evolved under the highest level of purifying selection at the  
132 protein sequence level ( $\omega_{GC12}<2$ , i.e. highly conserved): (i) the cerebellum (Wilcoxon  $rc=-$   
133 0.29,  $p=5.5e-6$ , Bonferroni corrected) and the cerebellar peduncle (Wilcoxon  $rc=-0.11$ ,  
134  $p=3.2e-4$ , bootstrap corrected for gene length and GC content), (ii) the amygdala (Wilcoxon  
135  $rc=-0.11$ ,  $p=4.1e-6$ , bootstrap corrected for gene length and GC content), and, more

136 surprisingly, (iii) the prefrontal cortex (Wilcoxon  $rc=-0.1$ ,  $p=5.7e-10$ , bootstrap corrected for  
137 gene length and GC content; Fig. 2a, see also Supplementary Table 3). Indeed, it has been  
138 suggested that the prefrontal cortex is one of the most divergent brain structure in human  
139 evolution [32], this diversity being associated with high-level cognitive function [33]. Only  
140 one brain structure was more divergent than expected: the superior cervical ganglion  
141 (Wilcoxon  $rc=0.22$ ,  $p=1e-6$ , bootstrap corrected for gene length and GC content). This  
142 structure provides sympathetic innervation to many organs and is associated with the archaic  
143 functions of fight-or-flight response. The divergent genes expressed in the superior cervical  
144 ganglion include *CARF*, which was found to be specifically divergent in the genus *Homo*.  
145 This gene encodes a calcium-responsive transcription factor that regulates the neuronal  
146 activity-dependent expression of *BDNF* [34] and a set of singing-induced genes in the song  
147 nuclei of the zebra finch, a songbird capable of vocal learning [35]. This gene had a raw  
148 dN/dS of 2.44 (7 non-synonymous vs 1 synonymous mutations in *Homo sapiens* compared to  
149 the common primate ancestor) and was found to be one of the most divergent protein-coding  
150 genes expressed in the human brain.

151 We then investigated the possible enrichment of conserved and divergent genes in  
152 brain-specific gene ontology terms. All pathways displayed high overall levels of  
153 conservation, but genes encoding proteins involved in glutamatergic and GABAergic  
154 neurotransmission were generally more conserved (Wilcoxon  $rc=-0.25$ ;  $p=9.8e-6$ , Bonferroni  
155 corrected) than those encoding proteins involved in dopamine and peptide neurotransmission  
156 and intracellular trafficking (Fig. 2b, see also Supplementary Fig. 6 and Supplementary Table  
157 3). The recently released ontology of the synapse provided by the SynGO consortium  
158 (<http://syngoportal.org>) was incorporated into this analysis, not only confirming the globally  
159 strong conservation of the synapse, but also revealing its close relationship to trans-synaptic  
160 signaling processes (Wilcoxon  $rc=-0.21$ ,  $p=4.5e-5$ , Bonferroni corrected) and to postsynaptic

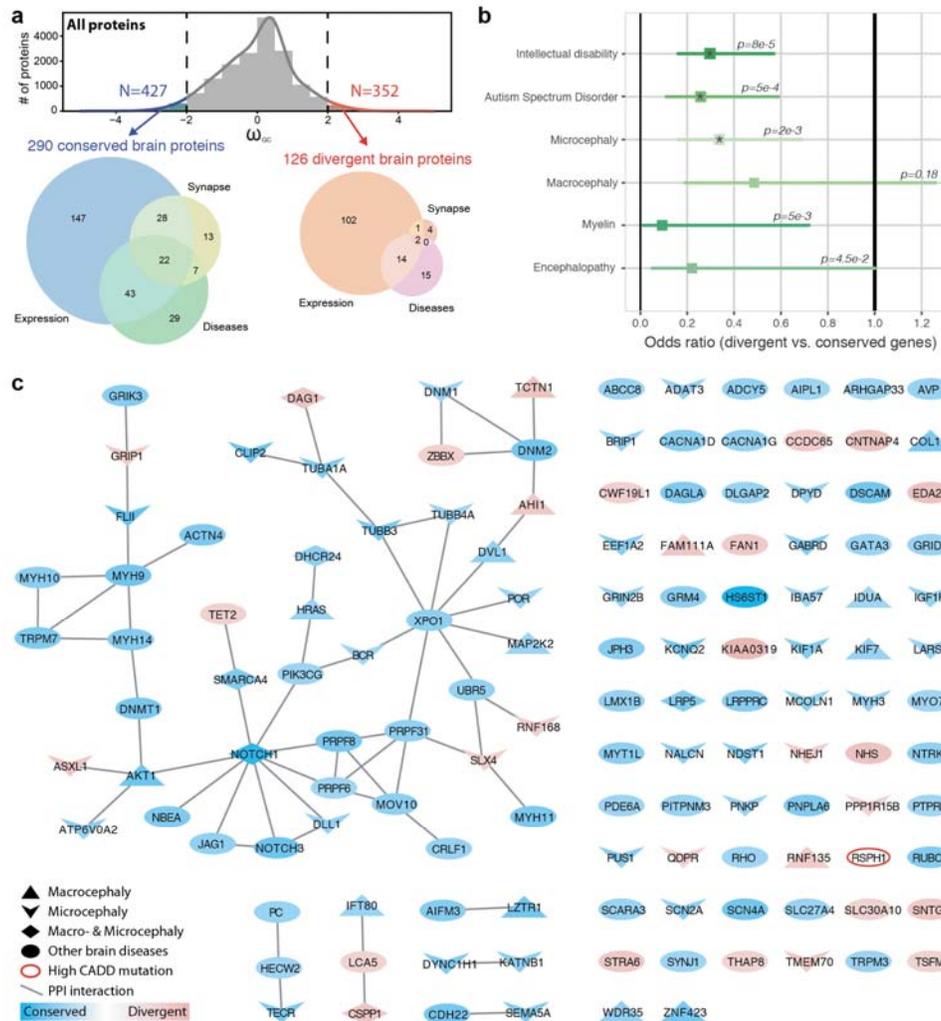


173 **d) SynGO sunburst plots showing nested statistically conserved (blue) biological**  
174 **processes and cellular components of the synapse.**

175

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

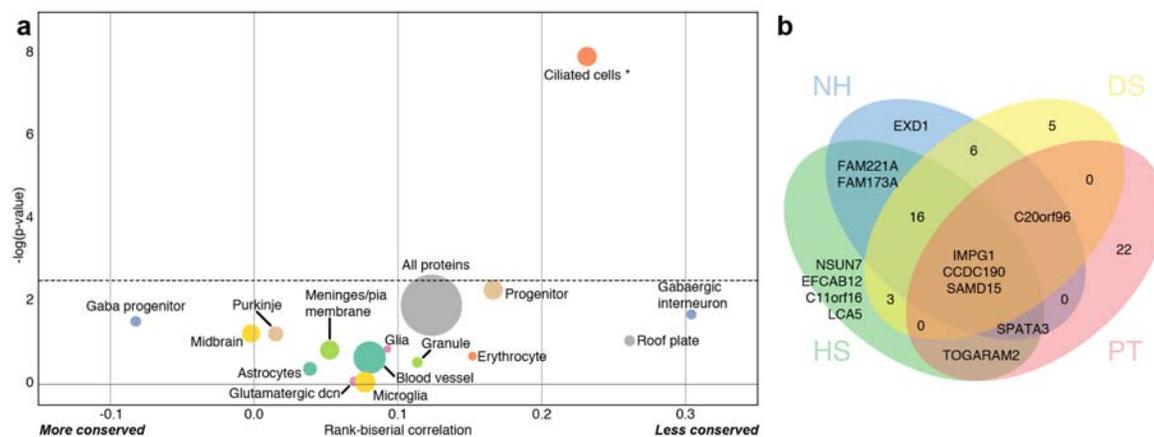
178 We focused on the genes situated at the extremes of the  $\omega_{GC12}$  distribution ( $>2SD$ ;  
179 Fig. 3a; Supplementary Table 4) and those fixed in the modern *Homo sapiens* population  
180 (neutrality index $<1$ ), to ensure that we analyzed the most-divergent protein-coding genes.  
181 Only 126 of these 352 highly divergent protein-coding genes were brain-related  
182 (impoverishment for brain genes, Fisher's exact test OR=0.66,  $p=1e-4$ ), listed as synaptic  
183 genes [36,37], specifically expressed in the brain ( $+2SD$  for specific expression) or related to  
184 a brain disease (extracted systematically from Online Mendelian Inheritance in Man -  
185 OMIM: <https://www.omim.org> and Human Phenotype Ontology - HPO:  
186 <https://hpo.jax.org/app/>). For comparison, we also extracted the 427 most strongly conserved  
187 protein-coding genes, 290 of which were related to the brain categories listed above  
188 (enrichment for brain genes, Fisher's exact test OR=1.26,  $p=0.0032$ ).



189

190 **Figure 3 Brain protein-coding genes and human diseases. (a) Distribution of  $\omega_{GC12}$  and**  
 191 **Venn diagrams describing the most conserved and divergent protein-coding genes**  
 192 **specifically expressed in the brain, related to the synapse, or brain diseases**  
 193 **(Supplementary Table 4). (b) Odds ratios for protein-coding gene sets related to brain**  
 194 **diseases (Fisher's exact test; Asterisks indicate  $p$ -values significant after Bonferroni**  
 195 **correction; horizontal lines indicate 95% confidence intervals) (c) Protein-protein**  
 196 **interaction (PPI) network for the most conserved and divergent protein-coding genes**  
 197 **associated with brain diseases. The *RSPHI* gene has accumulated variants with a high**  
 198 **combined annotation-dependent depletion (CADD) score, which estimates the**  
 199 **deleteriousness of a genetic variant.**

200 Using these 427 highly conserved and 352 highly divergent genes, we first used the  
 201 Brainspan data available from the specific expression analysis (SEA) to confirm that the  
 202 population of genes expressed in the cerebellum and the cortex was enriched in conserved  
 203 genes (Supplementary Figure 7). Despite this conservation, based on the adult Allen Brain  
 204 atlas, we identified a cluster of brain subregions (within the hypothalamus, cerebral nuclei,  
 205 and cerebellum) more specifically expressing highly divergent genes (Supplementary Figure  
 206 8). Analyses of the prenatal human brain laser microdissection microarray dataset [38] also  
 207 revealed an excess of divergent protein-coding genes expressed in the medial ganglionic  
 208 eminence (MGE; OR=2.78[1.05, 7.34], p=0.039; Supplementary Table 5) which is  
 209 implicated in production of GABAergic interneurons and their migration to neocortex during  
 210 development [39].  
 211



212  
 213 **Figure 4 Evolution of protein-coding genes expressed in different cerebellum cell types.**  
 214 **(a) Funnel plot summarizing the evolution of protein-coding genes specifically expressed**  
 215 **in different cell types within the cerebellum (Supplementary Table 6). (b) Venn diagram**  
 216 **summarizing the divergent protein-coding genes of *Homo sapiens* (HS), Neanderthals**  
 217 **(NH), Denisovans (DS), and *Pan troglodytes* (PT) specifically expressed in Cluster 47, so-**  
 218 **called “ciliated cells” [40].**

219

220 In single-cell transcriptomic studies of the mouse cerebellum [40,41], we found that cells  
221 expressing cilium marker genes, such as *DYNLRB2* and *MEIG1*, were the principal cells with  
222 higher levels of expression of the most divergent protein-coding genes (after stringent  
223 Bonferroni and bootstrap correction for gene length and GC content, Fig. 4a). Those “ciliated  
224 cells” were not anatomically identified in the cerebellum [40], but their associated cilium  
225 markers were found to be expressed at the site of the cerebellar granule cells [42]. These cells  
226 may, therefore, be a subtype of granule neurons involved in cerebellar function. The most  
227 divergent proteins in these ciliated cells code for the tubulin tyrosine ligase like 6 (*TTL6*),  
228 the DNA topoisomerase III alpha (*TOP3A*), the dynein cytoplasmic 2 light intermediate  
229 chain 1 (*DYNC2LI1*) and the lebercilin (*LCA5*) localized to the axoneme of ciliated cells.  
230 Given that most protein coding divergence occurs in testes and that the flagella of sperm and  
231 cilia of other cells are structurally related, is it possible that the enrichment of ciliated cells  
232 among the most divergent genes could be another feature of testis rather than brain  
233 divergence. However, only *TTL6* is highly expressed in testes, suggesting a neural relevance  
234 for *DYNC2LI1*, *LCA5*, and *TOP3A*. Interestingly, some of these protein coding genes are also  
235 involved in human brain-related ciliopathies such as Joubert syndrome [43] and microcephaly  
236 (see below). A similar single-cell transcriptomic analysis of the human cerebral cortex [41]  
237 revealed no such strong divergent pattern in any cell type (Supplementary Figure 9).

238 Finally, we assessed the potential association with brain functions, by extracting 19,244 brain  
239 imaging results from 315 fMRI-BOLD studies (T and Z score maps; see Supplementary  
240 Table 7 for the complete list) from NeuroVault [44] and comparing the spatial patterns  
241 observed with the patterns of gene expression in the Allen Brain atlas [45,46]. The  
242 correlation between brain activity and divergent gene expression was stronger in subcortical  
243 structures than in the cortex (Wilcoxon  $rc=0.14$ ,  $p=2.5e-248$ ). The brain activity maps that

244 correlate with the expression pattern of the divergent genes (see Supplementary Table 8 for  
245 details) were enriched in social tasks (empathy, emotion recognition, theory of mind,  
246 language; Fisher's exact test  $p=2.9e-20$ , OR=1.72, CI<sub>95%</sub>=[1.53, 1.93]; see Supplementary  
247 Figure 10 for illustration).

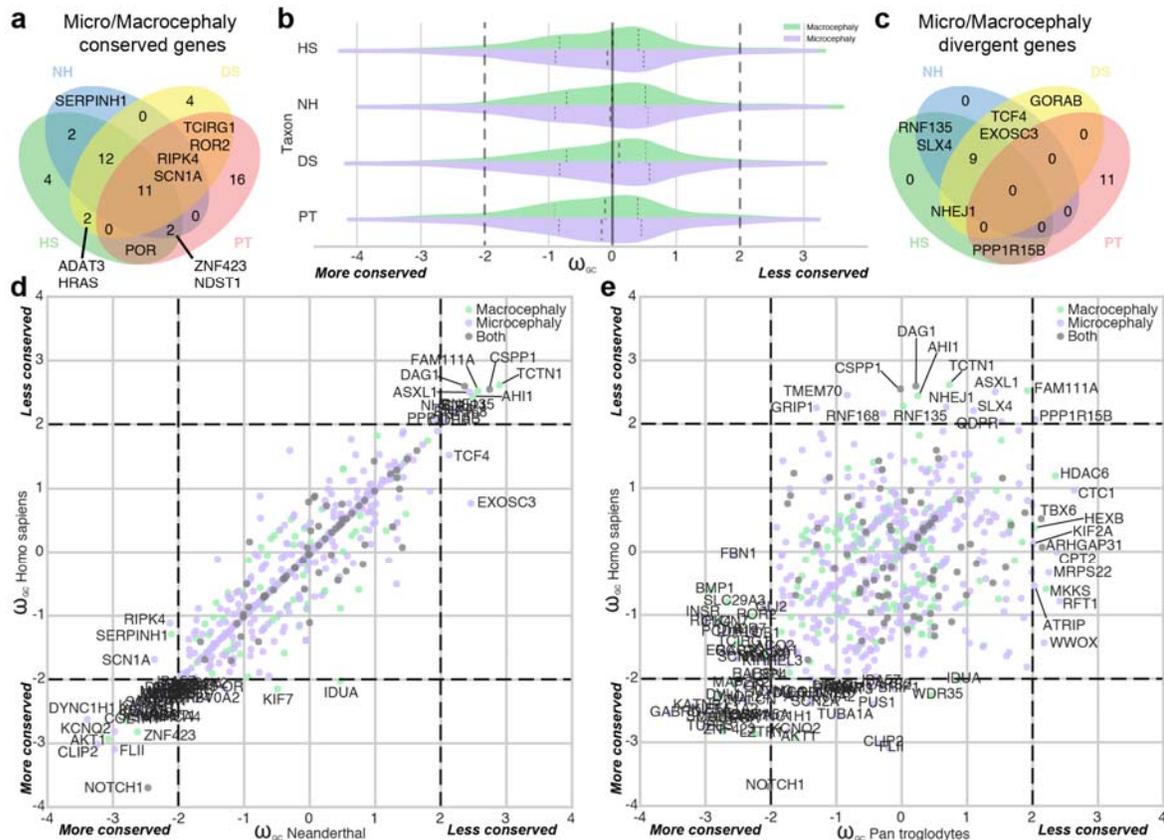
### 248 **Divergent protein-coding genes and their relationship to brain disorders**

249 Our systematic analysis revealed that highly constrained protein-coding genes were more  
250 associated with brain diseases or traits than divergent protein-coding genes, particularly for  
251 microcephaly ( $p=0.002$ , OR=0.37, CI<sub>95%</sub>=[0.16, 0.69], Bonferroni-corrected), intellectual  
252 disability ( $p=7.91e-05$ , OR=0.30 CI<sub>95%</sub>=[0.16, 0.57], Bonferroni-corrected) and autism  
253 ( $p=0.0005$ , OR=0.26, CI<sub>95%</sub>=[0.11, 0.59], Bonferroni-corrected) and for diseases associated  
254 with myelin (Fisher's exact test  $p=0.005$ , OR=0.09, CI<sub>95%</sub>=[0.01, 0.72], uncorrected) and  
255 encephalopathy (Fisher's exact test  $p=0.045$ , OR=0.22, CI<sub>95%</sub>=[0.05, 1.0], uncorrected;  
256 Figure 3b). The highly conserved protein-coding genes associated with brain diseases  
257 included those encoding tubulins (TUBA1A, TUBB3, TUBB4A), dynamin (DNM1),  
258 chromatin remodeling proteins (SMARCA4) and signaling molecules, such as AKT1, DVL1,  
259 NOTCH1 and its ligand DLL1, which were associated with neurodevelopmental disorders of  
260 different types (Supplementary Table 4). We also identified 31 highly divergent protein-  
261 coding genes associated (based on OMIM and HPO data) with several human diseases or  
262 conditions, such as micro/macrocephaly, autism or dyslexia.

263 A comparison of humans and chimpanzees with our common primate ancestor  
264 revealed several protein-coding genes associated with micro/macrocephaly with different  
265 patterns of evolution in humans and chimpanzees (Fig. 5). Some genes displayed a  
266 divergence specifically in the hominin lineage (*AHII*, *ASXLI*, *CSPP1*, *DAG1*, *FAM111A*,  
267 *GRIP1*, *NHEJ1*, *QDPR*, *RNF135*, *RNF168*, *SLX4*, *TCTN1*, and *TMEM70*) or in the  
268 chimpanzee (*ARHGAP31*, *ATRIP*, *CPT2*, *CTC1*, *HDAC6*, *HEXB*, *KIF2A*, *MKKS*, *MRPS22*,

269 *RFT1*, *TBX6*, and *WWOX*). The *PPP1R15B* phosphatase gene associated with microcephaly  
 270 diverged from the common primate ancestor in both taxa. None of the genes related to  
 271 micro/macrocephaly was divergent only in *Homo sapiens* (Fig. 5).

272



273

274 **Figure 5. Evolution of the protein-coding genes associated with micro- or macrocephaly**  
 275 **in humans. Comparison of  $\omega_{GC12}$  across taxa for the microcephaly- and macrocephaly-**  
 276 **associated genes. Venn diagrams for the conserved (a) and divergent (c) protein-coding**  
 277 **genes for *Homo sapiens* (HS), Neanderthals (NH), Denisovans (DS), and *Pan troglodytes***  
 278 **(PT). (b) Violin plots of  $\omega_{GC}$  for protein-coding genes associated with microcephaly**  
 279 **(purple), macrocephaly (green) or both (gray). Scatter plots of  $\omega_{GC}$  for the same genes,**  
 280 **comparing *Homo sapiens* with either Neanderthals (d) or *Pan troglodytes* (e).**

281



298           The genes associated with autism include *CNTNAP4*, a member of the neurexin  
299 protein family involved in correct neurotransmission in the dopaminergic and GABAergic  
300 systems [47]. *SNTG2* encodes a cytoplasmic peripheral membrane protein that binds to  
301 NLGN3 and NLGN4X, two proteins associated with autism [48], and several copy-number  
302 variants affecting *SNTG2* have been identified in patients with autism [49]. GRIP1 (glutamate  
303 receptor-interacting protein 1) is also associated with microcephaly and encodes a synaptic  
304 scaffolding protein that interacts with glutamate receptors. Variants of this gene have  
305 repeatedly been associated with autism [50].

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

312           Finally, several genes display very high levels of divergence in *Homo sapiens*, but  
313 their functions or association with disease remain unknown. For example, the zinc finger  
314 protein ZNF491 (raw dN/dS=4.7; 14 non-synonymous vs 1 synonymous mutations in *Homo*  
315 *sapiens* compared to the common primate ancestor) is specifically expressed in the  
316 cerebellum and is structurally similar to a chromatin remodeling factor, but its biological role  
317 remains to be determined. Another example is the *CCPI10* gene, encoding a centrosomal  
318 protein resembling ASPM, but not associated with a disease. Its function suggests that this  
319 divergent protein-coding gene would be a compelling candidate for involvement in  
320 microcephaly in humans. A complete list of the most conserved and divergent protein-coding  
321 genes is available in Supplementary Table 4 and on the companion website.

## 322 **Discussion**

### 323 **Divergent protein-coding genes and brain size in primates**

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

333         Conversely, our systematic detection approach identified the most divergent protein-  
334 coding genes in humans for micro/macrocephaly, the top 10 such genes being *FAM111A*,  
335 *AHII*, *CSPP1*, *TCTNI*, *DAG1*, *TMEM70*, *ASXLI*, *RNF168*, *NHEJ1*, *GRIP1*. This list of  
336 divergent protein-coding genes associated with micro/macrocephaly in humans can be used  
337 to select the best candidate human-specific gene/variants for further genetic and functional  
338 analyses, to improve estimates of their contribution to the emergence of anatomic difference  
339 between humans and other primates.

340         Some of these genes may have contributed to differences in brain size and to  
341 differences in other morphological features, such as skeleton development. For example, the  
342 divergent protein-coding genes *FAM111A* (raw dN/dS=2.99; 7 non-synonymous vs 1  
343 synonymous mutations in *Homo sapiens* compared to the common primate ancestor) and  
344 *ASXLI* (raw dN/dS=1.83; 12 non-synonymous vs 3 synonymous mutations in *Homo sapiens*  
345 compared to the common primate ancestor) are associated with macrocephaly and

346 microcephaly, respectively. Patients with dominant mutations of *FAM111A* are diagnosed  
347 with Kenny-Caffey syndrome (KCS). They display impaired skeletal development, with  
348 small dense bones, short stature, primary hypoparathyroidism with hypocalcemia and a  
349 prominent forehead [58]. The function of *FAM111A* remains largely unknown, but this  
350 protein seems to be crucial to a pathway governing parathyroid hormone production, calcium  
351 homeostasis, and skeletal development and growth. By contrast, patients with dominant  
352 mutations of *ASXL1* are diagnosed with Bohring-Opitz syndrome, a malformation syndrome  
353 characterized by severe intrauterine growth retardation, intellectual disability, trigonocephaly,  
354 hirsutism, and flexion of the elbows and wrists with deviation of the wrists and  
355 metacarpophalangeal joints [59]. *ASXL1* encodes a chromatin protein required to maintain  
356 both the activation and silencing of homeotic genes.

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

364 *AHII* was previously identified as a gene subject to positive selection during  
365 evolution of the human lineage [60,61], but, to our knowledge, neither *CSPP1* nor *TCTN1*  
366 has previously been described as a diverging during primate evolution. It has been suggested  
367 that the accelerated evolution of *AHII* required for ciliogenesis and axonal growth may have  
368 played a role in the development of unique motor capabilities, such as bipedalism, in humans  
369 [54]. Our findings provide further support for the accelerated evolution of a set of genes  
370 associated with ciliogenesis. Indeed, we found that three additional genes involved in Joubert

371 syndrome, *CSPP1*, *TLL6*, and *TCTNI*, were among the protein-coding genes that have  
372 diverged most during human evolution, and our single-cell analysis revealed that ciliated  
373 cells (a subtype of granule neurons) were the main category of cerebellar cells expressing  
374 divergent genes.

375

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

378 The emergence of a large cortex was undoubtedly an important step for human cognition, but  
379 other parts of the brain, such as the cerebellum, may also have made major contributions to  
380 both motricity and cognition. In this study, we showed that the protein-coding genes  
381 expressed in the cerebellum were among the most conserved in humans. However, we also  
382 identified a set of divergent protein-coding genes with relatively strong expression in the  
383 cerebellum and/or for which mutations affected cerebellar function. As discussed above,  
384 several genes associated with Joubert syndrome, including *AHII*, *CSPP1*, *TLL6*, and  
385 *TCTNI*, have diverged in humans and are important for cerebellar development. Furthermore,  
386 the most divergent protein-coding genes expressed in the brain include *CNTNAP4*, *FANI*,  
387 *SNTG2*, and *KIAA0319*, which also display high levels of expression in the cerebellum and  
388 have been associated with communication disorders, such as autism and dyslexia.

389 In humans, the cerebellum is associated with higher cognitive functions, such as  
390 visuo-spatial skills, the planning of complex movements, procedural learning, attention  
391 switching, and sensory discrimination [62]. It plays a key role in temporal processing [63]  
392 and in the anticipation and control of behavior, through both implicit and explicit  
393 mechanisms [62]. A change in the genetic makeup of the cerebellum would therefore be

394 expected to have been of great advantage for the emergence of the specific features of human  
395 cognition.

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

405

#### 406 **Limitations**

407 The present results have potential limits in their interpretations. Sources of error in the  
408 alignments (e.g. false orthologous, segmental duplications, errors in ancestral sequence  
409 reconstruction) are still possible and can result in inflated dN/dS. Moreover, methods to  
410 estimate the proteins evolution are expected to give downwardly biased estimates [64].  
411 However, our GC12 normalization have already proved to correct for most of those biases in  
412 systematic analyses [27] and our raw dN/dS values highly correlate with other independent  
413 studies on primates [65]. Moreover, for the enrichment analyses, we used bootstrapping  
414 techniques to better control for potential biases induced by differences in GC content and  
415 gene length, especially for genes implicated in brain disorders [66]. Finally, our data are  
416 openly available on the companion website and allow to check at the variant level which  
417 amino acids changed.

418

419

## 420 **Perspectives**

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

430

## 431 **Materials and Methods**

### 432 **Genetic sequences**

433 **Alignments with the reference genome:** We collected sequences and reconstructed  
434 sequence alignments with the reference human genome version hg19 (release 19,  
435 GRCh37.p13). For the primate common ancestor sequence, we used the Ensemble 6-way  
436 Enredo-Pecan-Ortheus (EPO) [26] multiple alignments v71, related to human (hg19),  
437 chimpanzee (panTro4), gorilla (gorGor3), orangutan (ponAbe2), rhesus macaque (rheMac3),  
438 and marmoset (calJac3). For the two ancestral hominins, Altai and Denisovan, we integrated  
439 variants detected by Castellano and colleagues [67] into the standard hg19 sequence  
440 (<http://cdna.eva.mpg.de/neandertal/>, date of access 2014-07-03). Finally, we used the whole-

441 genome alignment of all the primates used in the 6-EPO from the UCSC website  
442 (<http://hgdownload.soe.ucsc.edu/downloads.html>, access online: August 13<sup>th</sup>, 2015).

443 **VCF annotation:** We combined the VCF file from Castellano and colleagues [67] with the  
444 VCF files generated from the ancestor and primate sequence alignments. The global VCF  
445 was annotated with ANNOVAR [68] (version of June 2015), using the following databases:  
446 refGene, cytoBand, genomicSuperDups, esp6500siv2\_all, 1000g2014oct\_all,  
447 1000g2014oct\_afr, 1000g2014oct\_eas, 1000g2014oct\_eur, avsnp142, ljb26\_all, gerp++elem,  
448 popfreq\_max, exac03\_all, exac03\_afr, exac03\_amr, exac03\_eas, exac03\_fin, exac03\_nfe,  
449 exac03\_oth, exac03\_sas. We also used the Clinvar database  
450 (<https://ncbi.nlm.nih.gov/clinvar/>, date of access 2016-02-03).

#### 451 $\omega_{GC12}$ calculation

452 Once all the alignments had been collected, we extracted the consensus coding sequences  
453 (CCDS) of all protein-coding genes referenced in Ensembl BioMart Grc37, according to the  
454 HGNC (date of access 05/05/2015) and NCBI Consensus CDS protein set (date of access  
455 2015-08-10). We calculated the number of non-synonymous mutations  $N$ , the number of  
456 synonymous mutations  $S$ , the ratio of the number of nonsynonymous mutations per non-  
457 synonymous site  $dN$ , the number of synonymous mutations per synonymous site  $dS$ , and their  
458 ratio  $dN/dS$  —also called  $\omega$ —between all taxa and the ancestor, using the yn00 algorithm  
459 implemented in PamL software [69]. We avoided infinite and null results, by calculating a  
460 corrected version of  $dN/dS$ . If  $S$  was null, we set its value to one to avoid having zero as the  
461 numerator. The obtained values were validated through the replication of a recent systematic  
462 estimation of  $dN/dS$  between Homo Sapiens and two great apes [65] (Pan troglodytes and  
463 Pongo abelii; Pearson's  $r > 0.8$ ,  $p < 0.0001$ ; see Fig. S2). Finally, we obtained our  $\omega_{GC12}$  value  
464 by correcting for the GC12 content of the genes with a generalized linear model and by  
465 calculating a Z-score for each taxon [27]. GC content has been associated with biases in

466 mutation rates, particularly in primates [70] and humans [71]. We retained only the 11667  
467 genes with 1:1 orthologs in primates (extracted for GRCh37.p13 with Ensemble Biomart,  
468 access online: February 27<sup>th</sup>, 2017).

#### 469 **Gene sets**

470 We used different gene sets, starting at the tissue level and then focusing on the brain and key  
471 pathways. For body tissues, we used Illumina Body Map 2.0 RNA-Seq data, corresponding to  
472 16 human tissue types: adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung,  
473 lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells (for more  
474 information: [https://personal.broadinstitute.org/mgarber/bodymap\\_schroth.pdf](https://personal.broadinstitute.org/mgarber/bodymap_schroth.pdf); data  
475 preprocessed with Cufflinks, accessed May 5, 2015 at <http://cuffdiff.org> ). We also used the  
476 microarray dataset of Su and colleagues [30] (Human U133A/GNF1H Gene Atlas, accessed  
477 May 4, 2015 at <http://biogps.org>). Finally, we also replicated our results with recent RNAseq  
478 data from the GTEx Consortium [31] (<https://www.gtexportal.org/home/>).

479 For the brain, we used the dataset of Su and colleagues and the Human Protein Atlas data  
480 (accessed November 7, 2017 at <https://www.proteinatlas.org>). For analysis of the biological  
481 pathways associated with the brain, we used KEGG (accessed February 25, 2015, at  
482 <http://www.genome.jp/kegg/>), synaptic genes curated by the group of Danielle Posthuma at  
483 Vrije Universiteit (accessed September 1, 2014, at <https://ctg.cncr.nl/software/genesets>), and  
484 mass spectrometry data from Loh and colleagues [72]. Finally, for the diseases associated  
485 with the brain, we combined gene sets generated from Human Phenotype Ontology (accessed  
486 April 5, 2016, at <http://human-phenotype-ontology.github.io>) and OMIM (accessed April 5,  
487 2016, at <https://omim.org>), and curated lists: the 65 risk genes proposed by Sanders and  
488 colleagues [73] (TADA), the candidate genes for autism spectrum disorders from SFARI  
489 (accessed July 17, 2015 at <https://gene.sfari.org>), the Developmental Brain Disorder or DBD  
490 (accessed July 12, 2016 at <https://geisingeradmi.org/care-innovation/studies/dbd-genes/>), and

491 Cancer Census (accessed November 24, 2016 at [cancer.sanger.ac.uk/census](http://cancer.sanger.ac.uk/census)) data. Note that  
492 the combination of HPO & OMIM is the most exhaustive, making it possible to avoid  
493 missing potential candidate genes, but this combination does not identify specific  
494 associations.

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

500 For single-cell transcriptomics datasets, we identified the genes specifically highly expressed  
501 in each cell type, following the same strategy as used for the other RNAseq datasets. The  
502 single-cell data for the developing human cortex were kindly provided by Maximilian  
503 Haeussler (available at <https://cells.ucsc.edu>; access date: October 30, 2018). The single-cell  
504 transcriptional atlas data for the developing murine cerebellum [40] were kindly provided by  
505 Robert A. Carter (access date: January 29, 2019). For each cell type, we combined expression  
506 values across all available replicates, to guarantee a high signal-to-noise ratio. We then  
507 calculated the values for the associated genes in *Homo sapiens* according to the paralogous  
508 correspondence between humans and mice (Ensembl Biomart accessed on February 23,  
509 2019).

## 510 **Gene nomenclature**

511 We extracted all the EntrezId of the protein-coding genes for Grc37 from Ensembl Biomart.  
512 We used the HGNC database to recover their symbols. For the 46 unmapped genes, we  
513 searched the NCBI database manually for the official symbol.

#### 514 **McDonald-Kreitman-test (MK) and neutrality index (NI)**

515 We assessed the possible fixation of variants in the *Homo sapiens* population by first  
516 calculating the relative ratio of non-synonymous to synonymous polymorphism (pN/pS) from  
517 the 1000 Genomes VCF for all SNPs, for SNPs with a minor allele frequency (MAF) <1%  
518 and >5%. SNPs were annotated with ANNOVAR across 1000 Genomes Project (ALL+5  
519 ethnicity groups), ESP6500 (ALL+2 ethnicity groups), ExAC (ALL+7 ethnicity groups), and  
520 CG46 (see [http://annovar.openbioinformatics.org/en/latest/user-guide/filter/#popfreqmax-](http://annovar.openbioinformatics.org/en/latest/user-guide/filter/#popfreqmax-and-popfreqall-annotations)  
521 [and-popfreqall-annotations](http://annovar.openbioinformatics.org/en/latest/user-guide/filter/#popfreqmax-and-popfreqall-annotations) for more details). We then performed the McDonald–Kreitman  
522 test by calculating the neutrality index (NI) as the ratio of raw pN/pS and dN/dS values [74].  
523 We considered the divergent genes to be fixed in the population when  $NI < 1$ .

#### 524 **Protein-protein interaction network**

525 We plotted the protein-protein interaction (PPI) network, by combining eight human  
526 interactomes: the Human Integrated Protein-Protein Interaction Reference (HIPPIE)  
527 (accessed August 10, 2017 at <http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/>), the  
528 Agile Protein Interactomes DataServer (APID) (accessed September 7, 2017 at  
529 <http://cicblade.dep.usal.es:8080/APID/>), CORUM – the comprehensive resource of  
530 mammalian protein complexes (accessed July 13, 2017 at [http://mips.helmholtz-](http://mips.helmholtz-muenchen.de/corum/)  
531 [muenchen.de/corum/](http://mips.helmholtz-muenchen.de/corum/)), and five PPI networks from of the Center for Cancer Systems Biology  
532 (CCSB) (accessed July 12, 2016 at <http://interactome.dfci.harvard.edu/index.php?page=home>  
533 ): four high-quality binary protein-protein interaction (PPI) networks generated by a  
534 systematic primary yeast two-hybrid assay (Y2H): HI-I-05 from Rual and colleagues [75],  
535 Venkatesan-09 from Venkatesan and colleagues [76], Yu-11 from Yu and colleagues [77]  
536 and HI-II-14 from Rolland and colleagues [78], plus one high-quality binary literature dataset  
537 (Lit-BM-13) from Rolland and colleagues [78], comprising all PPIs that are binary and  
538 supported by at least two traceable pieces of evidence (publications and/or methods).

### 539 **NeuroVault analyses**

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

### 547 **Allen Brain data**

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

### 560 **Statistics**

561 **Enrichment analyses:** We first calculated a two-way hierarchical clustering on the  
562 normalized dN/dS values ( $\omega_{GC}$ ) across the whole genome (see Fig. 1b; note: 11,667 genes

563 were included in the analysis to ensure medium-quality coverage for *Homo sapiens*,  
564 Neanderthals, Denisovans, and *Pan troglodytes*; see Supplementary table 2). According to 30  
565 clustering indices [80], the best partitioning in terms of evolutionary pressure was into two  
566 clusters of genes: constrained ( $N=4825$ ; in HS, mean=-0.88 median=-0.80 SD=0.69) and  
567 divergent ( $N=6842$ ; in HS, mean=0.60 median=0.48 sd=0.63. For each cluster, we calculated  
568 the enrichment in biological functions in Cytoscape [81] with the BINGO plugin [82]. We  
569 used all 12,400 genes as the background. We eliminated redundancy, by first filtering out all  
570 the statistically significant Gene Ontology (GO) terms associated with fewer than 10 or more  
571 than 1000 genes, and then combining the remaining genes with the EnrichmentMap plugin  
572 [83]. We used a  $P$ -value cutoff of 0.005, an FDR  $Q$ -value cutoff of 0.05, and a Jaccard  
573 coefficient of 0.5.

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

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

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

584 It corresponds to the difference between the proportion of observations consistent with the  
585 hypothesis ( $f$ ) minus the proportion of observations contradicting the hypothesis ( $u$ ), thus

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

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

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

604 **Data & code availability:** All the data and code supporting the findings of this study are  
605 available from our resource website: <http://neanderthal.pasteur.fr>

606

607 **Acknowledgments**

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

622

623 **Author contributions**

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

627

## 628 **References**

- 629 1. Dunbar RIM, Shultz S. Why are there so many explanations for primate brain evolution?  
630 Phil Trans R Soc B. 2017;372: 20160244. doi:10.1098/rstb.2016.0244
- 631 2. Varki A, Geschwind DH, Eichler EE. Human uniqueness: genome interactions with  
632 environment, behaviour and culture. Nat Rev Genet. 2008;9: nrg2428.  
633 doi:10.1038/nrg2428
- 634 3. Neubauer S, Hublin J-J, Gunz P. The evolution of modern human brain shape. Sci Adv.  
635 2018;4: eao5961. doi:10.1126/sciadv.aao5961
- 636 4. Balsters JH, Cussans E, Diedrichsen J, Phillips KA, Preuss TM, Rilling JK, et al.  
637 Evolution of the cerebellar cortex: the selective expansion of prefrontal-projecting  
638 cerebellar lobules. NeuroImage. 2010;49: 2045–2052.  
639 doi:10.1016/j.neuroimage.2009.10.045
- 640 5. Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, et al.  
641 Inhibition of SRGAP2 Function by Its Human-Specific Paralogs Induces Neoteny  
642 during Spine Maturation. Cell. 2012;149: 923–935. doi:10.1016/j.cell.2012.03.034
- 643 6. Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, et al. Human-specific  
644 gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion.  
645 Sci N Y NY. 2015;347: 1–9. doi:10.1126/science.aaa1975
- 646 7. Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, et al. Human-  
647 Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch  
648 Regulation. Cell. 2018;173: 1370–1384.e16. doi:10.1016/j.cell.2018.03.067
- 649 8. Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, et al. The  
650 evolution and population diversity of human-specific segmental duplications. Nat Ecol  
651 Evol. 2017;1: 69. doi:10.1038/s41559-016-0069
- 652 9. Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Narvaiza I, Sudmant PH, et al.  
653 Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV  
654 susceptibility. Nature. 2016;536: 205–209. doi:10.1038/nature19075
- 655 10. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, et al.  
656 Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental  
657 Duplication. Cell. 2012;149: 912–922. doi:10.1016/j.cell.2012.03.033
- 658 11. Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, et al. Molecular  
659 evolution of FOXP2, a gene involved in speech and language. Nature. 2002;418: 869–  
660 872. doi:10.1038/nature01025
- 661 12. Montgomery SH, Mundy NI, Barton RA. ASPM and mammalian brain evolution: a case  
662 study in the difficulty in making macroevolutionary inferences about gene-phenotype  
663 associations. Proc R Soc B Biol Sci. 2014;281: 20131743. doi:10.1098/rspb.2013.1743
- 664 13. Enard W, Gehre S, Hammerschmidt K, Hölter SM, Blass T, Somel M, et al. A humanized  
665 version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell. 2009;137: 961–  
666 971. doi:10.1016/j.cell.2009.03.041

- 667 14. Atkinson EG, Audesse AJ, Palacios JA, Bobo DM, Webb AE, Ramachandran S, et al. No  
668 Evidence for Recent Selection at FOXP2 among Diverse Human Populations. *Cell*.  
669 2018;0. doi:10.1016/j.cell.2018.06.048
- 670 15. Miyata T, Kuma K, Iwabe N, Nikoh N. A possible link between molecular evolution and  
671 tissue evolution demonstrated by tissue specific genes. *Idengaku Zasshi*. 1994;69: 473–  
672 480.
- 673 16. Wang H-Y, Chien H-C, Osada N, Hashimoto K, Sugano S, Gojobori T, et al. Rate of  
674 Evolution in Brain-Expressed Genes in Humans and Other Primates. *PLoS Biol*. 2006;5:  
675 e13. doi:10.1371/journal.pbio.0050013
- 676 17. Tuller T, Kupiec M, Ruppin E. Evolutionary rate and gene expression across different  
677 brain regions. *Genome Biol*. 2008;9: R142. doi:10.1186/gb-2008-9-9-r142
- 678 18. King M-C, Wilson AC. Evolution at two levels in humans and chimpanzees. *Science*.  
679 1975;188: 107–116. doi:10.1126/science.1090005
- 680 19. Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, et al. An  
681 RNA gene expressed during cortical development evolved rapidly in humans. *Nature*.  
682 2006;443: 167–172. doi:10.1038/nature05113
- 683 20. Changeux J-P. Climbing Brain Levels of Organisation from Genes to Consciousness.  
684 *Trends Cogn Sci*. 2017;21: 168–181. doi:10.1016/j.tics.2017.01.004
- 685 21. Calarco JA, Xing Y, Cáceres M, Calarco JP, Xiao X, Pan Q, et al. Global analysis of  
686 alternative splicing differences between humans and chimpanzees. *Genes Dev*. 2007;21:  
687 2963–2975. doi:10.1101/gad.1606907
- 688 22. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the  
689 clinical impact of human mutation with deep neural networks. *Nat Genet*. 2018 [cited 24  
690 Jul 2018]. doi:10.1038/s41588-018-0167-z
- 691 23. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions  
692 in the human genome. *Nat Genet*. 2019;51: 88. doi:10.1038/s41588-018-0294-6
- 693 24. Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, et al.  
694 Accelerated Evolution of Nervous System Genes in the Origin of Homo sapiens. *Cell*.  
695 2004;119: 1027–1040. doi:10.1016/j.cell.2004.11.040
- 696 25. Huang Y, Xie C, Ye AY, Li C-Y, Gao G, Wei L. Recent Adaptive Events in Human  
697 Brain Revealed by Meta-Analysis of Positively Selected Genes. Robinson-Rechavi M,  
698 editor. *PLoS ONE*. 2013;8: e61280. doi:10.1371/journal.pone.0061280
- 699 26. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: Genome-wide  
700 mammalian consistency-based multiple alignment with paralogs. *Genome Res*. 2008;18:  
701 1814–1828. doi:10.1101/gr.076554.108
- 702 27. Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, et al. Genomic signatures of  
703 evolutionary transitions from solitary to group living. *Science*. 2015;348: 1139–1143.  
704 doi:10.1126/science.aaa4788
- 705 28. Yang Z, Bielawski J. Statistical methods for detecting molecular adaptation. *Trends Ecol*  
706 *Evol*. 2000;15: 496–503. doi:10.1016/s0169-5347(00)01994-7

- 707 29. O'Toole AN, Hurst LD, McLysaght A. Faster Evolving Primate Genes Are More Likely  
708 to Duplicate. *Mol Biol Evol.* 2018;35: 107–118. doi:10.1093/molbev/msx270
- 709 30. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the  
710 mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A.*  
711 2004;101: 6062–6067. doi:10.1073/pnas.0400782101
- 712 31. Consortium TGte. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue  
713 gene regulation in humans. *Science.* 2015;348: 648–660. doi:10.1126/science.1262110
- 714 32. Schoenemann PT, Sheehan MJ, Glotzer LD. Prefrontal white matter volume is  
715 disproportionately larger in humans than in other primates. *Nat Neurosci.* 2005;8: 242–  
716 252. doi:10.1038/nn1394
- 717 33. Frith C, Dolan R. The role of the prefrontal cortex in higher cognitive functions. *Cogn*  
718 *Brain Res.* 1996;5: 175–181. doi:10.1016/S0926-6410(96)00054-7
- 719 34. Tao X, West AE, Chen WG, Corfas G, Greenberg ME. A calcium-responsive  
720 transcription factor, CaRF, that regulates neuronal activity-dependent expression of  
721 BDNF. *Neuron.* 2002;33: 383–395.
- 722 35. Whitney O, Pfenning AR, Howard JT, Blatti CA, Liu F, Ward JM, et al. Core and region-  
723 enriched networks of behaviorally regulated genes and the singing genome. *Science.*  
724 2014;346: 1256780. doi:10.1126/science.1256780
- 725 36. Ruano D, Abecasis GR, Glaser B, Lips ES, Cornelisse LN, de Jong APH, et al.  
726 Functional Gene Group Analysis Reveals a Role of Synaptic Heterotrimeric G Proteins  
727 in Cognitive Ability. *Am J Hum Genet.* 2010;86: 113–125.  
728 doi:10.1016/j.ajhg.2009.12.006
- 729 37. Lips ES, Cornelisse LN, Toonen RF, Min JL, Hultman CM, Holmans PA, et al.  
730 Functional gene group analysis identifies synaptic gene groups as risk factor for  
731 schizophrenia. *Mol Psychiatry.* 2012;17: 996–1006.
- 732 38. Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, et al. Transcriptional  
733 landscape of the prenatal human brain. *Nature.* 2014;508: 199–206.  
734 doi:10.1038/nature13185
- 735 39. Brazel CY, Romanko MJ, Rothstein RP, Levison SW. Roles of the mammalian  
736 subventricular zone in brain development. *Prog Neurobiol.* 2003;69: 49–69.  
737 doi:10.1016/S0301-0082(03)00002-9
- 738 40. Carter RA, Bihannic L, Rosencrance C, Hadley JL, Tong Y, Phoenix TN, et al. A Single-  
739 Cell Transcriptional Atlas of the Developing Murine Cerebellum. *Curr Biol.* 2018;28:  
740 2910-2920.e2. doi:10.1016/j.cub.2018.07.062
- 741 41. Nowakowski TJ, Bhaduri A, Pollen AA, Alvarado B, Mostajo-Radji MA, Lullo ED, et al.  
742 Spatiotemporal gene expression trajectories reveal developmental hierarchies of the  
743 human cortex. *Science.* 2017;358: 1318–1323. doi:10.1126/science.aap8809
- 744 42. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide  
745 atlas of gene expression in the adult mouse brain. *Nature.* 2007;445: 168–176.  
746 doi:10.1038/nature05453

- 747 43. Coene KLM, Roepman R, Doherty D, Afroze B, Kroes HY, Letteboer SJF, et al. OFD1 is  
748 mutated in X-linked Joubert syndrome and interacts with LCA5-encoded lebercilin. *Am*  
749 *J Hum Genet.* 2009;85: 465–481. doi:10.1016/j.ajhg.2009.09.002
- 750 44. Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, et al.  
751 NeuroVault.org: a web-based repository for collecting and sharing unthresholded  
752 statistical maps of the human brain. *Front Neuroinformatics.* 2015;9: 8.  
753 doi:10.3389/fninf.2015.00008
- 754 45. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An  
755 anatomically comprehensive atlas of the adult human brain transcriptome. *Nature.*  
756 2012;489: 391–399. doi:10.1038/nature11405
- 757 46. Gorgolewski KJ, Fox AS, Chang L, Schäfer A, Arélin K, Burmann I, et al. Tight fitting  
758 genes: finding relations between statistical maps and gene expression patterns.  
759 *F1000Research.* 2014;5. doi:10.7490/f1000research.1097120.1
- 760 47. Karayannis T, Au E, Patel JC, Kruglikov I, Markx S, Delorme R, et al. *Cntnap4*  
761 differentially contributes to GABAergic and dopaminergic synaptic transmission.  
762 *Nature.* 2014;511: 236–240. doi:10.1038/nature13248
- 763 48. Jamain S, Quach H, Betancur C, Råstam M, Colineaux C, Gillberg IC, et al. Mutations of  
764 the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with  
765 autism. *Nat Genet.* 2003;34: 27–29. doi:10.1038/ng1136
- 766 49. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, et al.  
767 SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders  
768 (ASDs). *Mol Autism.* 2013;4: 36. doi:10.1186/2040-2392-4-36
- 769 50. Mejias R, Adamczyk A, Anggono V, Niranjana T, Thomas GM, Sharma K, et al. Gain-of-  
770 function glutamate receptor interacting protein 1 variants alter GluA2 recycling and  
771 surface distribution in patients with autism. *Proc Natl Acad Sci U S A.* 2011;108: 4920–  
772 4925. doi:10.1073/pnas.1102233108
- 773 51. Paracchini S, Thomas A, Castro S, Lai C, Paramasivam M, Wang Y, et al. The  
774 chromosome 6p22 haplotype associated with dyslexia reduces the expression of  
775 KIAA0319, a novel gene involved in neuronal migration. *Hum Mol Genet.* 2006;15:  
776 1659–1666. doi:10.1093/hmg/ddl089
- 777 52. Franquinho F, Nogueira-Rodrigues J, Duarte JM, Esteves SS, Carter-Su C, Monaco AP,  
778 et al. The Dyslexia-susceptibility Protein KIAA0319 Inhibits Axon Growth Through  
779 Smad2 Signaling. *Cereb Cortex N Y N 1991.* 2017;27: 1732–1747.  
780 doi:10.1093/cercor/bhx023
- 781 53. Dennis MY, Eichler EE. Human adaptation and evolution by segmental duplication. *Curr*  
782 *Opin Genet Dev.* 2016;41: 44–52. doi:10.1016/j.gde.2016.08.001
- 783 54. Hayward P. Joubert syndrome may provide clues about human evolution. *Lancet Neurol.*  
784 2004;3: 574. doi:10.1016/S1474-4422(04)00870-1
- 785 55. Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, Hudson RR, et al.  
786 Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*.  
787 *Science.* 2005;309: 1720–1722. doi:10.1126/science.1116815

- 788 56. Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, Vaez-Azizi LM, et  
789 al. Microcephalin, a gene regulating brain size, continues to evolve adaptively in  
790 humans. *Science*. 2005;309: 1717–1720. doi:10.1126/science.1113722
- 791 57. Yu F, Hill RS, Schaffner SF, Sabeti PC, Wang ET, Mignault AA, et al. Comment on  
792 “Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*.”  
793 *Science*. 2007;316: 370. doi:10.1126/science.1137568
- 794 58. Unger S, Górna MW, Le Béchech A, Do Vale-Pereira S, Bedeschi MF, Geiberger S, et al.  
795 FAM111A mutations result in hypoparathyroidism and impaired skeletal development.  
796 *Am J Hum Genet*. 2013;92: 990–995. doi:10.1016/j.ajhg.2013.04.020
- 797 59. Hoischen A, van Bon BWM, Rodríguez-Santiago B, Gilissen C, Vissers LELM, de Vries  
798 P, et al. De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat*  
799 *Genet*. 2011;43: 729–731. doi:10.1038/ng.868
- 800 60. Ferland RJ, Eyaid W, Collura RV, Tully LD, Hill RS, Al-Nouri D, et al. Abnormal  
801 cerebellar development and axonal decussation due to mutations in AHI1 in Joubert  
802 syndrome. *Nat Genet*. 2004;36: 1008–1013. doi:10.1038/ng1419
- 803 61. Gould DB, Walter MA. Mutational analysis of BARHL1 and BARX1 in three new  
804 patients with Joubert syndrome. *Am J Med Genet A*. 2004;131: 205–208.  
805 doi:10.1002/ajmg.a.30227
- 806 62. Koziol LF, Budding DE, Chidekel D. From movement to thought: executive function,  
807 embodied cognition, and the cerebellum. *Cerebellum Lond Engl*. 2012;11: 505–525.  
808 doi:10.1007/s12311-011-0321-y
- 809 63. Rao SM, Mayer AR, Harrington DL. The evolution of brain activation during temporal  
810 processing. *Nat Neurosci*. 2001;4: 317–323. doi:10.1038/85191
- 811 64. Eyre-Walker A, Keightley PD. Estimating the Rate of Adaptive Molecular Evolution in  
812 the Presence of Slightly Deleterious Mutations and Population Size Change. *Mol Biol*  
813 *Evol*. 2009;26: 2097–2108. doi:10.1093/molbev/msp119
- 814 65. Biswas K, Chakraborty S, Podder S, Ghosh TC. Insights into the dN/dS ratio  
815 heterogeneity between brain specific genes and widely expressed genes in species of  
816 different complexity. *Genomics*. 2016;108: 11–17. doi:10.1016/j.ygeno.2016.04.004
- 817 66. Zylka MJ, Simon JM, Philpot BD. Gene length matters in neurons. *Neuron*. 2015;86:  
818 353–355. doi:10.1016/j.neuron.2015.03.059
- 819 67. Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, et al.  
820 Patterns of coding variation in the complete exomes of three Neandertals. *Proc Natl*  
821 *Acad Sci U S A*. 2014;111: 6666–6671. doi:10.1073/pnas.1405138111
- 822 68. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants  
823 from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38: e164–e164.  
824 doi:10.1093/nar/gkq603
- 825 69. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*.  
826 2007;24: 1586–1591. doi:10.1093/molbev/msm088

- 827 70. Galtier N, Duret L, Glémin S, Ranwez V. GC-biased gene conversion promotes the  
828 fixation of deleterious amino acid changes in primates. *Trends Genet.* 2009;25: 1–5.
- 829 71. Kostka D, Hubisz MJ, Siepel A, Pollard KS. The Role of GC-Biased Gene Conversion in  
830 Shaping the Fastest Evolving Regions of the Human Genome. *Mol Biol Evol.* 2012;29:  
831 1047–1057. doi:10.1093/molbev/msr279
- 832 72. Loh KH. Proteomics: The proteomes of excitatory and inhibitory synaptic clefts. *Nat*  
833 *Methods.* 2016;13: 903–903. doi:10.1038/nmeth.4050
- 834 73. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al.  
835 Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71  
836 Risk Loci. *Neuron.* 2015;87: 1215–1233. doi:10.1016/j.neuron.2015.09.016
- 837 74. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*.  
838 *Nature.* 1991;351: 351652a0. doi:10.1038/351652a0
- 839 75. Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a  
840 proteome-scale map of the human protein-protein interaction network. *Nature.*  
841 2005;437: 1173–1178. doi:10.1038/nature04209
- 842 76. Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, et al.  
843 An empirical framework for binary interactome mapping. *Nat Methods.* 2009;6: 83–90.  
844 doi:10.1038/nmeth.1280
- 845 77. Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, Fan C, et al. Next-generation sequencing  
846 to generate interactome datasets. *Nat Methods.* 2011;8: 478–480.  
847 doi:10.1038/nmeth.1597
- 848 78. Rolland T, Taşan M, Charlotteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-  
849 scale map of the human interactome network. *Cell.* 2014;159: 1212–1226.  
850 doi:10.1016/j.cell.2014.10.050
- 851 79. French L. A FreeSurfer view of the cortical transcriptome generated from the Allen  
852 Human Brain Atlas. 2017. doi:10.6084/m9.figshare.1439749.v11
- 853 80. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining  
854 the relevant number of clusters in a data set. *J Stat Softw.* 2014;61: 1–36.
- 855 81. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A  
856 Software Environment for Integrated Models of Biomolecular Interaction Networks.  
857 *Genome Res.* 2003;13: 2498–2504. doi:10.1101/gr.1239303
- 858 82. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess  
859 overrepresentation of gene ontology categories in biological networks. *Bioinforma Oxf*  
860 *Engl.* 2005;21: 3448–3449. doi:10.1093/bioinformatics/bti551
- 861 83. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment Map: A Network-  
862 Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLOS ONE.*  
863 2010;5: e13984. doi:10.1371/journal.pone.0013984
- 864 84. Xu X, Wells AB, O'Brien DR, Nehorai A, Dougherty JD. Cell Type-Specific Expression  
865 Analysis to Identify Putative Cellular Mechanisms for Neurogenetic Disorders. *J*  
866 *Neurosci.* 2014;34: 1420–1431. doi:10.1523/JNEUROSCI.4488-13.2014

- 867 85. Kerby DS. The Simple Difference Formula: An Approach to Teaching Nonparametric  
868 Correlation. *Compr Psychol.* 2014;3: 11.IT.3.1. doi:10.2466/11.IT.3.1
- 869 86. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-  
870 specificity metrics. *Brief Bioinform.* 2016; bbw008. doi:10.1093/bib/bbw008