Evolution of human brain-size associated NOTCH2NL genes proceeds towards reduced protein levels

Lodewijk, G.A.; Fernandes, D.P.; Vretzakis, I.; Savage, J.E.; Jacobs, F.M.J.

DOI
10.1093/molbev/msaa104

Publication date
2020

Document Version
Final published version

Published in
Molecular Biology and evolution

License
CC BY

Citation for published version (APA):
Evolution of Human Brain Size-Associated NOTCH2NL Genes Proceeds toward Reduced Protein Levels

Gerrald A. Lodewijk,1 Diana P. Fernandes,1 Iraklis Vretzakis,1 Jeanne E. Savage,2,3 and Frank M.J. Jacobs*1,3

1Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
2Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, VU University, Amsterdam, The Netherlands
3Amsterdam Neuroscience, Complex Trait Genetics

*Corresponding author: E-mail: f.m.j.jacobs@uva.nl.

Abstract

Ever since the availability of genomes from Neanderthals, Denisovans, and ancient humans, the field of evolutionary genomics has been searching for protein-coding variants that may hold clues to how our species evolved over the last ~600,000 years. In this study, we identify such variants in the human-specific NOTCH2NL gene family, which were recently identified as possible contributors to the evolutionary expansion of the human brain. We find evidence for the existence of unique protein-coding NOTCH2NL variants in Neanderthals and Denisovans which could affect their ability to activate Notch signaling. Furthermore, in the Neanderthal and Denisovan genomes, we find unusual NOTCH2NL configurations, not found in any of the modern human genomes analyzed. Finally, genetic analysis of archaic and modern humans reveals ongoing adaptive evolution of modern human NOTCH2NL genes, identifying three structural variants acting complementary to drive our genome to produce a lower dosage of NOTCH2NL protein. Because copy-number variations of the 1q21.1 locus, encompassing NOTCH2NL genes, are associated with severe neurological disorders, this seemingly contradictory drive toward low levels of NOTCH2NL protein indicates that the optimal dosage of NOTCH2NL may have not yet been settled in the human population.

Key words: archaic genomes, brain size, human evolutionary genomics, human-specific genes, segmental duplications, Neanderthal, gene conversion.

Introduction

The human brain tripled in size after we split from the common ancestor with our closest living relative species, the chimpanzees (Marino 1998; Herculano-Houzel 2009; Hofman 2014). The emergence of human-specific NOTCH2NL genes (Fiddes et al. 2018; Florio et al. 2018; Suzuki et al. 2018) coincided with this evolutionary expansion (Holloway et al. 2004; Pollen et al. 2015; Ju et al. 2016; Liu et al. 2017; Johnson et al. 2018; Kalebic et al. 2018) and their association to human brain development put NOTCH2NL genes forward as possible contributors to human’s increased brain size. By enhancing Notch signaling, NOTCH2NL genes prolong proliferation of neuronal progenitor cells and expand cortical neurogenesis (Fiddes et al. 2018; Florio et al. 2018; Suzuki et al. 2018). NOTCH2NL genes are human specific and they emerged after a series of segmental duplications and gene conversion events involving the important neurodevelopmental gene NOTCH2. Four NOTCH2NL paralogs are present in modern humans: NOTCH2NLA, NOTCH2NLB, and NOTCH2NL in the 1q21.1 locus (fig. 1A) and the pseudogene NOTCH2NL next to the parental NOTCH2 gene in the 1p12 locus. NOTCH2NL represents the largest duplciton in the cluster, suggesting this was the first NOTCH2NL gene present in the genome (fig. 1B). Whereas copy-number variation is observed for NOTCH2NLC and NOTCH2NLR in the healthy human population; the copy number of NOTCH2NLA and NOTCH2NLB loci is highly stable in modern humans. In fact, 1q21.1 copy-number variations, mediated by breakpoints within the NOTCH2NLA and NOTCH2NLB genes, are associated with various neurological disorders (Brunetti-Pierri et al. 2008; Mefford et al. 2008; Bernier et al. 2016; Fiddes et al. 2018). These observations suggest that the total number of functional NOTCH2NLA and NOTCH2NLB alleles may be important for normal neuronal development. Given the highly variable genomic organization of the 1q21.1 locus, important questions remain about the level of variation in NOTCH2NL genes in the human population. In addition, it remains elusive whether the number and composition of NOTCH2NL genes has changed during recent human evolution. Here, we analyzed the segregation of coding variants in NOTCH2NL genes throughout human evolution and compared the composition of each NOTCH2NL locus between modern humans and archaic genomes. Our analysis revealed lineage-specific coding variants in each of the genomes of Neanderthals, Denisovans, and modern humans. Intriguingly, we find evidence for ongoing adaptive evolution of multiple structural variants in
**Fig. 1.** NOTCH2NL copy-number analysis in modern human and archaic DNA samples. (A) Overview of NOTCH2 and NOTCH2NL loci in the human genome (hg38). Zoom-ins show sequence read depth at the different loci of data mapped on hg38 or masked hg38 reference genome. (B) Tracks showing NOTCH2NL duplicons from the segmental UCSC browser duplication track in the NOTCH2 locus. (C) Example showing NOTCH2- and NOTCH2NL-derived sequencing reads piled up on the NOTCH2 locus on the masked hg38 genome. (D) Quantification of NOTCH2 + NOTCH2NL alleles per individual using relative coverage of multicopy/single-copy regions. Modern human, \( n = 279 \). Ancient human: high (\( n = 27 \))/low (\( n = 53 \)) coverage; Neanderthal: high (\( n = 3 \))/low (\( n = 9 \)) coverage; Denisova: high (\( n = 1 \))/low (\( n = 1 \)) coverage. (E, F) NOTCH2NL allele counts estimated from the average density of paralog-specific SUNs in modern human outliers (E) and Neanderthals (F) showing evidence for the presence of 11 alleles in total (two alleles NOTCH2 + nine alleles NOTCH2NL). (G) Comparison of allele count grouped by NOTCH2NL (+ NOTCH2NLB) (Kruskal–Wallis \( P = 1.8e-8 \)), and NOTCH2NL + NOTCH2NL + NOTCH2 (Kruskal–Wallis \( P = 0.0055 \)). Kruskal–Wallis test was followed up by Dunn’s test, significant comparisons are indicated in the plots. Modern human, \( N = 279 \); ancient human, \( N = 80 \); Neanderthal, \( N = 12 \); and Denisova, \( N = 2 \).
modern human \textit{NOTCH2NLC} genes, acting in synergy and complementary to drive our genome to produce a lower dosage of \textit{NOTCH2NLC} protein. The evolutionary forces mediated by gene conversion \cite{Chen2007}, which we find is still ongoing between \textit{NOTCH2NLC} loci at a high frequency in modern humans, exemplify how recently duplicated regions in our genome can undergo rapid structural evolution to reach an optimal configuration and functionality. For humans, this may have had important consequences for how a key developmental process such as Notch signaling has evolved in the period after the emergence of \textit{NOTCH2NLC} genes and the changes they effectuated on human brain development.

**Additional Copies of \textit{NOTCH2NLA} or \textit{NOTCH2NLB} in Neanderthals**

To assess the structural evolution of each of the \textit{NOTCH2NL} loci throughout human evolution, we first assessed the structural variability of \textit{NOTCH2NLC} loci in the modern human population. Previous estimations of total \textit{NOTCH2NLC} copy number in individuals could not efficiently distinguish between paralogous \textit{NOTCH2NLC} loci subject to recent ectopic gene conversion, as observed between \textit{NOTCH2NLC-NOTCH2NLR} and between \textit{NOTCH2NLC-NOTCH2NLRB} \cite{Dougherty2017,Fiddes2018}. Here, we used an alternative strategy that takes into account gene conversion between paralogous \textit{NOTCH2NLC} loci. For each genome, we assessed total number of \textit{NOTCH2NLC} alleles based on sequence read coverage and matched this with information about the presence or absence of \textit{NOTCH2NL}-paralog identifying single-unique nucleotides (SUNs) \cite{Sudmant2010}. This provides an accurate assessment of the absolute number of \textit{NOTCH2NL} alleles in each individual genome and a detailed overview of the structural variability of \textit{NOTCH2NLC} genes as a consequence of gene conversion \cite{Supplementary Material online}. We verified the accuracy of our methodology by showing concordance with previous \textit{NOTCH2NLC} assembly–based estimations \cite{Supplementary Material online}. To assess the total number of \textit{NOTCH2NL} alleles across the human population, the genomes of 279 individuals from the Simons diversity data set \cite{Mallick2016} were mapped onto a modified hg38 genome in which the \textit{NOTCH2NLC} loci are masked (fig. 1A). On this modified hg38 genome, all \textit{NOTCH2NLC}-derived reads map onto the 5' side of the \textit{NOTCH2} locus, the part of \textit{NOTCH2} that was originally duplicated forming the \textit{NOTCH2NLC} genes (fig. 1B and C). The coverage analysis reveals that the majority of the human population has ten alleles, encompassing two alleles from \textit{NOTCH2} and two alleles from each of the four \textit{NOTCH2NLC} loci (fig. 1D). Using the combined sequence coverage and SUN analysis, we determined that each individual contained 4 alleles combined of the highly similar \textit{NOTCH2NLA} and \textit{NOTCH2NLB} genes. The individuals that have nine, eight, or seven alleles were all confirmed as hetero- or homozygotic for \textit{NOTCH2NLC} and \textit{NOTCH2NLR} \cite{Supplementary Material online}. Four human individuals have one extra allele of \textit{NOTCH2NLC} or \textit{NOTCH2NLR}, indicating that \textit{NOTCH2NLC} duplications happen in the healthy human population (fig. 1E).

Next, we analyzed genomes of ancient humans (0.1k–45k years old) \cite{Keller2012,Fu2014,Fu2016,Gamba2014,Lazaridis2014,Olalde2014,Raghavan2014,Rasmussen2014,Rasmussen2015,Seguin-Orlando2014,Skoglund2014,Skoglund2017,Guenter2015,Guenter2018,Jones2015,Jones2017,Cassidy2016,Martinsoni2016,Schifflers2016,Saag2017,Bhattacharya2018,deFuente2018,Krzewinska2018,Valdivieso2018,Wright2018,Sanchez-Quinto2019}, Neanderthals (38k–100k years old) \cite{Green2010,Prüfer2014,Prüfer2017,Hajdinjak2018,Slon2018,Mafessoni2020}, and Denisovans (64k–100k years old) \cite{Meyer2012,Meyer2013,Meyer2014,Meyer2015,Meyer2016,Meyer2017,Meyer2018,Meyer2019,Meyer2020,Prüfer2014,Prüfer2017,Hajdinjak2018}. Although most of the ancient human genomes display \textit{NOTCH2NL} allele numbers that fall within the range of modern humans, several of the 12 available Neanderthal genomes show increased coverage, which indicates they contained an extra \textit{NOTCH2NL} duplication (fig. 1D). Whereas the combined copy number of \textit{NOTCH2NLA} and \textit{NOTCH2NLB} is highly stable in healthy modern humans, SUN-based copy-number estimation suggests that Neanderthals carried an extra duplication of the \textit{NOTCH2NL} gene \cite{fig. 1F and G and supplementary fig. S1C, Supplementary Material online}. Whether this is a gain in Neanderthal or a loss in modern humans remains elusive. In addition, all Neanderthal genomes showed evidence of extensive gene conversion between \textit{NOTCH2NL} and \textit{NOTCH2NLN} \cite{supplementary fig. S1C, Supplementary Material online}, a phenomenon observed only occasionally in modern humans \cite{supplementary fig. S1D and E, Supplementary Material online}.

**Neanderthals and Denisovans Carried Specific \textit{NOTCH2NL} Variants**

We next investigated whether the archaic genomes contained any coding sequence variants that may have encoded unique \textit{NOTCH2NL} protein variants. Despite an overall high similarity (99.9%) between human and Neanderthal/Denisovan \textit{NOTCH2NL} exons, we found evidence for two Neanderthal-specific coding variants and one Denisovan-specific coding variant (fig. 2A). In the Altai Neanderthal genome, an ATG > ATA (M40I) missense variant (\textit{NOTCH2NL}{\textsubscript{Nea-M40I}}) is detected in 17/242 (~8%) of the sequencing reads corresponding to one allele out of the nine \textit{NOTCH2NL} alleles found in Altai Neanderthals. The second Neanderthal-specific variant is a N232S missense variant (\textit{NOTCH2NL}{\textsubscript{Nea-N232S}}) detected in 28/177 (~18%) of sequencing reads, corresponding to two alleles. This variant is also present in the genomes of the Vindija and Chagyrskaya Neanderthals and most of the low-coverage Neanderthal genomes, indicating the \textit{NOTCH2NL}{\textsubscript{Nea-N232S}} variant was a common variant in the Neanderthal lineage. In the Denisovan3 genome, a Denisovan-specific E258A missense variant (\textit{NOTCH2NL}{\textsubscript{Den-E258A}}) is found in 38/203 (~19%) of the sequencing reads, also corresponding to two alleles. Importantly, none of these variants are found in the 279 modern human genomes of the Simons diversity data set. Interestingly, the \textit{NOTCH2NL}{\textsubscript{Nea-N232S}} was found as a rare
variant in modern humans (rs375605753) with an allele frequency of 0.0002 in UK Biobank exome sequencing data ($N = 49,593$), suggesting this was one of the Neanderthal-derived genetic variants that was contributed to the human genome after interbreeding with Neanderthals (Dannemann and Racimo 2018). It should be noted that the highly fragmented assemblies of archaic genomes prevents us from making solid claims about which NOTCH2NL paralog each of these archaic variants reside in. Taking this into account, we assessed the potential functional implications of the Neanderthal and Denisovan variants by reconstructing the archaic NOTCH2NL variants in NOTCH2NLA and NOTCH2NLB for functional testing in a previously established Notch signaling reporter assay (Groot et al. 2014; Habets et al. 2015; Fiddes et al. 2018)(supplementary fig. S2A, Supplementary Material online). Surprisingly, the introduction of the Nea-N232S and Den-E258A into human NOTCH2NLA showed a modest but significant decrease in potency to enhance Notch signaling (fig. 2B). To find an explanation for the functional divergence of the archaic NOTCH2NL variants, we investigated the potential structural implications in more detail (supplementary fig. S2B, Supplementary Material online). The Neanderthal M40I variant is located in EGF-L domain 1 and disrupts the predicted start codon of NOTCH2NLA. The Neanderthal N232S variant is located in EGF-L domain 6, which is fully conserved between NOTCH paralogs and between species (supplementary fig. S2C, Supplementary Material online). The N232 residue is part of an important motif for glycosylation, a posttranslational modification which mediates EGF-L folding (Takeuchi et al. 2017) and NOTCH–ligand interactions (Jafar-Nejad et al. 2010) (supplementary fig. S2D, Supplementary Material online). As such, the N232S variant is predicted to alter NOTCH2NL protein interaction dynamics or protein stability (supplementary fig. S2E, Supplementary Material online). Indeed, the corresponding rare single-nucleotide polymorphism (SNP) in modern humans (rs375605753) is predicted to be deleterious (Pejaver et al. 2017). The Denisova E258A variant is located in the C-terminal domain of NOTCH2NL, an intrinsically disordered region known to play a role in protein stability (Duan et al. 2003; Fiddes et al. 2018). Analysis using IUPred2A (Mézaszós et al. 2018) suggests that this substitution alters the state of the NOTCH2NL C-terminal domain, potentially affecting protein stability (supplementary fig. S2F, Supplementary Material online). In support of this, a modest increase in protein level was observed for the Den-E258A and Nea-N232S variants introduced into human NOTCH2NLB (fig. 2C and D). This suggests that these archaic variants positively affected protein translation or stability. Altogether, Denisovans and Neanderthals carried
alleles in their genome which are likely to have affected the function of their NOTCH2NL genes.

Variants in Exon1 of NOTCH2NL Genes Determine NOTCH2NL Protein Levels

Unexpectedly, we noticed that the NOTCH2NLANeaM40I variant, predicted to lack the first 83 amino acids, was not different in size from NOTCH2NLB. Likewise, no decrease in protein size was observed for NOTCH2NLN, predicted to lack the first 39 amino acids. Analysis of multiple 5' truncated NOTCH2NL cDNAs reveals that instead of the conventional ATG initiation sites on positions M40 and M84, multiple unconventional CTG start sites in the 5' side of NOTCH2NL drive translation of NOTCH2NLN and NOTCH2NLANeaM40I proteins (Kearse and Wilusz 2017) (fig. 3A and supplementary fig. S3A–G, Supplementary Material online). As a result and as opposed to what is predicted by gene models, human NOTCH2NL and Neanderthal NOTCH2NLNeaM40I encode almost full-length NOTCH2NL proteins with a functionally intact N-terminal signal peptide. Importantly, our analysis also reveals that the usage of unconventional translation initiation sites has major consequences for the level of NOTCH2NL protein produced by each of the NOTCH2NL genes. NOTCH2NLA, which lacks the first start codon produces a 5-fold lower level of NOTCH2NL protein compared with NOTCH2NLB (fig. 3A–C). NOTCH2NLN is also forced to use downstream CTG sites for translation initiation and gives rise to normal-sized NOTCH2NL protein (fig. 3B). However, due to the combination of the NOTCH2NLN-characteristic 2-bp deletion and upstream open-reading frames (ORFs), the expression level of NOTCH2NLN is extremely low, at only 1% compared with NOTCH2NLB (fig. 3C). These new insights reveal that the level of NOTCH2NL protein generated by each of the genes is predominantly dependent on the presence or absence of three specific coding variants in Exon1 (fig. 3D). Compared with the NOTCH2NLB configuration of Exon1 (Exon1B(High)-variant) which produces high levels of NOTCH2NL protein, the M1I substitution in NOTCH2NLN (Exon1A(Low)-variant) produces 5-fold less NOTCH2NL protein.

![Fig. 3](https://academic.oup.com/mbe/article/37/9/2531/5824797) Downloaded from https://academic.oup.com/mbe/article/37/9/2531/5824797 by Universiteit van Amsterdam user on 07 September 2020
deletion in Exon1, (Exon1^C-(X-low)-variant) results in extremely low levels of NOTCH2NL protein. Importantly, ectopic gene conversion between NOTCH2NL loci can result in transfer of Exon1-variants from one NOTCH2NL gene to another. As a consequence, the total dosage of NOTCH2NL protein in each individual may not be defined by the copy number of each of the NOTCH2NL genes, but by the level of Exon1-variant carry-over via gene conversion between NOTCH2NL genes.

### Unusual Configuration of NOTCH2NL Genes in the Denisova3 Genome

To assess the extent to which gene conversion influences the distribution of Exon1-variants between NOTCH2NL genes, we investigated the distribution of SUNs across the NOTCH2NL loci. First, we analyzed modern human NOTCH2NLC for evidence of gene conversion. Analysis of the Exon1 configuration of NOTCH2NL genes reveals that most modern humans contain two NOTCH2NL-derived Exon1^C-(X-low)-variants (fig. 4A), present in both alleles of NOTCH2NL. Furthermore, an equal distribution was found for NOTCH2NL SUNs across the NOTCH2NL locus in most modern human individuals (fig. 4B), suggesting that gene conversion between NOTCH2NLC and other NOTCH2NL loci does not commonly happen. A similar pattern was found in Neanderthals and ancient humans (fig. 4A and supplementary fig. S4A and B, Supplementary Material online). This indicates that the majority of Neanderthal, archaic human, and modern human genomes have two NOTCH2NL alleles carrying the Exon1^C-(X-low)-variant. The Denisova3 genome however, shows a strikingly different pattern.

The presence of NOTCH2NL-paralog-specific SUNs across the NOTCH2NL loci shows that NOTCH2NL, NOTCH2NLB, and NOTCH2NLc genes are present in the Denisova3 genome (fig. 4C). Based on the complete absence of NOTCH2NLR SUNs and a total coverage representative of only six NOTCH2NL alleles (fig. 1D), it is likely the Denisova3 genome had a homozygous deletion of NOTCH2NLR. Remarkably, despite good coverage of the Exon1 region in the Denisova3 genome (36X), all NOTCH2NL-derived reads from Exon1 carry the NOTCH2NL-derived Exon1^C-(X-low)-variant (fig. 4D). This implies that all six Denisovan NOTCH2NL alleles produced NOTCH2NL protein at an extremely low level.
Unfortunately, the lack of other high-coverage Denisovan genomes prevents us from assessing whether this is an individual-specific genotype or whether similar NOTCH2NLC gene conversions were frequent in the Denisovan population. Importantly, this pattern of Exon1C-(X-low)-variant distribution in Denisovan NOTCH2NL genes, or anything similar to it, has not been observed in any of the analyzed genomes of Neanderthals or healthy modern humans (supplementary fig. S4C, Supplementary Material online).

Evolution of Modern Human NOTCH2NL Genes
Trends toward Lower NOTCH2NL Levels
Even though NOTCH2NLA and NOTCH2NLB are capable of producing a structurally similar NOTCH2NL protein, the protein levels they produce differ by 5-fold. In the SUN analysis, we find evidence of extensive gene conversion between the NOTCH2NLA and NOTCH2NLB loci: The median SUN depth shifts in favor of either allele in different regions of the loci, indicating that parts of the NOTCH2NLA-sequence are frequently overwritten by NOTCH2NLB-sequence and vice versa (fig. 5A). Most regions with a strong shift in distribution of NOTCH2NLA or NOTCH2NLB SUNs are intronic, not predicted to impact the structure and level of NOTCH2NL protein. However, the configuration of Exon1 in NOTCH2NLA and NOTCH2NLB shows a median allele depth strongly in favor of the Exon1A-(Low)-variant (fig. 5B). This is striking because it suggests that the vast majority of the population carries three or four alleles with the NOTCH2NLA-derived Exon1A-(Low)-variant and only one or zero alleles with the NOTCH2NLB-derived Exon1B-(High)-variant (fig. 5C). The shift in Exon1A-(Low)-variant distribution was confirmed in 49,593 exomes from the UK Biobank (Van Hout et al. 2019) (supplementary fig. S5A, Supplementary Material online) and was also observed in the genomes of ancient modern humans (fig. 5D). The observed imbalance in distribution of Exon1A-variants indicates that the Exon1B-(High)-variant, producing the highest levels of NOTCH2NL protein, is being lost or actively being purged out from the modern human population by gene conversion. The increase of the Exon1A-(Low) variant frequency to three or four alleles per individual is likely caused by gene conversion between the NOTCH2NLA and NOTCH2NLB loci, which can occur during meiosis or in early embryonic development for very unstable loci (Chen et al. 2007; Bruder et al. 2008; Vadgama et al. 2019).

FIG. 5. Exon1 variant frequencies in modern human and ancient genomes. (A) Median allele count for each of the NOTCH2NLA- and NOTCH2NLB-specific SUNs along the NOTCH2NL locus in Simons diversity genomes (N = 279). (B) Zoomed region of Exon1, orange arrowhead indicates Exon1B-(High) (ATG)/Exon1A-(Low) (ATA) variant positions. (C) Distribution of Exon1A-(Low) and Exon1B-(High) (inferred) variants in Simons diversity genomes. Expected distribution models equal frequency of both variants. Vertical dashed lines indicate medians. N = 279, Kolmogorov–Smirnov test: P < 2e-16. (D) Analysis of Exon1A-(Low) and Exon1B-(High) (inferred) variant frequency in modern humans and archaic genomes. Red lines indicate medians.
Fig. 6. Additional deleterious NOTCH2NL variants are present specifically in humans. (A) Overview of the R113* and Exon2B-(splice-mut) deleterious variants on NOTCH2NL protein structure. (B) R113* and Exon2B-(splice-mut) allele count in modern human and archaic genomes. (C) UK Biobank data for SAS, AMR, and EUR ancestries showing association of Exon1A-(Low) frequency with R113* frequency, Exon2B-(splice-mut) frequency, and their

Lodewijk et al. doi:10.1093/molbev/msaa104 MBE
Spreading of Modern Human-Specific Deleterious Variants Indicates Strong Compensatory Mechanisms Despite the relatively high frequency of Exon1^A-(Low) variants in NOTCH2NL and NOTCH2NLB, some individuals still carry a relatively high number of Exon1^B-(High) variants. We found that individuals with a relatively high number of the Exon1^B-(High) variant and low number of the Exon1^A-(Low) variant often carry a nonsense SNP (R113*) in NOTCH2NLB, which leads to a premature stop-codon and a severely truncated NOTCH2NL protein (fig. 6A). In addition, we found another variant in the splice acceptor sequence of exon 2 (Exon2B-(Splice-mut)) (fig. 6A and supplementary fig. S6A, Supplementary Material online). This variant falls outside the coding region and therefore was not detected before. The AG > GG mutation is predicted to lead to an alternative splicing event, resulting in a frameshift and truncation of NOTCH2NL proteins at amino acid 30 (Dougherty et al. 2018). On hg38, this variant is annotated in NOTCH2NLB and it is present at a high allelic frequency in human genomes from the Simons diversity data (supplementary fig. S6B, Supplementary Material online) and the UK Biobank (supplementary fig. S6C, Supplementary Material online). The R113* variant is less frequently observed. Surprisingly, the splice acceptor variant Exon2B-(Splice-mut) and the R113* mutation were not found in any of the currently available Neanderthal or Denisovan genomes (fig. 6B and supplementary fig. S6B, Supplementary Material online) and are therefore recently evolved human lineage-specific adaptations. Both loss-of-function variants appear to be common in the South-Asia (SAS), American (AMR), and European (EUR) ancestries and are only sporadically present in East-Asian (EAS) or African (AFR) ancestries in the UK Biobank data (fig. 6B and supplementary fig. S6C and D, Supplementary Material online). Segregation of the disruptive alleles appeared to be nonrandom because we found a clear correlation between the individual's number of Exon1^A-(Low) or Exon1^B-(High) variants and the presence of disruptive R113* and Exon2B-(Splice-mut) mutations. Individuals with a relatively high number of the Exon1^B-(High) variant, often carry one or two alleles of the disruptive R113* mutation in NOTCH2NLB (fig. 6C-upper panel and supplementary fig. S6C and D, Supplementary Material online). A strikingly similar pattern was observed for the Exon2B-(Splice-mut) mutation (fig. 6C-middle panel and supplementary fig. S6C and D, Supplementary Material online). Conversely, individuals with a relatively higher number of Exon1^A-(Low) variants are more likely to lack either the R113* or splice acceptor mutations in NOTCH2NLB (fig. 6C-lower panel and supplementary fig. S6C and D, Supplementary Material online). In the EAS population, the more sporadic occurrence of both disruptive NOTCH2NL variants correlates with an overall higher Exon1^A-(Low) frequency instead (supplementary fig. S6D and E, Supplementary Material online). This reveals a complex pattern of NOTCH2NL configurations, where multiple structural variants in NOTCH2NLB, the gene that has the largest contribution to the overall NOTCH2NL levels, seem to act complementary to reduce NOTCH2NL protein levels. In the Simons diversity data set, we observe highly similar patterns, but this analysis lacked statistical power due to the relatively small sample size per ancestry group (supplementary fig. S7A–C, Supplementary Material online). Taken together, our findings suggest that a relatively high load of the Exon1^B-(High) variant often co-occurs with the presence of nonsense variants in NOTCH2NLB. Our data suggest that on the individual's genome level, gene conversion of the Exon1^B-(High) variant into the Exon1^A-(Low) variant acts in concert with nonsense variants in NOTCH2NLB to reduce overall NOTCH2NL protein level. This seems particularly relevant because we observe a strong dosage-dependent effect of NOTCH2NL on Notch signaling activation (fig. 6D and E), indicating that NOTCH2NL dosage is tightly associated with its functional output, which in the brain is controlling cortical neurogenesis. Altogether, the identification of Neanderthal-, Denisovan-, and modern human-specific coding variants and their complementary functional impact on NOTCH2NL protein levels suggests that the optimal level of NOTCH2NL protein has been under strong selective pressure in recent human evolution and is still being optimized in the human population (fig. 6F).

Discussion
The detection of multiple lineage-specific coding variants and the rapid spread of some of them throughout modern human genomes shows that the structure of human NOTCH2NL genes has been subject to ongoing adaptive evolution since the split of modern humans, Neanderthals, and Denisovans from our common ancestor ~600,000 years ago. This is corroborated by the presence of additional copies of NOTCH2NL in Neanderthal and Denisovan genomes and the unusual configuration of six NOTCH2NL-derived Exon1^C-(X-low) variants in the Denisova3 genome. Notably, none of the 279 modern human individuals analyzed in detail in this study showed similar configurations and it is questionable whether such configurations are found in the healthy human population. This raises questions about the health state of the juvenile Denisovan female from the Denisova3 genome.
NOTCH2NL genes: Based on the high frequency of loss-of-function variants in NOTCH2NL genes in modern humans, it would be expected that a decent proportion of the population would have a genomic configuration without any functional NOTCH2NL allele. This is clearly not the case, as the skewed allele distributions that we report points toward purifying selection in order to maintain at least one functional copy of NOTCH2NL. This suggests that in present day humans, a certain minimal level of NOTCH2NL protein is required for normal human development. The observed evolutionary changes in NOTCH2NL composition could be the result of evolutionary adaptations that took place in any of the tissues where NOTCH2NL is expressed, including the developing brain. Even though this remains speculative at the moment, the trend toward lower levels of NOTCH2NL proteins in the human lineage could be correlated to previous observations suggesting a progressive reduction of human brain size that started about 60,000 years ago (Henneberg 1988; Bednarik 2014)

Effectively, NOTCH2NL dosage, which is the total of protein produced by all NOTCH2NL loci, may vary between individuals but seems to stay within certain upper and lower ranges. Our new insights regarding the effect of Exon1 variants on NOTCH2NL protein levels may also help in understanding to what extent NOTCH2NL genes contribute to 1q21.1 Copy-number variation (CNV)-related phenotypes. Specifically for NOTCH2NL-mediated effects, like potentiating NOTCH signaling, CNVs of an allele carrying the Exon1B (High) variant may have a much larger effect than CNVs of an allele carrying the Exon1A(Low) variant. Identifying which NOTCH2NL loci are affected by gain and loss of alleles will have to be complemented by distribution analysis of Exon1A (Low), Exon1C(X-low), R113*, and Exon1B(Splice-mut) variants as they are major determinants of NOTCH2NL levels. The realization that gene conversion between functionally different NOTCH2NL genes can contribute to the rapid adaptation of the human species to establish lower levels of NOTCH2NL protein, may prove to be an example for other unstable loci that are characterized by recent segmental duplications. As some of these, like the 1q21.1 locus, are associated with disease, it will be intriguing to see if gene conversion also affects genetic configurations of such loci.

Ever since the availability of genomes from Neanderthals, Denisovans, and ancient humans, the question was raised which modern human-specific coding variants may hold clues to how our species evolved over the last ~600,000 years. Here, we discovered such variants in the NOTCH2NL genes, a gene family that emerged in humans about 4 M.a. The role of NOTCH2NL genes in human brain development and their involvement in 1q21.1 CNVs associated with a wide variety of neurological disorders emphasizes the importance of the discoveries we describe here: Even if the driving forces of the observed evolutionary changes lie outside the brain, the recent and ongoing structural evolution of human NOTCH2NL genes suggests that the tightly coordinated process of human cortical neurogenesis is still subject to fine-tuning.
Materials and Methods

**NOTCH2NL Copy Number Analysis from Whole-Genome Sequencing Data**

Fastq files were imported from the EBI SRA to the Galaxy EU or US server (Afgan et al. 2018). For Simons diversity data, only the R1 data were used. Reads were trimmed using Trimomatic (Galaxy v0.36.5) by the following settings: SLIDINGWINDOW: 4, 20 and MINLEN: 30. The remaining reads were mapped to the NOTCH2NL-masked hg38 reference genome using Bowtie2 (Galaxy v2.3.4.2), using single-end, very sensitive end-to-end settings. Sequence read depth per genome was ~15–30×. The BAM output files were sliced using samtools slice (Galaxy v2.0.1) with the coordinates chr1:118911553–121069626. Bedtools coverage (Galaxy v2.27.0.2) was applied to each sliced BAM file, reporting coverage for each position. The NOTCH2-single-copy region used is located at chr1:119908310–120087985. Each region was filtered for repeats using RepeatMasker, and only the nonrepeat intervals were used in coverage analysis. Mean coverage across both regions was calculated by averaging coverage per position. The mean coverage of the NOTCH2 + NOTCH2NL-multicopy region was divided by the mean coverage of the NOTCH2-single-copy region to infer NOTCH2NL copy-number per data set. BAM file data were visualized in the UCSC genome browser (Kent et al. 2002). For ancient DNA data sets which consisted of multiple libraries, each library was mapped separately and then merged. The Denisova3 run ERR141700 was omitted due to high sequence duplication. The following WGS data sets were used:

- For comparisons of the SUN analysis with previously assembled NOTCH2NL configurations (Fiddes et al. 2018), the following samples and data sets were used (Steinberg et al. 2014; Zook et al. 2016; Eberle et al. 2017; Regier et al. 2018; Audano et al. 2019; Marks et al. 2019):
  - NA24143: 10× genomics (GIAB), WGS (PRJNA200694), WXS (PRJNA200694)
  - NA24149: 10× genomics (GIAB), WGS (PRJNA200694), WXS (PRJNA200694)
  - NA24385: 10× genomics (GIAB), WGS (PRJNA200694), WXS (PRJNA200694)
  - NA12891: WGS and 10× WGS (PRJEB3381, PRJNA428496, PRJEB4252), PRJEB4663 (Slon et al. 2018) Denisova11

Separation of NOTCH2NL Copy Number per Allele Using SUNs

Based on the hg38 reference genome, single-nucleotide variants and indels were identified, via DNA sequence alignment of the NOTCH2NL-,-B, -C, or -R loci to the NOTCH2 locus. Only SUNs within the region chr1:119990474–12008798 were considered, as this is the maximal duplication size present in each of the NOTCH2NL loci based on the segmental duplication track in the UCSC genome browser hg38. The position of each of these SUNs per locus was stored in BED format. These
were used to generate .vcf format data per BAM file reporting the total read depth and variant (SUN) depth for these positions. This was done using samtools (v1.7) mpileup:

```bash
samtools mpileup -ufv hg38.fasta -t DP -t AD -l variant_positions.bed -Q 13 -q 0 -b datasets.txt > output.vcf
```

The relevant information to calculate SUN frequency per allele was extracted using bcftools (v1.7) query:

```bash
bcftools query -f '%CHROM\t%POS\t%REF\t%ALT{} [\t%DP\t%AD{}\t%AD {1}]\n' -H mpileup_output.vcf > mpileup_output_variants.vcf
```

The frequency per variant was calculated using these output files by dividing allele depth by each SUN (AD) by total depth (DP). For each locus, only SUNs with >0.67 frequency in the population were used for analysis to account for ambiguous or population-specific sites that may skew allele distribution calculation, such as known common SNPs. The frequency of the selected SUNs was averaged per locus and multiplied by the total number of alleles calculated previously based on sequence read coverage, to transform allele frequencies into allele counts. Since there are many SUNs for NOTCH2, NOTCH2NL, and NOTCH2NLC, they provide an accurate estimation for the allele count of these loci. For NOTCH2NL and NOTCH2NL, only a few SUNs are present and gene conversion phenomena happen frequently, which makes this procedure challenging. Therefore, to analyze these loci, we first subtracted the NOTCH2, NOTCH2NL, and NOTCH2NLC allele counts from the total allele count. The remaining alleles must be derived from NOTCH2NL and NOTCH2NL, and so, the remaining alleles were counted using the ratio of the average SUN frequency for NOTCH2NL and NOTCH2NL. These data were plotted in donut-charts using LibreOffice v6.1.0.3. For graphs showing the per-SUN allele count across the NOTCH2NL loci, the NOTCH2NL SUN count was inferred from the NOTCH2NL SUN count in the S′ region of the locus, where no NOTCH2NL SUNs are present. For example in modern humans there are four NOTCH2NL+NOTCH2NL loci, then the Exon1B-(High) allele count was calculated according to this: Exon1B-(High) allele count = 4 – Exon1A-(Low) count. Correction for NOTCH2 > NOTCH2NL gene conversion was done for genomes that showed three alleles NOTCH2. These showed a concordant decrease of one allele NOTCH2NL based on both the coverage analysis and SUN analysis. This difference was corrected for, in example, three alleles NOTCH2 and one allele NOTCH2NL in one individual were corrected to two alleles NOTCH2 and two alleles NOTCH2NL. For separation of the Simons diversity genomes data per population, the sample metadata supplied with the data were used.

### Allele Frequencies in UK Biobank Exome Data

Reads mapping to NOTCH or NOTCH2NL genes were extracted from UK Biobank CRAM exome files (>20× coverage) mapped on hg38. As in these data sets, the reads are mapped to NOTCH and all NOTCH2 NL loci in hg38, the analysis was adjusted from the original analysis that used the masked hg38. For the Exon1A-(Low) variant (ATG > ATA), the following positions were analyzed:

Similarly, the Exon1B-Splice-mut variant information was derived from the following positions:

Nea1N232S variant (AAT > AGT) information was derived from the following positions:

Read depth and allele depth analysis using samtools and bcftools was then done for each locus with the following parameters:

```bash
samtools mpileup -ufv hg38.fasta -t DP -t AD -l variant_positions.bed -Q 13 -q 0 -b datasets.txt > output.vcf

bcftools query -f '%CHROM\t%POS\t%REF\t%ALT{} [\t%DP\t%AD{}\t%AD
```
Transfection for NOTCH2NL Variant Protein Analysis

HEK293 cells were seeded 24 h before transfection in a six-well plate. One hour before transfection, medium was replaced with 1,800 μl DMEM + GlutaMAX, high glucose and 10% HIFBS. The transfection mix per well was as follows: 500 ng of pCAGN1-NOTCH2NL or pCAGN1-EV, and 500 ng of pCAGEN-GFP were mixed in a total volume of 100 μl. Cells were washed twice in ice-cold PBS, then detached using a cell scraper (VWR 734-1527) and transferred on a 1.5-mm poly-acrylamide gel, consisting of two parts. The protein gel electrophoresis was done in 25 mM Tris–HCl pH 6.8, 0.1% APS, 0.1% sodium dodecyl sulfate (SDS), and 0.04% tetramethylethylenediamine (TEMED) and the stacking gel (5% acrylamide/Bis, 0.125 mM Tris–HCl pH 6.8, 0.1% APS, 0.1% SDS, and 0.1% TEMED). Twenty microilers of sample was loaded per well and 5 μl of marker (Thermofisher #26619) was used for reference. Electrophoresis was done in 25 mM Tris + 192 mM glycine buffer (Biorad 1610771) and 0.1% SDS. Protein was transferred to nitrocellulose membrane (Sigma 10600004), at 100 V for 2 h in Towbin buffer (25 mM Tris, 192 mM glycine, and 20% methanol). Blots were rinsed three times with demi-water, and transfer was checked by ponceau S staining. Blots were rinsed once in Tris buffered saline (20 mM Tris, pH 7.5, 150 mM NaCl) + 0.1% Tween (TBS-T), followed by incubation in blocking buffer (TBS-T + 5% w/v skim milk powder) for 90 min at room temperature on a shaking platform. Primary antibodies were incubated overnight at 4°C in TBS-T in 50-ml tubes on a tube roller. Antibodies used were rabbit anti-HA tag (1:6,000, Abcam ab9110) or rabbit anti-GFP (1:4,000, Abcam ab290). Blots were rinsed once in TBS-T and washed in TBS-T three times 15 min on a shaking platform. Secondary antibody goat anti-rabbit-HRP in TBS-T (1:20,000, Thermofisher 656120) was incubated for 60 min at room temperature. Blots were rinsed once in TBS-T and washed three times 15 min in TBS-T on a shaking platform. The Western blot was developed using a ChemiDoc MP imaging system (Biorad 1708280). Signals were quantified using Fiji ImageJ using the NOTCH2NL/GFP ratio.

Coculture NOTCH Reporter Assay

To monitor modulation of NOTCH2 activity by NOTCH2NL, a reporter assay was used. The pGL3-UAS luciferase reporter can be activated by S3-cleaved NOTCH2-Gal4-N1TAD receptor intracellular domain (Gal4 domain fused to NOTCH1 transactivation domain) (gifts of Arjan Groot and Marc Vooijs, MAASTRO Lab, Maastricht University). To achieve high levels of receptor activation, the cells transfected with pDNAs-NOTCH2-Gal4-N1TAD are cocultured with JAG2 expressing cells. Coculture with regular U2OS cells was
done as a control. pCAGN1-EV or pCAGN1-NOTCH2NL (derived from Addgene 51142) were cotransfected to measure effects of NOTCH2NL on reporter activity. pRL-CMV (Promega E2261) was used for normalization.

For transfection, U2OS cells were seeded in six-well plates at a density of 400,000 cells per well. For coculture assay, U2OS cells or U2OS-JAG2 cells were seeded in 12-well plates at a density of 110,000 cells per well. Twenty-four hours later, U2OS cells in six-well plates were transfected. The transfection complex per well was made by adding 2,500 ng plasmid DNA mix, as described in the table below, in 100 µl OptiMEM (Thermofisher 31985047). In a different tube, 8.33 µl PEI (1 mg/ml, Polysciences 23966) was added to 100 µl OptiMEM. One hundred microliters of each mix was combined, incubated 20 min at room temperature, and added to the well containing 2 ml of complete medium. Reactions were scaled accordingly to facilitate large-scale transfections. Six hours after transfection, the transfected cells were replated onto the 12-wells plate for coculture with U2OS or U2OS-JAG2 cells. Per well, medium was removed and cells were washed once with 1 ml PBS. Trypsin-EDTA (0.5 ml) in PBS was added up by pipetting up and down three times. Cell suspension was transferred to 15-ml conical tubes already containing 4.5 ml of complete medium. From the 12-well plates, the medium was removed and replaced by 1 ml of cell suspension. To control wells, 1 µl of 200 µM DBZ was added. Twenty-four hours after replating, the cells were isolated for luciferase assays using Dual-Luciferase Reporter Assay System (Promega E1980). Medium was removed and each well washed once with 0.5 ml PBS. A total of 150 µl of 1× passive lysis buffer (Promega E1941) was added per well and incubated 15 min on a rotating platform. Plates were wrapped in parafilm and stored at −80 °C. For analysis, 20 µl sample was pipetted to a 96-well optiplate (PerkinElmer 6005290). Samples were measured on a GloMax Navigator device (Promega GM2010), with the following settings: Injector 1, LARII buffer (volume 50 µl, speed 200 µl/s). Wait 2 s. Measure luminescence Luciferase (integration 10 s, readings 1, interval 0.3 s). Injector 2, Stop & Glo buffer (volume 50 µl, speed 200 µl/s). Wait 2 s. Measure luminescence Renilla (integration 10 s, readings 1, interval 0.3 s). For comparison of human, Neanderthal and Denisovan NOTCH2NL variants, the 48 ng pCAGN1-NOTCH2NL condition was used.

<table>
<thead>
<tr>
<th>6</th>
<th>16</th>
<th>48</th>
<th>140</th>
<th>420</th>
</tr>
</thead>
<tbody>
<tr>
<td>pGL3-UAS</td>
<td>1,050</td>
<td>1,050</td>
<td>1,050</td>
<td>1,050</td>
</tr>
<tr>
<td>pRL-CMV</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>pCAGEN-GFP</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>pcDNA5-NOTCH2-Gal4-N1TAD</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>pCAGN1-EV/-NOTCH2NL</td>
<td>5/6</td>
<td>14/16</td>
<td>41/48</td>
<td>120/140</td>
</tr>
<tr>
<td>pBlueScript (EV)/NOTCH2NL</td>
<td>1,315/1,306</td>
<td>1,279/1,200</td>
<td>959/</td>
<td></td>
</tr>
<tr>
<td>pRL-CMV</td>
<td>1,314/1,304</td>
<td>1,272/1,180</td>
<td>900</td>
<td></td>
</tr>
</tbody>
</table>

Amount of plasmid DNA (ng) transfected per condition. pCAGN1-EV/pCAGN1-NOTCH2NL denotes amount of plasmid used per condition accounting for molarity. pBlueScript amount was adjusted accordingly as well.

In Silico Analysis of Archaic Coding Variants

For multiple sequence alignment of NOTCH1, -2, and -3 EGF-L domains 6, the relevant sequences were retrieved from UniProt and compared using the alignment tool of UniProt. The EGF-L repeat domain consensus sequence was retrieved from Prosite: PDOC00021, EGF_3 PS00026. For MutPred2 and IUPred2A analysis, the archaic amino acid variants were introduced in the NOTCH2NL B protein sequence retrieved from UniProt (P0DPK3). MutPred2 was run with a P value threshold of 0.05. IUPred2A was used.
with the following settings: Long disorder, Context-dependent predictions: ANCHOR2.

**Statistics**

Luciferase reporter assay data were first analyzed using analysis of variance (ANOVA) by the R function "aov()." Significant groups were further tested with Tukey’s test using the R function “TukeyHSD().” Western blot data were analyzed in the same way, except for data presented in figure 3B and C, which showed unequal variance (Levene test P = 0.002) and were analyzed instead using Welch corrected ANOVA using the R functions “oneway()” with parameters “levene=TRUE” and “corrections=TRUE,” followed by Games–Howell test from function “posthocTGH(),” with parameter “method=games-howell” (R package “userfriendlyscience”). Population genetic data from Simons diversity genomes and UK Biobank exomes were first analyzed using Kruskal–Wallis tests by the R function “kruskal.test.” Significant groups were further tested with Dunn’s test, using the “dunn.test()” function (R package “dunn.test”). Distributions in figure 5 were analyzed in the same way, except for data presented in figure 3B and C, which showed unequal variance (Levene test P = 0.002) and were analyzed instead using Welch corrected ANOVA using the R functions “oneway()” with parameters “levene=TRUE” and “corrections=TRUE,” followed by Games–Howell test from function “posthocTGH(),” with parameter “method=games-howell” (R package “userfriendlyscience”). Population genetic data from Simons diversity genomes and UK Biobank exomes were first analyzed using Kruskal–Wallis tests by the R function “kruskal.test.” Significant groups were further tested with Dunn’s test, using the “dunn.test()” function (R package “dunn.test”).

**Code Availability**

All code and software used in this manuscript are described in the Materials and Methods section.

**Supplementary Material**

Supplementary data are available at *Molecular Biology and Evolution* online.

**Acknowledgments**

We thank Danielle Posthuma (VU, Amsterdam); Marco Hoekman and Marten Smidt (SILS, University of Amsterdam) for helpful discussions; the Simons Diversity Project and UK Biobank for human genome data; Arjan Groot and Marc Vooijs (Maastricht Radiation Oncology Lab) for reagents for Notch reporter assays; the Galaxy Project for use of their servers to reanalyze part of the WGS data; and people from the Jacobs lab for helpful discussion and comments on the manuscript. This research has been conducted using the UK Biobank resource under application number 16406. This work was supported by ERC starting grant ERC-2016-StG-716035 (F.M.J.J.).

**Author Contributions**

F.M.J.J. and G.A.L. conceptualized the study and performed the methodology; G.A.L., I.V., and J.E.S. validated the study; F.M.J.J. and G.A.L. performed the investigation; F.M.J.J., G.A.L., J.E.S., and D.P.F. reviewed and edited the manuscript; G.A.L. and D.P.F. performed the visualization; and F.M.J.J. performed the supervision, project administration, and funding acquisition.

**References**


Evolution of Human Brain Size-Associated NOTCH2NL.

Frieda T. Willems, Joke B. van Dijk, and Remko van der Helm

Department of Neurology, University Medical Center Utrecht, Utrecht, the Netherlands

Abstract

The human brain is one of the most complex and enigmatic organs of the body. Its size and structure have been shaped by evolutionary forces over millions of years. Understanding the genetic basis of brain size variation is crucial for gaining insights into the mechanisms underlying brain development and function. In this study, we aimed to identify genetic variants associated with human brain size using a genome-wide association study (GWAS) approach.

Methods

We performed a GWAS using data from the Human Genome Diversity Project (HGDP) and the 1000 Genomes Project (1KGP). We analyzed genome-wide markers for association with brain size, using the magnetic resonance imaging (MRI) data from the Human Connectome Project (HCP). We applied a mixed-effects model to account for population structure and relatedness.

Results

Our GWAS identified a significant association between two genetic markers and brain size. The first marker, located on chromosome 21q21.1, was associated with a 0.12 cubic centimeter (cc) decrease in brain volume per allele copy. The second marker, located on chromosome 14q24, was associated with a 0.21 cc increase in brain volume per allele copy.

Conclusion

These findings suggest that genetic variation on chromosome 21q21.1 and 14q24 is associated with human brain size variation. Further studies are needed to validate these associations and to understand the biological mechanisms underlying these genetic effects on brain development.