Differential Gene Expression in the Human Brain Is Associated with Conserved, but Not Accelerated, Noncoding Sequences

Kyle A. Meyer,[1] Tomas Marques-Bonet,[2,3,4] and Nenad Sestan*[,1,5]

[1]Department of Neuroscience and Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT

[2]Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain

[3]Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys, Barcelona, Spain

[4]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

[5]Departments of Genetics and Psychiatry, Section of Comparative Medicine, Program in Cellular Neuroscience, Neurodegeneration and Repair, and Yale Child Study Center, Yale School of Medicine, New Haven, CT

*Corresponding author: E-mail: nenad.sestan@yale.edu.

Associate editor: Naruya Saitou

Abstract

Previous studies have found that genes which are differentially expressed within the developing human brain disproportionately neighbor conserved noncoding sequences (CNSs) that have an elevated substitution rate in humans and in other species. One explanation for this general association of differential expression with accelerated CNSs is that genes with pre-existing patterns of differential expression have been preferentially targeted by species-specific regulatory changes. Here we provide support for an alternative explanation: genes that neighbor a greater number of CNSs have a higher probability of differential expression and a higher probability of neighboring a CNS with lineage-specific acceleration. Thus, neighboring an accelerated element from any species signals that a gene likely neighbors many CNSs. We extend the analyses beyond the prenatal time points considered in previous studies to demonstrate that this association persists across developmental and adult periods. Examining differential expression between non-neural tissues suggests that the relationship between the number of CNSs a gene neighbors and its differential expression status may be particularly strong for expression differences among brain regions. In addition, by considering this relationship, we highlight a recently defined set of putative human-specific gain-of-function sequences that, even after adjusting for the number of CNSs neighbored by genes, shows a positive relationship with upregulation in the brain compared with other tissues examined.

Key words: **conserved noncoding sequence, gene expression, brain.**

Article

Introduction

Noncoding regulatory elements play a central role in the spatiotemporal control of gene expression, and variation in regulatory sequences can underlie the evolution of species-specific phenotypes (Shapiro et al. 2004; Wray 2007; Carroll 2008; Linnen et al. 2013; Shibata et al. 2015; Indjeian et al. 2016). Using conservation as an indicator of putative regulatory elements, studies have searched for conserved noncoding sequences (CNSs) that harbor an excess of substitutions in humans (Pollard et al. 2006a, 2006b; Prabhakar et al. 2006a; Bird et al. 2007; Kim and Pritchard 2007; Lindblad-Toh et al. 2011), identifying accelerated CNSs (ACNSs) that are candidates for being regulators of species-specific changes in gene expression. *In vivo* enhancer assays have provided information on a number of human-accelerated CNSs (HACNSs). For example, human-specific mutations in one HACNS are capable of driving (Prabhakar et al. 2008), or perhaps derepressing (Sumiyama and Saitou 2011), the expression of a reporter gene in the developing limb. Other enhancer assays have identified HACNSs that produce expression patterns in the nervous system which are different or absent compared with the patterns directed by non-accelerated counterparts from other species (Capra et al. 2013; Kamm et al. 2013a, 2013b; Boyd et al. 2015). Following up on a reporter gene assay, Boyd et al. (2015) provided evidence that a human-accelerated regulatory enhancer HARE5, regulates the expression of *FZD8*. Through comparison of transgenic mouse lines in which mouse *Fzd8* was driven by either HARE5 or the chimpanzee counterpart, they demonstrated that *Fzd8* driven by HARE5 leads to an increase in neural progenitor proliferation and a larger neocortex.

At a more global and correlative level, analyses of genes that neighbor (i.e., are the nearest gene to) CNSs suggest that many HACNSs, and accelerated noncoding sequences in general, may influence expression in the brain. Gene ontology (GO) analysis indicates that genes neighboring HACNSs, compared with those neighboring CNSs as a whole, are disproportionately involved in neuronal adhesion (Prabhakar et al. 2006a). A meta-analysis that combined HACNSs with two other noncoding sets, human-accelerated regions (HARs) (Pollard et al. 2006a) and human-accelerated promoters (Haygood et al. 2007), found that genes neighboring these

© The Author 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

sequences, in contrast to genes whose coding regions exhibit acceleration, were enriched for neural development-related functions and had a higher tendency for elevated expression in the brain (Haygood et al. 2010). On the other hand, comparison of putative regulatory elements identified by ChIP-seq profiling of promoter and enhancer marks during early cortical neurogenesis in humans, rhesus macaques, and mice (Reilly et al. 2015)—as well as profiling in adult human, chimpanzee, and macaque brains (Vermunt et al. 2016)—did not find any enrichment of human-accelerated sequences in peaks that were unique to the human samples.

Focusing on spatial differential expression within the human brain, a handful of studies have examined the frequency at which differentially expressed (DEX) genes are the nearest gene to a HACNS (Johnson et al. 2009; Lambert et al. 2011; Miller et al. 2014). An analysis of multiple regions from the mid-fetal human brain found that, of genes nearest to CNSs, those that are DEX among brain regions are more likely than non-DEX genes to be the nearest gene to a HACNS (Johnson et al. 2009). An independent study of gene expression in two cortical regions during fetal development also found an association between differential expression and accelerated noncoding sequences (Lambert et al. 2011). Importantly, the authors note that this pattern does not appear to be specific to human-accelerated sequences, as genes neighboring CNSs with elevated substitution rates in other species are also more likely to be DEX between the two human brain regions studied. Thus, species-specific regulatory changes, at least as represented by ACNSs and in the context of nervous system development, may have preferentially operated within the set of genes that had pre-existing differential expression patterns.

Here we re-examine the relationship between differential expression and ACNSs. We demonstrate that when a gene neighbors an ACNS it often neighbors a high number of non-ACNSs. Accounting for this relationship reveals that genes neighboring ACNSs are DEX within the brain at a rate that is consistent with what is expected given the number of CNSs to which each gene is nearest (referred to as "adjacent" CNSs). The positive relationship between differential expression in the brain and the number of adjacent CNSs exists across time periods and is also present in most pairwise expression comparisons between non-neural adult tissues. Finally, by adjusting for the number of adjacent CNSs, we provide support for the hypothesis that the regulatory evolution captured by a recently defined set of putative human-specific elements has disproportionately targeted genes that are upregulated in the human brain.

Results

Genes Neighboring ACNSs Tend to be near More CNSs

To test whether genes that are DEX among brain regions neighbor accelerated regions more frequently than non-DEX genes do, previous studies (Johnson et al. 2009; Lambert et al. 2011) have focused on genes that are nearest to at least one CNS. From this subset, the proportion of genes that were nearest to an accelerated element was compared between DEX and non-DEX genes. We first took a similar approach in order to verify that we obtained consistent results using our processing steps and differential expression classification.

We classified a gene as DEX among brain regions based on differential expression between at least two brain regions, performing separate classifications with the Johnson et al. (2009), Lambert et al. (2011), and Kang et al. (2011) data sets (supplementary table S1, Supplementary Material online). For the Johnson et al. and Kang et al. studies, which profiled expression in both neocortical and non-neocortical samples, we considered either differential expression among brain regions, with neocortical areas treated as a single brain region, or differential expression among neocortical areas. The Kang et al. data set was restricted to samples from time period 6 (supplementary table S2, Supplementary Material online), a late mid-fetal period that is crucial for the formation of neuronal circuits (see Silbereis et al. 2016) and that overlaps with the time points analyzed in the previous studies. HACNSs, chimpanzee-accelerated CNSs (CACNSs), and mouse-accelerated CNSs (MACNSs), as identified by Prabhakar et al. (2006a), were intersected with gene coordinates to determine whether a gene (more specifically, either bound of its longest transcript) was the nearest gene to any sequence in each group of ACNSs. To compare these sets of genes with the larger set of genes nearest to any CNS, we generated a superset of CNSs (182,682 in total), of which HACNSs, CACNSs, and MACNSs make up a small proportion (0.5%, 0.6%, and 2.5%, respectively), by filtering 8-way vertebrate phastCons elements (Siepel et al. 2005) to those that reside in noncoding sequences and have a conservation score of at least 400.

Following earlier studies, we intersected the nearest gene and differential expression classifications to assess whether DEX genes disproportionately neighbor ACNSs. When CNS-neighboring genes were restricted to DEX genes, the proportion of genes neighboring a HACNS was greater in this subset than in the overall set of CNS-neighboring genes (supplementary table S3, Supplementary Material online). As noted previously (Lambert et al. 2011), DEX genes within the mid-fetal human brain showed a similar enrichment for neighboring sets with acceleration in other species, represented here by CACNSs and MACNSs. Thus, despite not attempting to closely match details of previous analyses (see "Materials and Methods" section), we find that this approach for examining the relationship between differential expression and neighboring an ACNS yields results that are consistent with previous findings and that hold across mid-fetal brain samples from independent studies.

A limitation of the above method, however, is that it only considers whether a gene neighbors at least one CNS, which loses a substantial amount of information because the number of adjacent CNSs displays a wide range. Of genes that are nearest to at least one non-ACNS (which we will refer to as OCNSs for "other" CNSs), over three quarters neighbor more than one OCNS (fig. 1A and supplementary fig. S1A, Supplementary Material online). For each ACNS type, the proportion of genes that neighbor more than one element

is lower but still above a quarter (supplementary fig. S1*B*, Supplementary Material online). Importantly, if a gene is the nearest gene to a HACNS, it is likely that it is also the nearest gene to many OCNSs (fig. 1*B*). This relationship with OCNSs is also present for genes nearest to CACNSs (fig. 1*C*) and MACNSs (supplementary fig. S2*A*, Supplementary Material online).

Genes Neighboring More CNSs Are More Likely to be DEX in the Brain

The increased tendency for ACNS-neighboring genes to neighbor a large number of OCNSs raises the possibility that the number of adjacent OCNSs is responsible, at least in part, for the positive correlation between neighboring an ACNS and differential expression. In order for this to be the case, the number of adjacent OCNSs must show a positive relationship with differential expression. Indeed, grouping genes by the number of adjacent OCNSs suggests that genes neighboring more OCNSs are more frequently DEX. For example, of genes that are nearest to at least ten OCNSs (18% of the analyzed genes), 33% were classified as DEX among brain regions in period 6 of the Kang et al. data set, while only 9% of genes nearest to fewer than ten OCNSs were DEX. In general, binning genes by the number of adjacent OCNSs reveals that the proportion of DEX genes tends to increase for bins containing genes that are nearest to more OCNSs (supplementary fig. S3, Supplementary Material online).

The positive relationship between the number of adjacent OCNSs and differential expression, together with the covariation between the number of adjacent ACNSs and the number of adjacent OCNSs, motivates the development of a model to assess whether the rate of differential expression for genes neighboring ACNSs exceeds what is expected given the number of OCNSs that each gene neighbors. For this purpose, we constructed a probit regression model, taking differential expression status as the binary response variable and incorporating counts of OCNSs and of each type of ACNS that a gene neighbors as predictors (see "Materials and Methods" section). In this generalized linear model (GLM) variant, each predictor contributes to an overall linear predictor, which represents a z-score. To link the continuous z-score to a dichotomous response variable, the z-score is mapped, via the standard cumulative normal distribution, onto probability space. The coefficient for each predictor, conditioned on the other predictors of the model, is the predictor's per unit contribution to the z-score. Thus, a positive coefficient reflects an increase in probability, though the extent of the increase depends on the initial z-score value. Consequently, examining the change in marginal probability for a coefficient, with other parameters at reasonable values, is useful for interpreting the coefficient.

We first fit this model with differential expression status among mid-fetal brain regions, classified from period 6 of the Kang et al. data set, as the binary response variable. For each set of elements (OCNSs, HACNSs, CACNSs, and MACNSs), the predictors included the log-transformed count of the number of elements that a gene neighbors. The probit coefficient for the OCNS predictor was 0.30–0.34 (95% credible interval [CI]), corresponding to the marginal probability of differential expression increasing by 14–16 percentage points for a gene neighboring ten OCNSs (fig. 2*A*). Genes nearest to more OCNSs also had a higher probability of being DEX among neocortical areas (fig. 2*A*; probit coefficient of 0.26–0.34, 95% CI), which, starting at the lower base rate of differential expression among neocortical areas, maps to a marginal probability increase of 2–3 percentage points.

Due to our procedure for assigning genes as targets of CNSs, we were concerned that much of the relationship between the number of adjacent CNSs and differential expression may reflect variation in locus length, in a manner similar to the annotation bias described by Taher and Ovcharenko (2009). Specifically, a gene with more intergenic space and longer introns, all else being equal, will be assigned as the nearest gene to more CNSs by chance, and the locus length of a gene may be positively correlated with whether a gene is DEX. We assessed the influence of locus length by considering one of two predictors. First, we considered the log-transformed locus length of each gene. This predictor had a positive relationship with differential expression among brain regions, but the coefficient for the OCNS count predictor remained well above zero (0.19–0.26, 95% CI). Second, we generated nearest gene counts for 15 sets of random noncoding coordinates (which were similar to the CNS set in terms of length and number) and summarized across these sets to

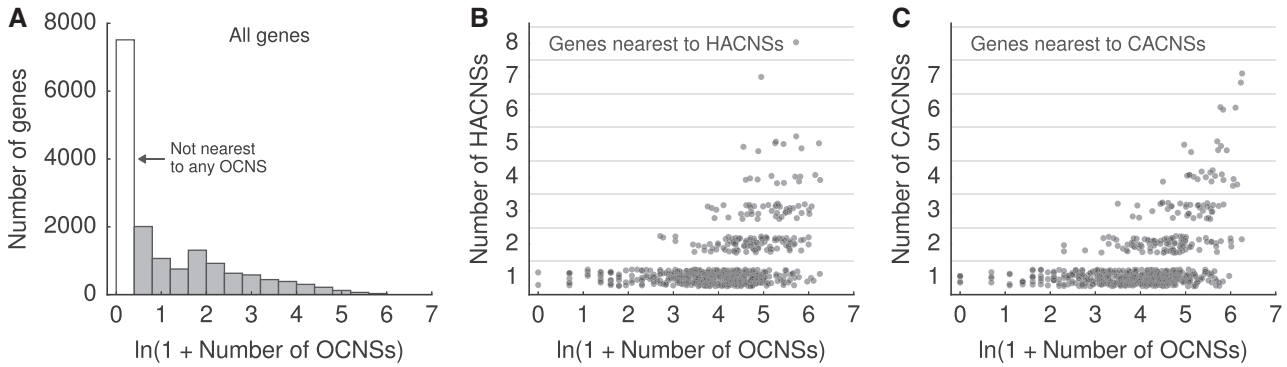

Fig. 1. Number of OCNSs neighbored by genes. (*A*) The distribution of the number of OCNSs neighbored by all genes. (*B*–*C*) The distribution of the number of OCNSs neighbored by genes that also neighbor at least one HACNS (*B*) or CACNS (*C*). The Spearman correlation coefficients for the relationships shown in *B* and *C* are 0.50 and 0.47, respectively. Points are jittered along the *y*-axis.

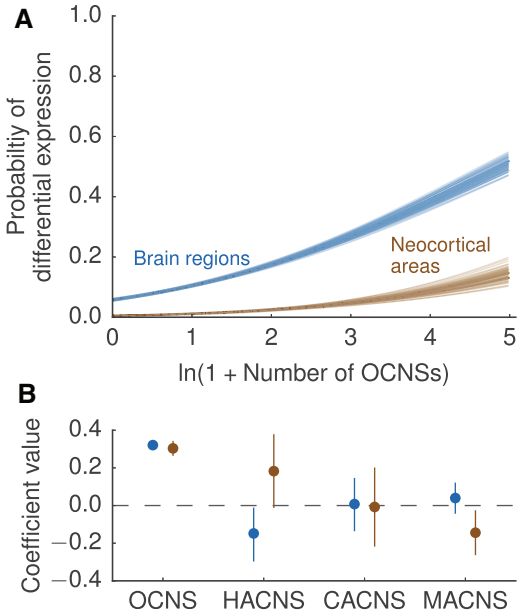

FIG. 2. Differential expression in the mid-fetal human brain of genes neighboring CNSs. (*A*) The probability of differential expression among brain regions (blue) or among neocortical areas (brown), given the number of adjacent OCNSs, was calculated using a probit regression model with predictors that indicated the number of elements (log-transformed) in each category—OCNSs, HACNSs, CACNSs, or MACNSs—that a gene neighbors. Separate models were run for the neocortical and regional response variables. Each set of lines represents 100 randomly selected realizations from the posterior distribution. The x-axis is restricted to the range of OCNSs for which a gene was observed to neighbor at least one OCNS but not any ACNS. (*B*) The probit regression coefficients for each CNS predictor. Intervals represent 95% CIs.

form an aggregate predictor. When compared with the locus length predictor, this random analog of the OCNS count predictor, which we refer to as the "target size" predictor, provides a more direct measure of the probability that a gene is assigned as a nearest gene for a set of random noncoding regions. The probit coefficient for the target size covariate was positive (0.16–0.24, 95% CI), but, as with the locus length predictor, the coefficient for the OCNS count predictor remained positive (0.17–0.23, 95% CI), with an increase of 8–11 percentage points in the marginal probability of differential expression when the target size predictor is set to its mean value. These results suggest both that target size is important to consider as a covariate and that the OCNS association is not simply a consequence of target size.

The OCNS count predictor, as a whole, carries information beyond what is captured in a random target size predictor, but the impact of the target size adjustment may depend on the location of the OCNSs. We categorized OCNSs into three groups: 1) located in the intron of a gene, 2) located <100 kb from the nearest gene, or 3) located 100 kb or more from the nearest gene. Predictors formed from these sets were substituted for the single OCNS count predictor. Ignoring the target size predictor, all groups showed a positive correlation with differential expression, with intronic OCNSs showing a stronger relationship (supplementary fig. S4, Supplementary Material online; probit coefficient of 0.23–0.28, 95% CI) than both OCNSs at <100 kb (0.15–0.22, 95% CI) and those at or beyond 100 kb (0.05–0.11, 95% CI). When the target size predictor was also considered, the coefficients of predictors for OCNSs within introns and within 100 kb survived adjustment (0.14–0.19 and 0.09–0.16, respectively, 95% CIs), while the coefficient of the predictor for OCNSs at 100 kb or farther no longer supported a positive relationship with differential expression status (−0.04 to 0.02, 95% CI).

To check the sensitivity of these estimates to the phastCons conservation score threshold used to define CNSs, we varied this cutoff, which was initially set to 400 to match the criteria used in the identification of HACNSs (Prabhakar et al. 2006a). From a score of 150, which captures the lowest score included in this phastCons set, the OCNS coefficient retained a relatively steady value up to a threshold of 400, where the OCNS coefficient began to drop with increasing thresholds, indicating that information relevant for the association is lost at these higher thresholds (supplementary fig. S5, Supplementary Material online). At the lower end, the number of adjacent elements is strongly correlated across sets with different score thresholds. For example, if we define the nearest gene count predictors from one of two groups, either CNSs that have scores above 150 but below 300 or those that have scores above 300 but below 400, the Spearman correlation coefficient between these predictors is 0.90.

Because the response vector encodes whether each gene was classified as DEX in at least one of six regions, it is possible that one region is primarily contributing to the correlation between differential expression and the number of adjacent OCNSs. However, constructing alternative response vectors where a single region is dropped suggests that the signal does not depend on any one region (supplementary fig. S6, Supplementary Material online).

Differential Expression of Genes Neighboring ACNSs

The model described earlier allows us to estimate the extent to which genes neighboring HACNSs show a higher tendency of being DEX than expected for the number of adjacent OCNSs. Returning to the first model reported—where differential expression was classified using samples from period 6 of the Kang et al. data set and nearest gene count predictors were included for OCNSs, HACNSs, CACNSs, and MACNSs—estimates of the association between the number of adjacent HACNSs and differential expression among brain regions (probit coefficient of −0.30 to −0.01, 95% CI) and among neocortical areas (−0.01 to 0.38, 95% CI) showed wide intervals that do not provide clear evidence of a positive association (fig. 2*B*). A similar pattern was also observed for the CACNS and MACNS predictors (fig. 2*B*). In contrast, when predictors for all elements except HACNSs (i.e., OCNSs, CACNSs, and MACNSs) were dropped, the HACNS coefficient showed a strong positive relationship with the differential expression response variable for brain regions (0.81–1.05, 95% CI) and for neocortical areas (0.76–1.06, 95% CI), as the HACNS predictor then conveyed indirect information about the number of adjacent OCNSs.

The OCNS count predictor also appears to be relevant for the relationship reported between HARs and differential expression (Lambert et al. 2011). A log-transformed count predictor of the number of adjacent HARs showed a positive relationship with regional (probit coefficient of 0.77–1.12, 95% CI) and neocortical differential expression (0.78–1.21, 95% CI) when the OCNS count predictor was not included. However, as with genes neighboring ACNSs, genes neighboring HARs tend to be nearest to many OCNSs (supplementary fig. S2*B*, Supplementary Material online). When we adjusted for the number of adjacent OCNSs, filtered to those that do not overlap with a HAR, we obtained wide coefficient estimates that are consistent with no association (regional differential expression status: −0.33 to 0.02, neocortical differential expression status: −0.11 to 0.36, 95% CIs), similar to what was observed for HACNSs.

Although the coefficient estimates are inherently tied to the chosen predictor transformation, the general pattern does not depend on a specific representation of the ACNS predictors (e.g., being a count rather than binary predictor or being log-transformed), provided that some form of a count predictor for OCNSs is incorporated into the model (supplementary table S4, Supplementary Material online). The results also do not appear to be sensitive to the method used to call differential expression, nor specific to the Kang et al. data set (supplementary fig. S7, Supplementary Material online).

Association between Regional Differential Expression and CNSs Persists across Time

The results described to this point have focused on differential expression during a mid-fetal time period that overlaps with the time periods from previous studies. Expanding the analyses to consider samples from all the time periods of the Kang et al. data set (supplementary table S2, Supplementary Material online), we found that the strength of the association between regional differential expression and the OCNS count predictor varied across time periods but remained positive (fig. 3; for all periods, the lowest 95% CI bound of any OCNS coefficient was 0.12). The largest drop in the value of the OCNS coefficients occurred between prenatal periods 1 and 2, coinciding with the largest increase in the number of DEX genes (from 282 to 1,428) between any consecutive time periods.

Because a gene that is DEX in one time period is predictive of whether that gene is DEX in the next period, the association in later time periods may be driven by DEX genes from earlier time periods. To test this, we reran the regression for each time period with a predictor that indicated the differential expression status for the previous time period. The association decreased, as expected, but persisted in most time periods (fig. 3). Together, these analyses suggest that, while the OCNS coefficients do show notable variation across time periods, the tendency for DEX genes to neighbor a higher number of OCNSs is not exclusive to any time period or group of time periods.

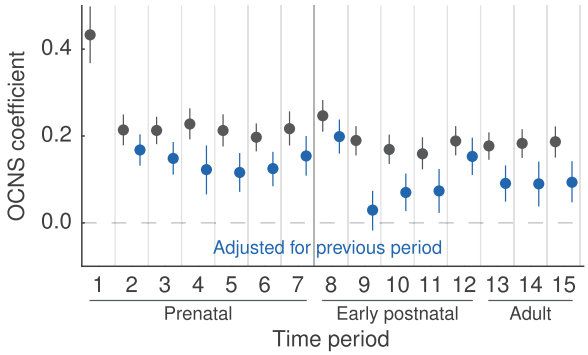

Fig. 3. Relationship of OCNSs and differential expression among human brain regions across time. The coefficients for the OCNS count predictor were estimated with a probit regression model. The adjusted coefficients (blue) were generated from an extended model that also included an indicator of whether each gene was DEX in the last time period. Intervals represent 95% CIs.

Association of CNSs with Human Inter-Tissue Differential Expression

To assess whether the relationship between the number of adjacent OCNSs and differential expression that we observed within the brain is also present for differential expression between tissues, we used a subset of tissues available in the GTEx expression data set (The GTEx Consortium 2015), grouping the samples according to the adult Kang et al. time periods (supplementary table S2, Supplementary Material online).

Using the nearest gene count predictors described earlier, we performed separate regressions for each pairwise tissue comparison, where the response variable was an indicator vector that represented whether a gene was upregulated in a given tissue compared with another (fig. 4). Nearly three quarters of the pairwise comparisons showed a positive relationship with the number of adjacent OCNSs, the overall average coefficient being 0.07 (0.17 for contrasts with upregulation in the cerebral cortex or cerebellum, 0.04 for others). The patterns of OCNS coefficient values appear to be consistent across these adult time periods (supplementary fig. S8, Supplementary Material online) and do not seem to simply mirror the number of upregulated genes (supplementary fig. S9, Supplementary Material online). Although comparing the coefficients directly is complicated by the dependence among the response vectors, OCNS coefficients for comparisons where the upregulated member is the cerebral cortex or cerebellum comprised some of the highest values for all GTEx tissues examined. In contrast to OCNSs, a consistent positive relationship was not observed for the HACNS, CACNS, or MACNS predictors (supplementary figs. S10 and S11, Supplementary Material online).

In addition to gene expression data, we used ChIP-seq data from the Roadmap Epigenomics Mapping Consortium to examine H3K4me1 and H3K4me3 signals in 26 tissues, comparing OCNSs to ACNSs, as well as OCNSs and ACNSs to the target size predictor (i.e., random noncoding sequences). In 24 out of 26 of the tissues, we observed that H3K4me1 signals, a correlate of enhancer activity, were shifted toward higher

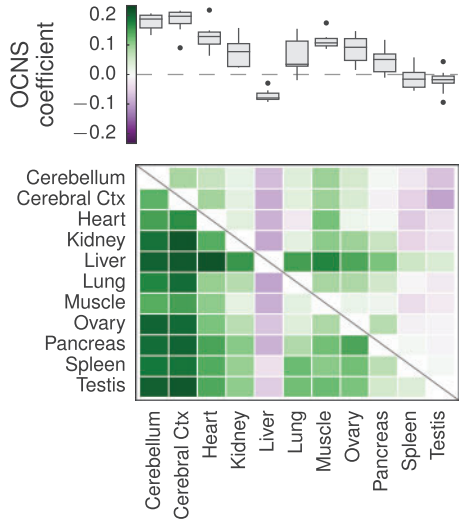

Fig. 4. Association of OCNSs with inter-tissue differential expression. OCNS probit coefficients were estimated for pairwise upregulation vectors generated with GTEx samples, restricted to a subset of tissues and to samples that fall within period 13 of the Kang et al. data set. (Heat maps for period 14 and period 15 samples are shown in supplementary fig. S8, Supplementary Material online) For each heat map cell, the response vector indicated whether genes were upregulated in the column tissue compared with the row tissue. The boxplots summarize all values for a column, with the box covering the interquartile range and the whiskers extending beyond the box with a length of 1.5 times the interquartile range. Ctx: cortex.

values in OCNSs relative to HACNSs, as indicated by a Mann-Whitney U test where a Bonferroni-adjusted *P* value below 0.05 was classified as a positive result (supplementary table S5 and supplementary fig. S12, Supplementary Material online). A similar relationship was observed for CACNSs but not MACNSs, as the H3K4me1 signal distribution for OCNSs was found to be shifted rightward of the distribution for MACNSs in only 5 out of 26 of the tissues (supplementary table S5 and supplementary fig. S12, Supplementary Material online). The factors resulting in the elevated H3K4me1 signals in MACNSs relative to HACNSs and CACNSs are unclear, although the larger size of the MACNS set may be relevant. In contrast to H3K4me1 signals, H3K4me3 signals, a correlate of promoter activity, showed fewer differences between OCNSs and ACNSs, with an increase in the signal in OCNSs compared with HACNSs, CACNSs, or MACNSs being found in 7, 15, and 5 of the 26 tissues, respectively.

When compared with H3K4me1 signals in random noncoding sequences, H3K4me1 signals in OCNSs, HACNSs, CACNSs, and MACNSs were elevated in most tissues (26, 21, 20, and 24 of 26 tissues, respectively; supplementary table S5, Supplementary Material online). Similar results were also observed for H3K4me3 signals. Thus, while the regulatory profiles of OCNSs, HACNSs, CACNSs, and MACNSs appear to differ from each other, all of these sets exhibit regulatory-associated ChIP-seq signals beyond what would be expected for random noncoding sequences.

Human-Specific Gain-of-Function Sequences Are Associated with Upregulation in the Brain after OCNS Adjustment

ACNSs, as well as HARs, are identified by lineage-specific divergence in sequences that are generally conserved across a set of species, a process that likely couples the distributions of OCNSs and accelerated elements. To explore a case where this coupling may be less pronounced, we analyzed two sets of sequences that were recently defined with a machine learning approach that classified segments of the genome as constrained or unconstrained using human population data (Schrider and Kern 2015). By intersecting these sets with interspecies conservation data, the authors identified sequences that show conservation in other species but are unconstrained in humans (referred to as losses-of-function or LOFs). With the reverse logic, sequences that show constraint in humans but not across other species were classified as gains-of-function (GOFs). Because the definition of GOF depends on the absence rather than presence of a CNS, we reasoned that the distribution of GOFs is more decoupled from CNSs as a whole. Consistent with this idea, the median number of adjacent OCNSs for genes nearest to at least one noncoding LOF (ncLOF) is 57, while, for genes nearest to at least one noncoding GOF (ncGOF), the median number of adjacent OCNSs is 16. For comparison, the median number of adjacent OCNSs for genes nearest to at least one HACNS is 66.

Incorporating nearest gene count predictors for ncGOFs and ncLOFs into the probit model described earlier, we found that both ncGOFs and ncLOFs showed a positive relationship with differential expression among brain regions across all time periods when the number of adjacent OCNSs and the target size predictor were ignored (fig. 5*A*). Following adjustment for the OCNS and target size predictors, neither ncGOFs nor ncLOFs showed a clear enrichment for being neighbored by genes that are DEX within the brain (fig. 5*A* and supplementary fig. S13, Supplementary Material online). In contrast, when upregulation between tissues rather than differential expression among brain regions was taken as the response variable, the number of adjacent ncGOFs was associated with a higher probability of upregulation in the brain (fig. 5*B and C* and supplementary fig. S14, Supplementary Material online). As an example, 30% of the 427 genes that neighbor at least one ncGOF were classified as upregulated in cerebral cortex samples compared with testis samples, an increase of 2– 8 percentage points (95% CI) in the marginal probability of upregulation beyond what is expected for genes that do not neighbor an ncGOF.

LOFs and GOFs were defined based on intersection with phastCons elements from the 100-way vertebrate data set, whereas the CNSs considered here were generated from the 8-way vertebrate set to match the CNSs used to define ACNSs. However, we do not expect this to substantially influence the results because, although the sets differ in the number of elements, the nearest gene counts display a very similar pattern (Spearman correlation coefficient of 0.91). Indeed, when the OCNS predictor was generated from the 100-way set, ncGOFs showed a comparable positive

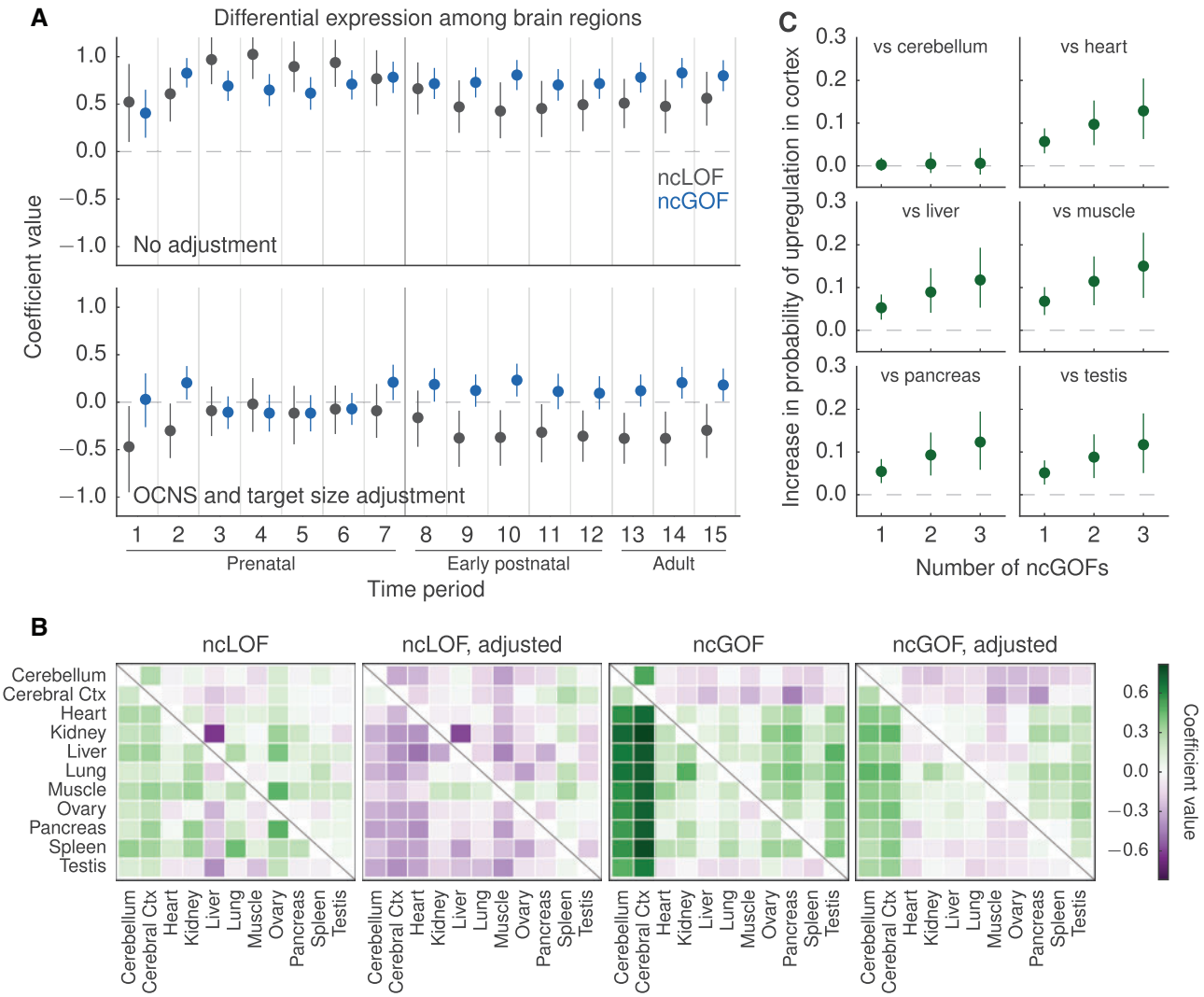

Fig. 5. Association of human-specific LOF and GOF candidates with differential expression. (*A*) Probit coefficients for the noncoding subset of human-specific LOF and GOF candidates identified by Schrider and Kern (2015). For each time period, coefficients for ncLOF and ncGOF count predictors (log-transformed) were generated from a model with differential expression status among brain regions as the response variable. In the bottom panel, the OCNS count and target size predictors were included. (*B*) Probit coefficient values of ncGOF and ncLOF predictors for inter-tissue expression comparisons. The pairwise upregulation vectors were generated with GTEx samples, as in figure 4. (*C*) Increase in the probability of upregulation in the cerebral cortex compared with another selected tissue, given the number of ncGOFs a gene neighbors. The adjusted ncGOF coefficients were used to calculate the probabilities, with other predictors set to zero. Intervals in *A* and *C* represent 95% CIs.

correlation with upregulation in the brain (supplementary fig. S15, Supplementary Material online).

In contrast to the association of ncGOFs with upregulation in the brain, analysis of H3K4me1 and H3K4me3 signals did not indicate that the regulatory activity of ncGOFs is more pronounced in the brain. Although H3K4me1 and H3K4me3 signals were found to be consistently elevated in ncGOFs compared with random noncoding sequences, neural tissues do not seem to differ from other tissues in this regard, and a similar pattern was observed for ncLOFs (supplementary table S6 and supplementary fig. S16, Supplementary Material online). It is also worth noting that in the majority of tissues examined, ncLOFs displayed a higher H3K4me1 signal than ncGOFs did.

Schrider and Kern (2015) analyzed genes containing and neighboring GOFs using GREAT (McLean et al. 2010), a tool which adjusts for the bias introduced by locus length (Taher and Ovcharenko 2009), and found that this set was enriched for GO terms related to neurotransmission, particularly GABAergic signaling. Given that upregulation in the brain was the expression pattern for which we observed the highest coefficients for the ncGOF count predictor, we expected to recover the GABA-related enrichment if the analysis were limited to brain-upregulated genes rather than all genes. To verify this, we restricted the genes to the set of genes that were upregulated in the cerebral cortex compared with non-neural tissues, and, from this set, tested for biological processes that were overrepresented in the set of genes neighboring ncGOFs. As a crude method of accounting for the number of adjacent OCNSs and locus length, we ran parallel analyses with either genes nearest to at least ten OCNSs or genes nearest to at least ten units of the target size predictor. Terms enriched in these sets were removed

from the terms obtained for ncGOF-neighboring genes. Finally, the analyses were run separately for each adult time period, and a GO term was taken as a hit only if it occurred in at least two of the three periods. Following this procedure, we found that genes neighboring an ncGOF are enriched for processes related to adherens junction organization (GO:0034332, mean *P* value 0.001), synaptic transmission (GO:0007268, mean *P* value 0.001), cell junction assembly (GO:0034329, mean *P* value 0.005), and GABAergic synaptic transmission (GO:0051932, mean *P* value 0.007). When the same procedure was performed for ncLOFs, no significant terms were found that occurred in at least two of the time periods.

Discussion

The analyses here were motivated by previous reports that genes which are DEX among human mid-fetal brain regions are more likely than non-DEX genes to neighbor ACNSs. After adjusting for the number of adjacent OCNSs, we do not find evidence that genes nearest to HACNSs have a higher tendency to be DEX. The results from the previous studies can be explained by the following relationship: when the predictor for being the nearest gene to OCNSs is reduced to a binary indicator, being the nearest gene to a HACNS indicates that a gene is also likely nearest to many OCNSs. This relationship extends to ACNSs in general and provides an alternative explanation for the observation that genes nearest to a chimpanzee- or mouse-accelerated sequence are also overrepresented among genes that are DEX within the developing and adult human brain.

The above explanation consists of two claims: 1) that previous estimates of the association between accelerated sequences and differential expression are largely capturing the association with CNSs in general, and 2) that ACNSs are not associated with differential expression. The findings here provide weaker support for the second claim than for the first claim for the following reasons. First, the coefficient estimates for the ACNS predictors are relatively uncertain compared with the estimates for the OCNS predictor, with some upper bounds extending to large positive values. Second, while we examined samples from various time points and tissues, ACNSs (or a particular set of ACNSs) may show an association in some other context.

We have used multiple data sets to demonstrate that a positive relationship exists between the number of adjacent CNSs and a gene being DEX among brain regions. This association persists across time periods, and pairwise comparisons of a subset of adult tissues suggest that the strength of the association between the number of adjacent OCNSs and upregulation in the brain is at the high end of what is observed for pairwise tissue comparisons in general. Our results complement a recent analysis which categorized genes into one of four bins based on the number of adjacent CNSs and found that genes nearest to a greater number of CNSs have more conserved expression between human and mouse (Babarinde and Saitou 2016). Together, these findings suggest that accounting for the number of CNSs that a gene neighbors may be necessary to detect a positive correlation between ACNSs and interspecies differential expression. If this factor is not considered, any association between ACNSs and divergent expression between species would likely be masked by the strong association of CNSs with conserved expression.

Finally, we show that genes neighboring a recently identified set of human-specific GOF regions are overrepresented for genes that are upregulated in the human brain compared with non-neural tissues, an association that survives the adjustment for the number of adjacent OCNSs. This observation lends support to the hypothesis that this group of elements is involved in regulating genes with functions in the nervous system (Schrider and Kern 2015). Inter-species gene expression data sets will be useful for determining whether genes nearby these elements do indeed exhibit human-specific expression patterns. More generally, the approach described here may be useful for analyzing other sets of noncoding sequences, including conserved sequences with different levels of phylogenetic conservation, such as Hominidae-specific CNSs (Saber et al. 2016).

What are the main factors contributing to the positive correlation of the CNS count predictor and differential expression within the brain? In modeling this relationship, we have made the common simplification of assigning the nearest gene as the target of a CNS. However, we expect the cases in which this classification is false to weaken, rather than strengthen, the association. Beyond the issue of identifying the target gene or genes, the gene expression data were from tissue samples with heterogeneous cell populations, which undoubtedly masks some cell-type-specific differential expression. Furthermore, the results here are observational and do not provide direct evidence that any of these CNSs are regulating expression within the brain. In addition to potential technical confounding variables, the number of adjacent CNSs may vary with some unspecified genomic feature that is contributing to the signal. In fact, this may be CNSs with different depths of conservation. Because the nearest gene count predictor is a gene-level aggregate predictor, counts generated from sequences that are conserved across distantly related species may correlate well with counts that are generated from sequences that are only conserved across more closely related species, in which case sequences that are not as deeply conserved may contribute to the signal.

Previous findings on CNSs, however, provide support for a regulatory role. Although studies have used various methods, groups of species, and thresholds to identify CNSs, the different sets likely have a large degree of overlap in terms of functional characteristics. For example, ultraconserved elements do not appear functionally distinct from the larger blocks of CNSs within which they often reside (Visel et al. 2008). In general, genes neighboring CNSs, especially clusters of CNSs, show enrichment for functions related to development and transcriptional regulation (Bejerano et al. 2004; Sandelin et al. 2004; Woolfe et al. 2004; Plessy et al. 2005; Pennacchio et al. 2006; Babarinde and Saitou 2013), although at least part of the development-related enrichment is expected due to a sampling bias related to locus length (Taher and Ovcharenko 2009; McLean et al. 2010). Many studies have provided functional evidence for the enhancer activity

of specific CNSs by testing them with reporter gene assays in zebrafish (e.g., Woolfe et al. 2004; Blader et al. 2004) or mice (e.g., Pennacchio et al. 2006; Zerucha et al. 2000; Nobrega et al. 2003). A large proportion of characterized CNSs appear to regulate expression in the nervous system (Pennacchio et al. 2006), including an element critical for controlling the identity and connectivity of corticospinal neurons (Shim et al. 2012). The density of nearby CNSs seems to be a good indicator of whether a gene is developmentally regulated (Prabhakar et al. 2006b), as well as whether a gene is a target rather than bystander gene (Kikuta et al. 2007). Based on gene expression in human tissue samples, a bias for expression in the nervous system has been reported for genes neighboring CNSs (Babarinde and Saitou 2016), including a subset of paralogous CNSs (Matsunami and Saitou 2013) and ultraconserved elements (Ovcharenko 2008). Beyond correlating differential gene expression with adjacent CNSs, ChIP-seq experiments have enabled a more direct investigation of the interplay between sequence conservation and putative regulatory elements that are active in a tissue of interest (Nord et al. 2013; Wenger et al. 2013; Babarinde and Saitou 2016; Emera et al. 2016).

The interpretation of findings on CNSs depends on the question of what underlies their conservation over large evolutionary timescales (Boffelli et al. 2004; Harmston et al. 2013). In some cases, conservation at the sequence level seems decoupled from conserved regulatory function. There are several examples of CNSs that possess different functional activity in different species (Blader et al. 2004; Nelson and Wardle 2013). CNSs also may acquire additional functions (Hiller et al. 2012; Wenger et al. 2013), perhaps influenced by flanking, non-conserved sequences (McEwen et al. 2009; Goode et al. 2011). Moreover, a high level of conservation for noncoding regulatory elements is unexpected given the degeneracy of transcription factor-binding sites and flexibility of inter-module spacing, which can lead to conserved regulatory function despite a lack of sequence conservation (Fisher et al. 2006; Hare et al. 2008; Kalay and Wittkopp 2010; Villar et al. 2014). Several potential mechanisms have been proposed to explain the observed degree of sequence conservation, including a dense overlap of transcription factor-binding sites (Poulin et al. 2005), the importance of these sequences in early developmental stages (Nelson and Wardle 2013), their multi-functional nature (McEwen et al. 2009), and their interaction with each other (Robyr et al. 2011; Dimitrieva and Bucher 2012), but it is not clear that any of these are sufficient to explain why CNSs exhibit such a high level of conservation (Harmston et al. 2013).

Regardless of the factors that are responsible for the conservation of CNSs or for the positive relationship between CNSs and differential expression, our results demonstrate the importance of considering the number of adjacent CNSs when quantifying the relationship between noncoding features and differential expression. More generally, covariates that incorporate information about the number of CNSs surrounding a gene may deserve greater attention when studying gene-level variables other than differential expression, such as the classification of genes as associated with a particular biological process or disease.

Materials and Methods

Code and build scripts for all analyses, including the downloading and preparation of the data sets, are available in a Git repository at https://gitlab.com/kmeyer/cns-count-analyses. In addition to specific tools referenced below, these analyses relied on the R language (R Core Team 2016), Snakemake (Köster and Rahmann 2012), and many components of the SciPy stack, including Matplotlib (Hunter 2007).

Gene Expression Data

Gene expression levels were obtained from a microarray study of brain regions throughout human development (supplementary tables S1 and S2, Supplementary Material online) (Kang et al. 2011). The total data set consisted of 1,331 samples. Genes were filtered to protein-coding genes known to Gencode 19. Normalized gene expression values were also downloaded for the Johnson et al. (2009) and Lambert et al. (2011) studies.

RNA-seq data for tissues from the GTEx project (The GTEx Consortium 2015) were downloaded from the consortium's website (http://www.gtexportal.org; last accessed October 23, 2015). Analyses considered samples from 11 tissues: cerebellum, cerebral cortex, heart (left ventricle), kidney (cortex), liver, lung, skeletal muscle, ovary, pancreas, spleen, and testis samples. For comparison, each sample was classified as belonging to one of the three adult stages from the Kang et al. data set (supplementary table S2, Supplementary Material online), and the genes analyzed were restricted to those present in the microarray used in the Kang et al. study.

Identification of Candidate Regulatory Element Sets

The locations of HACNSs, CACNSs, and MACNSs were retrieved from the supporting online material of the Prabhakar et al. (2006a) study. The set of CNSs was generated according to the reported filtering criteria of the original analysis. Specifically, an element in the eight-way vertebrate phastCons data set (retrieved from http://genome.ucsc.edu; last accessed April 6, 2015) was retained if it had a conservation score $\geq$ 400 and if it did not overlap with human mRNAs, human spliced ESTs, retroposed genes, or duplicated blocks. Note that the CNS set in the original analysis was generated with additional filtering steps based on non-human constraint and statistical power. We used the set of HARs generated by Lindblad-Toh et al. (2011) and filtered the coordinates to those that did not overlap with exons. All coordinates were converted to hg19 coordinates using UCSC Genome Browser's LiftOver executable.

Human-specific LOF and GOF sets (Schrider and Kern 2015) were downloaded from the popCons data repository (http://www.github/kern-lab/popCons; last accessed April 14, 2016). Coordinates that overlapped with exons were removed. An OCNS set was generated that did not contain any LOF or GOF coordinates. A second set of OCNSs was also generated from the 100-way vertebrate phastCons elements (retrieved from http://genome.ucsc.edu; last accessed June 17, 2016), as phastCons elements from this species set,

rather than the 8-way set, were used in the original filtering of LOF and GOF candidates.

Determination of the Nearest Genes to CNSs

To find the nearest gene for each element, the coordinates were intersected with the longest transcripts of protein-coding genes from Gencode 19 using BEDTools (Quinlan and Hall 2010). If an element's coordinates were found within the start and end coordinates of a transcript, the corresponding gene was counted as a nearest gene. Otherwise, the gene with the minimum distance to an element, based on either bound of its largest transcript, was taken as the nearest gene. These nearest gene assignments were then used to tally the total number of times that each gene was the nearest gene to any element from a given set.

Classification of Genes as DEX

Before classifying genes in the Kang et al. data set as DEX, genes were filtered to those that had an average detection above background *P* value across all samples of 0.01 or lower. After filtering, two different linear models were constructed using the limma package (Smyth 2004): one where the neocortical areas were taken as a single region, resulting in 6 brain regions, and another where only the 11 neocortical areas were considered. With both these model structures, each brain region or area was nested within its respective time period. These models also included covariates for the sample individual, treated as a random effect, and the sample RNA integrity number (RIN). Pairwise contrasts were formed for all region factors within that period. To be classified as DEX among brain regions, a gene was required to have a $\log_2$-fold change above 1, tested in limma using the TREAT method (McCarthy and Smyth 2009), and an FDR-adjusted *P* value at or below 0.01 for at least one contrast. A similar procedure was used to classify genes in the Johnson et al. data set as DEX between regions, but all samples were taken as belonging to a single time period. For the Lambert et al. data set, which consisted of two brain regions from two individuals, region and individual were used as covariates, with the latter treated as a random effect.

As an alternative method, an ANOVA model was constructed that considered period 6 samples and included a factor for either 6 brain regions or 11 neocortical areas, with sample RIN as a covariate. Following the criteria of Kang et al. (2011), a gene was called DEX if it had an FDR-adjusted *P* value below 0.01, at least one sample with a $\log_2$-transformed signal intensity above 6, and an average $\log_2$-fold change above 1 between at least two regions.

To classify genes as DEX between tissues in the GTEx data set, genes were first filtered to include only those that had a minimum count of ten in at least three samples. The expression counts were transformed with the voom package (Law et al. 2014) for modeling with limma. The sequencing batch, individual, and RIN were included as covariates, with the individual taken as a random effect. Pairwise contrasts were made between each tissue.

Esthe Probability of Regional Differential Expression

Probit regression was used to model the relationship between the number of elements a gene neighbors and differential expression. Let y be an N-length binary vector in which y_i indicates the differential expression status of the ith gene. The value of the latent variable z determines the value of y:

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, \ldots, N.$$

z follows a normal distribution with a standard deviation of one,

$$z_i \sim \mathrm{N}(\beta_0 + X_i\beta, 1),$$

where X is an $N \times K$ predictor matrix that includes a column for each element type (e.g., OCNS, HACNS, CACNS, and MACNS) indicating the number of elements to which a gene was the nearest. For each gene i and CNS predictor k, the count predictor was transformed as $\ln(1 + X_{ik})$. This transformation compresses the large upper range of the OCNS counts and reduced the correlation between the binned residual value, calculated with R package arm (Gelman and Su 2014), and the OCNS predictor value (supplementary fig. S17, Supplementary Material online). We also examined several other transformations of CNS predictors (supplementary table S4, Supplementary Material online).

The prior for the intercept term as well as all predictor coefficients were modeled as normally distributed and centered at zero with a standard deviation of three. The model parameters were estimated with Markov chain Monte Carlo (MCMC) using the Stan modeling language (Stan Development Team 2015). The autocorrelation of parameter values across iterations, as well as the $\hat{R}$ statistic across several chains, were examined for each run to verify that there was no indication of poor mixing. In addition to estimating parameters of the above model with MCMC, the parameters were estimated with maximal likelihood where no prior information was encoded. Both methods gave similar coefficient estimates (supplementary fig. S18, Supplementary Material online), suggesting that the results are not sensitive to the chosen priors.

The probability of differential expression was generated using the posterior samples of the regression coefficients. The linear predictor

$$\eta^{\mathrm{OCNS}} = \beta_0 + x\beta^{\mathrm{OCNS}}$$

was calculated for different values of x that were evenly spaced from zero to the highest observed OCNS count predictor value where all ACNS count predictors were zero. η^{OCNS} was transformed to probability space as $p^{\mathrm{OCNS}} = \Phi(\eta^{\mathrm{OCNS}})$, where Φ is the normal cumulative distribution function. The marginal probability difference for a given value of x was then calculated as $\Phi(\eta^{\mathrm{OCNS}}) - \Phi(\beta_0)$. Marginal probability differences reported for other predictors were calculated similarly.

As an extension of the GLM described earlier, a generalized additive model (GAM) was also fit using the R package mgcv (Wood 2011). The GAM yielded results that were consistent with the GLM findings (supplementary fig. S19, Supplementary Material online). GAM estimates were also

used to check that the relationship between the OCNS predictor and differential expression was not primarily driven by the expression level of a gene. When the median expression of each gene (standardized to be zero-centered and have a standard deviation of one) was included in the GAM, the estimated relationship between the OCNS predictor and differential expression remained similar (supplementary fig. S20, Supplementary Material online).

Adjusting for Gene Target Size

We used one of two predictors to adjust for target size variation across genes. The first method was to include the locus length, as defined by Taher and Ovcharenko (2009), as a predictor. After restricting the genes to non-overlapping genes, the locus length for a gene was calculated by extending the bounds of each gene halfway across the intergenic space, where intergenic space was defined as the bases between the longest transcripts of neighboring genes. The log-transformed number of base pairs with these extended bounds was used as the predictor.

The second method was to include a predictor that was generated from mapping random noncoding sequences. We selected a list of random coordinates in the genome so that, after filtering on the same set of features used to generate CNSs, the list was of comparable size. With these elements, nearest gene counts were tallied using the method described earlier. This process was repeated 15 times, resulting in 15 vectors of nearest gene counts. The predictor was formed by taking the median of log-transformed counts for each gene. When reporting the coefficient of this predictor, it was unstandardized so that the value was comparable to coefficients for the OCNS predictor. Otherwise, the predictor was transformed to be centered and have a standard deviation of one.

Estimating the Probability of Upregulation for Pairwise Tissue Comparisons

For pairwise tissue comparisons, a single-response probit regression was set up with similar predictors as described earlier. A response variable was generated for each comparison that indicated whether a gene was upregulated in one tissue compared with another. A separate regression was run for each response, with coefficients estimated by maximum likelihood.

ChIP-Seq Data Sets

H3K4me1 and H3K4me3 ChIP-seq data for 26 tissues were downloaded from the NIH Roadmap Epigenomics Mapping Consortium's web portal (http://egg2.wustl.edu/roadmap; last accessed November 17, 2016). Using the consolidated epigenomes, the average signal fold-changes for coordinates were calculated with UCSC Genome Browser's bigWigAverageOverBed executable.

GO Analyses of ncGOFs

GO analyses were performed with topGO (Alexa et al. 2006). The statistical test and algorithm were left at their default values of Fisher's exact test and "weight01", respectively, and the minimum number of nodes was set to ten. Terms with a *P* value below 0.01 were considered enriched. The *P* values for each GO term were not adjusted for multiple testing because the algorithm takes into account the graph structure, resulting in *P* value calculations that are not independent across terms (Alexa et al. 2006).

The gene universe was restricted to genes that were called as upregulated in the cerebral cortex in pairwise comparisons with non-neural GTEx tissues. Enrichment analyses were performed independently for three groups of GTEx samples that were created by assigning samples to one of the three adult periods of the Kang et al. data set. The same cerebral cortex-upregulated genes were tested with respect to one of three classifications: whether each gene neighbors 1) at least one ncGOF, 2) at least ten OCNSs, or 3) at least ten elements of the target size predictor. Results of these three sets were intersected for each time period to identify terms that were enriched in ncGOF-neighboring genes but not in genes from the other two categories.

Supplementary Material

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

Acknowledgments

We thank Belen Lorente-Galdos, Gabriel Santpere, André Sousa, and Andrew Tebbenkamp for helpful discussions and critical comments on earlier versions of the article. We are grateful to the Yale High Performance Computing Center for providing resources used to run some of the analyses. This work was supported by the National Science Foundation Graduate Research Fellowship Program (DGE-1122492 to K.A.M); by MINECO grants BFU2014-55090-P (FEDER), BFU2015-7116-ERC, and BFU2015-6215-ERC to T.M.B; and by the National Institutes of Health (MH103339, MH110926, and MH106934 to N.S, MH106874 to T.M.B. and N.S.).

References

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22:1600–1607.

Babarinde IA, Saitou N. 2013. Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. *Genome Biol Evol*. 5:2330–2343.

Babarinde IA, Saitou N. 2016. Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics. *Mol Biol Evol*. 33:1807–1817.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.

Bird C, Stranger B, Liu M, Thomas D, Ingle C, Beazley C, Miller W, Hurles M, Dermitzakis E. 2007. Fast-evolving noncoding sequences in the human genome. *Genome Biol* 8:R118.

Blader P, Lam CS, Rastegar S, Scardigli R, Nicod J-C, Simplicio N, Plessy C, Fischer N, Schuurmans C, Guillemot F, et al. 2004. Conserved and acquired features of *neurogenin1* regulation. *Development* 131:5627–5637.

Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nat Rev Genet*. 5:456–465.

Boyd JL, Skove SL, Rouanet JP, Pilaz L-J, Bepler T, Gordân R, Wray GA, Silver DL. 2015. Human-chimpanzee differences in a *FZD8* enhancer

alter cell-cycle dynamics in the developing neocortex. *Curr Biol*. 25:772–779.
Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. 2013. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci*. 368:20130025.
Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134:25–36.
Dimitrieva S, Bucher P. 2012. Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. *Bioinformatics* 28:i395–i401.
Emera D, Yin J, Reilly SK, Gockley J, Noonan JP. 2016. Origin and evolution of developmental enhancers in the mammalian neocortex. *Proc Natl Acad Sci U S A*. 113:E2617–E2626.
Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* 312:276–279.
Gelman A, Su Y-S. 2014. arm: Data analysis using regression and multilevel/hierarchical models. Available from: http://CRAN.R-project.org/package=arm, last accessed September 9, 2015.
Goode DK, Callaway HA, Cerda GA, Lewis KE, Elgar G. 2011. Minor change, major difference: Divergent functions of highly conserved *cis*-regulatory elements subsequent to whole genome duplication events. *Development* 138:879–884.
Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid *even-skipped* enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genet*. 4:e1000106.
Harmston N, Barešić A, Lenhard B. 2013. The mystery of extreme non-coding conservation. *Philos Trans R Soc Lond B Biol Sci*. 368:20130021.
Haygood R, Babbitt CC, Fedrigo O, Wray GA. 2010. Contrasts between adaptive coding and noncoding changes during human evolution. *Proc Natl Acad Sci U S A*. 107:7853–7857.
Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet*. 39:1140–1144.
Hiller M, Schaar BT, Bejerano G. 2012. Hundreds of conserved non-coding genomic regions are independently lost in mammals. *Nucleic Acids Res*. 40:11463–11476.
Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 9:90–95.
Indjeian VB, Kingman GA, Jones FC, Guenther CA, Grimwood J, Schmutz J, Myers RM, Kingsley DM. 2016. Evolving new skeletal traits by *cis*-regulatory changes in bone morphogenetic proteins. *Cell* 164:45–56.
Johnson MB, Kawasawa YI, Mason CE, Krsnik Ž, Coppola G, Bogdanović D, Geschwind DH, Mane SM, State MW, Šestan N. 2009. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* 62:494–509.
Kalay G, Wittkopp PJ. 2010. Nomadic enhancers: Tissue-specific *cis*-regulatory elements of *yellow* have divergent genomic positions among *Drosophila* species. *PLoS Genet*. 6:e1001222.
Kamm GB, López-Leal R, Lorenzo JR, Franchini LF. 2013a. A fast-evolving human NPAS3 enhancer gained reporter expression in the developing forebrain of transgenic mice. *Philos Trans R Soc Lond B Biol Sci*. 368:20130019.
Kamm GB, Pisciottano F, Kliger R, Franchini LF. 2013b. The developmental brain gene *NPAS3* contains the largest number of accelerated regulatory sequences in the human genome. *Mol Biol Evol*. 30:1088–1102.
Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, et al. 2011. Spatio-temporal transcriptome of the human brain. *Nature* 478:483–489.
Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, et al. 2007. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res*. 17:545–555.
Kim SY, Pritchard JK. 2007. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet*. 3:e147.
Köster J, Rahmann S. 2012. Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics* 28:2520–2522.
Lambert N, Lambot M-A, Bilheu A, Albert V, Englert Y, Libert F, Noel J-C, Sotiriou C, Holloway AK, Pollard KS, et al. 2011. Genes expressed in specific areas of the human fetal cerebral cortex display distinct patterns of evolution. *PloS One* 6:e17753.
Law C, Chen Y, Shi W, Smyth G. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 15:R29.
Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478:476–482.
Linnen CR, Poh Y-P, Peterson BK, Barrett RDH, Larson JG, Jensen JD, Hoekstra HE. 2013. Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* 339:1312–1316.
Matsunami M, Saitou N. 2013. Vertebrate paralogous conserved non-coding sequences may be related to gene expressions in brain. *Genome Biol. Evol*. 5:140–150.
McCarthy DJ, Smyth GK. 2009. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* 25:765–771.
McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Elgar G. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet*. 5:e1000762.
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol*. 28:495–501.
Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, et al. 2014. Transcriptional landscape of the prenatal human brain. *Nature* 508:199–206.
Nelson AC, Wardle FC. 2013. Conserved non-coding elements and *cis* regulation: Actions speak louder than words. *Development* 140:1385–1395.
Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* 302:413–413.
Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V, et al. 2013. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* 155:1521–1531.
Ovcharenko I. 2008. Widespread ultraconservation divergence in primates. *Mol Biol Evol*. 25:1668–1676.
Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* 444:499–502.
Plessy C, Dickmeis T, Chalmel F, Strähle U. 2005. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends Genet*. 21:207–210.
Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, et al. 2006a. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet*. 2:e168.
Pollard KS, Salama SR, Lambert N, Lambot M-A, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–172.
Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA. 2005. *In vivo* characterization of a vertebrate ultraconserved enhancer. *Genomics* 85:774–781.
Prabhakar S, Noonan JP, Pääbo S, Rubin EM. 2006a. Accelerated evolution of conserved noncoding sequences in humans. *Science* 314:786–786.
Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA. 2006b. Close sequence comparisons are sufficient to identify human *cis*-regulatory elements. *Genome Res* 16:855–863.
Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, FitzPatrick DR, Afzal V, et al. 2008. Human-specific gain of function in a developmental enhancer. *Science* 321:1346–1350.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/, last accessed November 6, 2016.

Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP. 2015. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* 347: 1155–1159.

Robyr D, Friedli M, Gehrig C, Arcangeli M, Marin M, Guipponi M, Farinelli L, Barde I, Verp S, Trono D, et al. 2011. Chromosome conformation capture uncovers potential genome-wide interactions between human conserved non-coding sequences. *PLoS One* 6:e17634.

Saber MM, Adeyemi Babarinde I, Hettiarachchi N, Saitou N. 2016. Emergence and evolution of Hominidae-specific coding and non-coding genomic sequences. *Genome Biol Evol*. 8:2076–2092.

Sandelin A, Bailey P, Bruce S, Engström P, Klos J, Wasserman W, Ericson J, Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* 5:99.

Schrider DR, Kern AD. 2015. Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain. *Genome Biol Evol*. 7:3511–3528.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428:717–723.

Shibata M, Gulden FO, Sestan N. 2015. From trans to cis: Transcriptional regulatory networks in neocortical development. *Trends Genet*. 31:77–87.

Shim S, Kwan KY, Li M, Lefebvre V, Sestan N. 2012. *Cis*-regulatory control of corticospinal system development and evolution. *Nature* 486:74–79.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15:1034–1050.

Silbereis JC, Pochareddy S, Zhu Y, Li M, Sestan N. 2016. The cellular and molecular landscapes of the developing human central nervous system. *Neuron* 89:248–268.

Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 3:1–25.

Stan Development Team. 2015. Stan: A C ++ library for probability and sampling. Available from: http://mc-stan.org/, last accessed March 18, 2015.

Sumiyama K, Saitou N. 2011. Loss-of-function mutation in a repressor module of human-specifically activated enhancer HACNS1. *Mol Biol Evol*. 28:3005–3007.

Taher L, Ovcharenko I. 2009. Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics* 25:578–584.

The GTEx Consortium. 2015. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348:648–660.

Vermunt MW, Tan SC, Castelijns B, Geeven G, Reinink P, de Bruijn E, Kondova I, Persengiev S, Netherlands Brain Bank, Bontrop R, et al. 2016. Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat Neurosci*. 19:494–503.

Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans—Mechanisms and functional implications. *Nat Rev Genet*. 15:221–233.

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet*. 40:158–160.

Wenger AM, Clarke SL, Notwell JH, Chung T, Tuteja G, Guturu H, Schaar BT, Bejerano G. 2013. The enhancer landscape during early neocortical development reveals patterns of dense regulation and co-option. *PLoS Genet*. 9:e1003728.

Wood SN. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc B*. 73:3–36.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2004. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*. 3:e7.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet*. 8:206–216.

Zerucha T, Stühmer T, Hatch G, Park BK, Long Q, Yu G, Gambarotta A, Schultz JR, Rubenstein JLR, Ekker M. 2000. A highly conserved enhancer in the *Dlx5/Dlx6* intergenic region is the site of cross-regulatory interactions between *Dlx* genes in the embryonic forebrain. *J Neurosci*. 20:709–721.