Published in final edited form as:
*Neuron*. 2022 January 05; 110(1): 70–85.e6. doi:10.1016/j.neuron.2021.10.014.

# An expansion of the non-coding genome and its regulatory potential underlies vertebrate neuronal diversity

**Michael Closser**[1], **Yuchun Guo**[2], **Ping Wang**[3], **Tulsi Patel**[1], **Sumin Jang**[1], **Jennifer Hammelman**[2], **Joriene De Nooij**[4], **Rachel Kopunova**[1], **Esteban O. Mazzoni**[5], **Yijun Ruan**[3], **David K. Gifford**[2,*], **Hynek Wichterle**[1,**]

[1]Departments of Pathology and Cell Biology, Neuroscience, and Neurology, Center for Motor Neuron Biology and Disease, Columbia Stem Cell Initiative, Columbia University Irving Medical Center, New York, NY 10032, USA.

[2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA

[3]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030, USA.

[4]Department of Neurology, Center for Motor Neuron Biology and Disease, Columbia Stem Cell Initiative, Columbia University Irving Medical Center, New York, NY 10032, USA

[5]Department of Biology, New York University, New York, NY 10003, USA.

## Summary

The proper assembly and function of the nervous system requires the generation of a uniquely diverse population of neurons expressing cell-type-specific combination of effector genes that collectively define neuronal morphology, connectivity, and function. How countless, partially overlapping, yet cell-type-specific patterns of gene expression are controlled at the genomic level remains poorly understood. Here we show that neuronal genes are associated with highly complex gene regulatory systems composed of independent cell-type- and cell-stage-specific regulatory elements that reside in expanded non-coding genomic domains. Mapping enhancer-promoter interactions revealed that motor neuron enhancers are broadly distributed across the large chromatin domains. This distributed regulatory system is not a unique property of motor neurons, but is broadly employed throughout the nervous system. The number of regulatory elements dramatically increased during the transition from invertebrates to vertebrates, suggesting that acquisition of new enhancers might be a fundamental process underlying the evolutionary increase in cellular complexity.

[*]Corresponding Author David K. Gifford, gifford@mit.edu, 617-253-6039. [**]Lead Contact Hynek Wichterle, hw350@columbia.edu, 212-342-3929.

AUTHOR CONTRIBUTIONS

M.C. and H.W. conceived the study and wrote the manuscript. M.C., P.W., T.P., J.D.N., E.O.M. and R.K. performed experiments., M.C., Y.G, S.J. and J.H. analyzed data. H.W, D.G and Y.R obtained funding and oversaw study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

## eTOC Blurb

Closser et al. show that neuronal genes are associated with highly complex regulatory systems in expanded non-coding genomic domains. Neuronal enhancers are broadly distributed and sparsely utilized in cell-type- and cell-stage-specific patterns. The acquisition of new enhancers might be a fundamental process underlying the evolutionary increase in cellular complexity.

## Introduction

At the most fundamental level, unique cell identities in multicellular organisms are defined by the expression of specific combinations of effector genes (e.g. receptors, cytoskeletal proteins, adhesion molecules, ion channels, neurotransmitters) that endow individual cells with distinct morphological, physiological, and biochemical properties. Considering that the number of genes has not significantly increased during the evolution of multicellular organisms, the emergence of new cell types has to be accompanied by the evolution of more elaborate gene regulatory systems that facilitate implementation of novel cell type specific gene expression programs (King and Wilson, 1975). The complexity of gene regulation is especially daunting in the vertebrate central nervous system (CNS), where upwards of thousands of distinct neuronal subtypes express unique combinations of largely overlapping subsets of effector genes (Macosko et al., 2015; Saunders et al., 2018; Tasic et al., 2016). Genetic studies in invertebrates established that gene expression is typically controlled by specialized cell-type-specific regulatory elements (Flames and Hobert, 2009; Hobert et al., 2010; Konstantinides et al., 2018; Kratsios et al., 2011; Stefanakis et al., 2015). Similarly, vertebrate regulatory elements are active in restricted subsets of cells and/or at discrete developmental stages, indicative of their high degree of specialization (Lindtner et al., 2019; Nord et al., 2013; Rhee et al., 2016; Sandberg et al., 2016; Shim et al., 2012; Visel et al., 2013). Reliance on such a selective regulatory system could have a significant impact on genome organization in species with highly complex nervous systems (Nelson et al., 2004). Specifically, regulation of genes broadly expressed across multiple neuronal cell types in the vertebrate CNS would demand a commensurate increase in the number and complexity of cis-regulatory elements associated with these genes. How this anticipated increase in vertebrate regulatory information is encoded, organized, and utilized at the genomic level remains largely unknown.

The extreme cellular diversity of the nervous system complicates interpretation of bulk tissue genomic analysis, and single cell approaches currently lack resolution and robustness required for detailed chromatin interaction mapping. Here we took advantage of methods for efficient reprogramming of stem cells to motor neurons to generate, to our knowledge, the first detailed map of enhancer-promoter interactions in a defined neuronal cell type. We identified distal enhancers by analyzing the distribution of genomic sites associated with the mediator complex and increased levels of H3K27ac, and assigned actively engaged enhancers to their appropriate target genes using enhancer-promoter chromatin interaction analysis. We discovered that postmitotic neuronal genes are controlled by a complex regulatory system composed of enhancers distributed over genomic territories twice the size of regulatory systems mapped in embryonic stem cells and motor neuron progenitors. Loss- and gain-of-function studies established that enhancers distributed over the regulatory space converge on individual promoters, controlling target genes in a largely cell-type and cell-stage specific manner. The expanded regulatory system associated with genes expressed in postmitotic neurons is not limited to motor neurons, but is broadly utilized across the CNS, as confirmed by a meta-analysis of published DNA accessibility and chromatin interaction data. At the genomic level, the observed increase in enhancer-promoter interaction distance is manifested as expanded intergenic domains flanking genes expressed in neurons, compared to genes expressed in non-neuronal tissues. Phylogenic comparison of intergenic space across model organisms with well-annotated genomes revealed that expansion of regulatory space in the proximity of neuronal genes coincided with the transition from invertebrates to vertebrates. Together our studies suggest, that an increase in neuronal diversity demands a commensurate increase in the complexity of the regulatory landscape around neuronal genes, contributing to the evolutionary expansion of the non-coding genome.

## Results

### Expansion of complexity in gene regulatory domains surrounding neuronal genes

To elucidate principles underlying the control of cell type specific gene expression in the vertebrate nervous system, we first examined the genome-wide distribution of putative regulatory elements associated with the top 500 genes induced in eight embryonic and adult neuronal and nine adult non-neuronal mouse tissues and cell types. Cis-regulatory elements were mapped based on sequence conservation and ENCODE DNA accessibility data (Shen et al., 2012; Yue et al., 2014). Following proximity assignment, we observed 2-3 fold increase in the number of putative regulatory elements associated with genes expressed in the neuronal relative to non-neuronal tissues (Figure 1A and 1B). Next, we asked whether neuronal genes are also associated with larger domains of non-coding DNA to accommodate the increased regulatory complexity. We compared the sizes of intergenic domains surrounding neuronal genes and found that they are indeed 2-3 times larger than the median intergenic size (Figure 1C, Figure S1A). In contrast, genes expressed in non-neuronal tissues are associated with intergenic regions that do not significantly differ from the median size.

Most enhancer-promoter interactions are constrained within larger chromatin domains termed “topological domains” (TADs), which are thought to be largely stable between different cell types (Dixon et al., 2012). Given the expanded intergenic domains associated with neuronal genes, we asked whether neuronal genes reside in larger topological domains than non-neuronal genes. Interestingly, the overall size of topological domains associated with the different cell-type specific genes was not significantly different. Instead, we observed that neuronal genes are located in topological domains with much lower gene density compared to other cell types (Figure S1B and S1C). Within a single TAD, genes and their regulatory elements are posited to reside in sub-TAD structures known as insulated domains, defined by CTCF-Cohesin occupied boundary elements (Dowen et al., 2014; Phillips-Cremins et al., 2013). We next asked whether the size of putative insulated domains was expanded in motor neurons by mapping the distance between CTCF-Cohesin co-bound sites flanking individual genes. Consistent with the expansion in intergenic domains sizes, we found that motor neuron genes reside in significantly expanded insulated regulatory domains (MN genes = 218kb, ES genes = 102kb, All genes = 135kb, $p < 0.005$, Figure S1D).

We reasoned that the expansion in regulatory complexity surrounding neuronal genes could be required to implement the highly complex expression patterns that exist in the mammalian nervous system. If so, neuronal genes expressed in fewer cell types would be predicted to be associated with less complex regulatory regions than those expressed more broadly. To examine this possibility, we analyzed neuronal genes expressed selectively in cortical pyramidal excitatory neurons, parvalbumin inhibitory interneurons, sensory neurons, and motor neurons. We observed that the size of intergenic regions as well as the numbers of conserved and accessible sites was significantly larger in the proximity of neuronal effector genes expressed in all four cell types compared to cell type specific genes (Figure S1E-G, $p < 4.9e-7$, 2.3e-8, 2.1e-6, respectively). Together these data suggest that individual neuronal genes reside in large, gene poor, non-coding domains containing increased numbers of regulatory elements that are likely required for the establishment of complex gene expression patterns across diverse neuronal cell types.

## Cell-type- and cell-stage-specific enhancers populate expanded regulatory space associated with neuronal genes

Next, we asked whether the regulatory system associated with broadly expressed neuronal genes is utilized in a cell type-specific manner. In such case, only small subsets of the potential regulatory elements should be actively used in any individual neuronal cell type. To evaluate cell type specificity of regulatory elements, we compared chromatin accessibility profiled by an Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq, (Buenrostro et al., 2013)) in four distinct primary embryonic and adult neuronal populations. We profiled primary sensory neurons purified from embryonic mouse dorsal root ganglia (“SN” labeled by TrkC-tdTomato (Bai et al., 2015)) and motor neurons purified from embryonic mouse spinal cords (“MN” labeled by Hb9-GFP (Wichterle et al., 2002)) and compared them with available ATAC-seq profiles of adult cortical excitatory neurons (“EXC” labeled by Camk2a-cre) and a subtype of cortical interneurons (“PV” labeled by Pvalb-cre) (Mo et al., 2015). We examined accessible regions surrounding neuronal genes that are broadly expressed in these cell types, by identifying the top 500 genes which were induced at least 2 fold across all four neuronal cell types. This gene set includes neuronal splicing factors (Rbfox1, Elavl4), neurotransmitter receptors (Grik2), synaptic proteins (Syt1, Nrnx3), guidance molecules (Dcc, Gap43), neuronal cytoskeleton proteins (Ank2, Nefl) and cell adhesion molecules (Ncam1, Pcdh7) (Figure 1D and 1E and Figure S1H). Analysis of DNA accessibility revealed that the majority of accessible sites around these broadly expressed neuronal genes are highly cell type specific. Global mapping of accessible regions identified ~25,000 total significant peaks within intergenic regions associated with this gene set. Remarkably, less than 2% of the accessible sites are shared across all four neuronal types, with the remaining 98% either present in only one of the cell types (53%) or shared between a subset of cell types (45%) (Figure 1D, 1E and 1F, and Figure S1H).

## Selector transcription factors globally control neuronal identity

Although genome conservation, DNA accessibility, and gene proximity are broadly used to make inferences about enhancers and their target genes (Shen et al., 2012; Yue et al., 2014), not all putative regulatory elements actively engage with promoters to control gene expression and some regulatory elements skip the most proximal gene to regulate a more distal promoter (Li et al., 2012). To obtain a more realistic map of regulatory interactions controlling neuronal expression programs, we used chromatin interaction analysis to assign active enhancers to their target genes. To facilitate a high-resolution chromatin interaction mapping in a well-defined neuronal population, we took advantage of transcriptionally programmed murine stem cell-derived spinal motor neurons (Mazzoni et al., 2013) that can be produced in high purity and essentially unlimited quantities. We first mapped locations of cis-regulatory regions by performing chromatin immunoprecipitation (ChIP-seq) for histone acetylation (H3K27ac) and for mediator complex binding (Med1), two hallmarks of active enhancers (Heintzman et al., 2009; Kagey et al., 2010; Whyte et al., 2013). To assign active enhancers to appropriate target genes, we performed Chromatin Interaction Analysis with Paired End Tagging (ChIA-PET) of RNA polymerase II bound promoters (Fullwood et al., 2009). We then employed a computational method for chromatin interaction calling based on density clustering (Guo et al., 2019; Rodriguez and Laio, 2014) to map enhancer-promoter interactions, identifying 4,483 unique distal enhancers (> 5 kb from the nearest promoter) engaged in high-confidence interactions with 2,512 Pol II occupied promoters (FDR<.01; Figure 2A, 2B and S2A).

To focus our analysis on the most relevant set of enhancer-promoter interactions, we selected for further analysis the subset of distal enhancers interacting with genes significantly induced (FPKM > 1, log2 fold change > 4, p < .01, Figure S2B) upon motor neuron specification. We mapped 2,268 high-confidence enhancer-promoter interactions engaged with 856 genes induced in motor neurons. To validate the regulatory interactions discovered in cultured stem cell-derived motor neurons, we extended our analysis to primary motor neurons purified from embryonic day 10.5 Hb9-GFP mouse spinal cords. ATAC-seq analysis of primary motor neurons revealed that ~86% of the distal enhancers discovered in vitro exhibited increased accessibility in motor neurons in vivo, suggesting that the majority of the discovered sites are involved in the control of the spinal motor neuron gene expression program both in vitro and in vivo (Figure 2A, 2B and Figure S2C).

Lim homeodomain (Lim-HD) containing transcription factors Isl1 and Lhx3 are motor neuron selector genes, sufficient to impose motor neuron identity on neural progenitors (Lee and Pfaff, 2003). To determine whether these factors control motor neuron effector genes directly or whether they act through a set of secondary transcription factors, we performed de novo motif discovery in engaged distal enhancers. Lim-HD motifs were the most highly enriched motifs present in distal enhancers interacting with motor neuron genes (p = 4e-71, 63.5%) (Figure 2C). Analysis of Isl1 and Lhx3 binding by ChIP-seq confirmed that 53% (1197) of the distal enhancers interacting with motor neuron genes were occupied by Isl1/Lhx3 complexes (Figure 2A, 2B and 2D) and 65% (556) of the induced genes interacted with at least one Isl1/Lhx3 bound enhancer (Figure 2D). Importantly, nearly all Isl1/Lhx3 bound enhancers (~95%) that interact with genes induced in motor neurons also exhibited increased genomic accessibility in primary motor neurons, indicative of their role in normal motor neuron specification in vivo (Figure 2A, 2B, 2E and Figure S2D). Some transcription factors were shown to function as both activators and repressors of gene expression (Dasen et al., 2003; Sandberg et al., 2016). Isl1 and Lhx3 appear to function predominantly as transcriptional activators since 96% of differentially expressed genes interacting with Isl1/Lhx3 bound enhancer were induced in motor neurons and only 4% were downregulated (Figure 2D).

Isl1/Lhx3 bound enhancers interacted with genes encompassing broad aspects of neuronal identity, including genes that are restricted to motor neurons (Mnx1/Hb9), genes expressed in subsets of neurons (Isl1, Isl2, ChAT) or effector genes expressed broadly across the nervous system (Slit2, Robo2, Nefl/m) (Figure 2D, 2E and 2F). These observations suggest that the selector transcription factors control not only the motor neuron specific genes or gene modules associated with specific neuronal functions, but operate as global regulators of the motor neuron gene expression program. This was confirmed in a genome-wide analysis comparing expression patterns of Isl1/Lhx3 target genes in the motor neurons, sensory neurons, excitatory neurons and cortical inhibitory neurons (Fig 1). 71 % of motor neuron specific genes and 62% of the shared neuronal genes (including broadly expressed transcription factors and neuronal effector genes) were targeted by Isl1 and Lhx3 bound enhancers (Figure 2F and 2G). Together, these data demonstrate that Isl1 and Lhx3 selector transcription factors function as global regulators of motor neuron identity, directly interacting with the majority of effector genes induced in postmitotic motor neurons (Hobert, 2016; Kratsios et al., 2011; Mann and Carroll, 2002).

To determine whether the global regulation of gene expression by selector transcription factors extends beyond motor neurons, we performed motif enrichment analysis in motor neuron and sensory neuron specific putative regulatory regions around cell type specific genes and genes shared between the two cell types (Figure S3A and S3B). Sensory neuron-specific regions were enriched for Runx and Lim-HD binding motifs consistent with the role of Runx1, Runx3 and Isl1 in sensory neuron identity (Chen et al., 2006; Kramer et al., 2006; Sun et al., 2008), while motor neuron specific regions were most enriched for Lim-HD and Hox binding motifs consistent with the role of Isl1, Lhx3 and Hox transcription factors controlling motor neuron identity (Dasen et al., 2003; Dasen et al., 2005; Lee and Pfaff, 2003; Pfaff et al., 1996) (Figure S3C). Importantly, these cell type specific motifs were enriched not only in the vicinity of cell type specific genes, but also in the vicinity of shared genes expressed by both sensory and motor neurons (Figure S3C). These observations support the model in which each neuronal cell type will utilize distinct sets of selector transcription factors binding largely non-overlapping sets of cell-specific regulatory regions to control both neuron type specific and broadly expressed genes, analogous to recently suggested models of pan-neuronal and convergent effector gene regulation in *C. elegans* (Stefanakis et al., 2015) and *D. melanogaster* (Konstantinides et al., 2018).

## Distributed organization of neuronal enhancers spanning large non-coding genomic space

Having mapped chromatin interactions in motor neurons, we examined in detail the organization of the distal regulatory elements in relation to their target genes. Tissue-specific genes are often controlled by clustered “super-enhancers”, in which spatial proximity of individual transcription factor binding sites facilitates mediator recruitment and contributes to robust induction of target genes (Hnisz et al., 2013; Whyte et al., 2013). To determine how enhancers are organized in motor neurons and at what point during embryonic development the neuron specific organization of regulatory elements emerges, we compared ChIA-PET chromatin interactions in postmitotic motor neurons, motor neuron progenitors (pMN), and embryonic stem cells (mESC) (Kieffer-Kwon et al., 2013; Zhang et al., 2013) (Figure 3A). Analysis of genes induced in the individual cell types revealed that the enhancer-promoter interactions associated with motor neuron (MN) genes span nearly twice the genomic distance of interactions mapped in pMNs or mESCs (median distances of 157 kb vs 94 kb vs 86kb, p < 2.2e-16; Figure 3A, 3B and 3C) and are associated with longer intergenic regions populated by a greater number of conserved elements (Figure 3D and 3E and S4A). Importantly, the increase in interaction distances is specific to genes selectively induced in motor neurons as enhancer-promoter interactions controlling stably expressed genes span comparable distances in all three cell types (Figure 3B, p < 2.2e-16). Our examination also revealed that of the 447 motor neuron genes associated with more than one distal regulatory element, these enhancers are not clustered, but distributed over large genomic regions with a median span of ~112.2 kb. This contrasts the organization of enhancers around mESC and pMN genes that are clustered significantly closer together (median distance of pMN =~38.5kb, mESC = 39.4kb, p <. 2.2e-16, Figure 3A and 3C). For example, the neurofilament gene *Nefm* is associated with seven distal enhancer interactions in postmitotic motor neurons spanning ~415kb of genomic space. In contrast, the neural progenitor gene *Nestin* (*Nes*), encoding an intermediate filament protein, is associated with four distal enhancer elements in motor neuron progenitors spanning ~41kb and embryonic stem cell specific gene, *Nanog*, is associated with four distal enhancer elements spanning 101kb (Figure 3F). Thus, the expanded distributed regulatory architecture observed in postmitotic motor neurons emerges during the transition from neuronal progenitors to postmitotic neurons.

The unique organization of motor neuron enhancers, that is not apparent in their immediate progenitors, raised the question whether similar regulatory features are characteristic of other parts of the developing and mature CNS. To answer this question, we analyzed published HiC chromatin interaction datasets from the developing human cortical progenitor zone and the postmitotic cortical plate (de la Torre-Ubieta et al., 2018; Won et al., 2016). Analogous to our observations in mouse embryonic motor neurons, the enhancer-promoter distances and intergenic sizes associated with genes expressed in cortical plate neurons were significantly longer than distances associated with genes expressed in their progenitors (Figure S4B, p<.001; Figure 1C, 3D and S4A). Together, these independent experimental techniques (ChIA-PET and HiC) performed on different neuronal and non-neuronal samples support the conclusion that distributed cell type specific enhancers are broadly utilized to control postmitotic neuronal gene expression.

## Distributed enhancers cooperate to regulate target gene expression

To probe the functional significance of distributed enhancers spanning hundreds of kilobases, we manipulated distal enhancers associated with *Isl1*, a gene essential for motor neuron specification. The *Isl1* promoter interacts with five distal regulatory regions in stem cell derived motor neurons spanning approximately 1 Mb of genomic space. We identified three elements with the signature of active enhancers with high levels of Isl1/Lhx3 binding, H3K27ac and Mediator occupancy. We functionally tested two of these enhancers. As a positive control we targeted a known *Isl1* enhancer E+222 (~222 kb downstream of TSS, previously named CREST1 (Kim et al., 2015; Uemura et al., 2005)). In addition, we targeted a newly discovered distal enhancer E+622 (~622 kb downstream of TSS). Both of these enhancers are highly occupied by Isl1/Lhx3 and accessible in vivo (Figure 2A and 4A). Given that Isl1 function is essential for motor neuron differentiation (Pfaff et al., 1996) and that the Isl1 transcription factor regulates its own expression (Figure 2A) we first performed a series of loss of function experiments in the context of transcriptionally programmed motor neurons expressing a transgenic Isl1 from an independent doxycycline regulated genomic locus (Mazzoni et al., 2013; Mazzoni et al., 2011). This experimental paradigm allowed accurate dissection of *Isl1* enhancer function while compensating for the potential endogenous *Isl1* expression deficits (Figure S5A and S5B). CRISPR/Cas9-mediated deletion of either E+222 or E+622 in isolation resulted in a significant decrease in *Isl1* transcription in stem cell derived motor neurons (61 and 82%, respectively; Figure 4B) suggesting that each enhancer contributes independently to the robustness in *Isl1* gene expression (Dickel et al., 2018; Hay et al., 2016; Lindtner et al., 2019; Osterwalder et al., 2018). Contribution of both enhancers to *Isl1* expression was further verified by compound deletion of both E+222 and E+622, resulting in a ~95% reduction in endogenous *Isl1* mRNA (Figure 4B).

Next, we asked whether the two enhancers are active in the same cells or whether each enhancer regulates *Isl1* expression in a different subset of motor neurons. Control and enhancer mutant cell lines were directed to differentiate into motor neurons by retinoic acid and smoothened agonist treatment (Wichterle et al., 2002) and endogenous Isl1 protein levels were measured in individual motor neurons by immunocytochemistry. Consistent with the population gene expression data, we observed a decrease in Isl1 protein levels in enhancer mutant motor neurons (E+222:25%, E+622:32%, E+222/E+622 49%, Figure 4C, 4D and S5C). Importantly, this decrease was observed as a global reduction of immunofluorescence levels across the entire motor neuron population. Furthermore, this effect was not due to perturbation of motor neuron specification, as evidenced by comparable levels of motor neuron progenitor marker Olig2 expression across the control and mutant cell lines (Figure 4C). Thus, while motor neuron enhancers are distributed across hundreds of kilobases of genomic space, these initial loss of function experiments suggest that individual enhancers converge and cooperate at promoters to ensure robust expression of target genes.

Isl1/Lhx3 bound enhancers regulating the *Isl1* gene in motor neurons lack DNA accessibility in sensory neurons (Figure 5A and 5B), suggesting that these enhancers are cell type specific. To functionally test the specificity of the two enhancers, we cloned the E+222 and E+622 regions into a reporter plasmid upstream of a minimal promoter and a destabilized GFP (minP::GFP). The plasmids were co-electroporated with ubiquitously expressed pCAGGS-mCherry into the ventral and the dorsal neural tube of developing chick embryos to target motor neurons or sensory neurons, respectively. Two days later, embryos were dissected and the location of reporter expressing cells was examined on transverse sections.

Both reporters exhibited strong expression in ventral motor neurons, but neither enhancer activated reporter expression in sensory neurons located in dorsal root ganglia (Figure 5C). Consistent with the transient co-expression of Isl1/Lhx3 in nascent motor neurons and our previous description of dynamic reorganization of motor neuron enhancers upon Lhx3 downregulation (Rhee et al., 2016), the reporters were expressed most strongly in medially located newly born motor neurons and extinguished in more mature laterally located motor neurons that downregulated Lhx3 expression (Figure 5C). Interestingly, while E+222 driven reporter was expressed exclusively in nascent spinal motor neurons (Kim et al., 2015; Rhee et al., 2016; Uemura et al., 2005), enhancer E+622 also activated reporter expression in a population of dorsal Isl1 expressing interneurons, indicating that the regulatory region contains additional transcription factor binding sites utilized by these cells (Figure 5C).

The above experiments demonstrate the high degree of cell type specificity among regulatory elements associated with neuronal genes. However, the complexity of the neuronal regulatory system extends beyond cell type specificity to temporal changes in regulatory element utilization during neuronal maturation (Nord et al., 2013; Visel et al., 2013). Previously, we have shown that motor neuron effector genes whose expression is initially controlled by Isl1/Lhx3 bound enhancers transition to a distinct set of enhancers controlled by Onecut1 and Isl1 transcription factors in maturing motor neurons (Rhee et al., 2016). Accordingly, global analysis of DNA accessibility in motor neurons purified from embryonic day 10.5 and 13.5 spinal cords revealed that of 16,490 accessible regions associated with motor neuron genes only 21% (3,444) are shared between the two time points while the remaining 79% of accessible regions are stage-specific (Figure 5D). These results suggest that besides cell type specific enhancers, neuronal genes are associated with temporally dynamic sets of regulatory elements active at different developmental and maturation stages within the same neuron type, thus imposing further demand on the complexity and size of the regulatory genomic space associated with neuronal genes.

## Evolutionary expansion of neuronal regulatory space

Together our data raise the intriguing possibility that the complexity of regulatory regions associated with neuronal genes might scale with the cellular complexity of the nervous system. To explore this idea, we investigated the intergenic space associated with neuronal genes across seven species with well annotated genomes, including *Homo sapiens* (human), *Mus musculus* (mouse), *Danio rerio* (zebrafish), *Takifugu rubripes* (fugu), *Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (worm) and *Ciona intestinalis* (seq squirt). We identified a set of conserved orthologous genes across the seven examined species and compared the intergenic space associated with orthologues of mouse neuronal and non-neuronal genes. The analysis revealed that neuronal orthologues in all vertebrate species exhibit a relative increase (LogFC 0.4 - 1.2) in the size of the associated intergenic space compared to non-neuronal orthologues. In contrast, intergenic space associated with neuronal and non-neuronal orthologues in invertebrates is comparable in size (LogFC −0.2 - 0.11) (Figure 6A).

In order to directly compare the numbers of putative regulatory elements in vertebrates and invertebrates, we took advantage of available single cell ATAC-seq data, mapping genome accessibility across multiple cell types and tissues in the mouse (*Mus musculus*) and the developing fruit fly (*Drosophila melanogaster*) (Cusanovich et al., 2018a; Cusanovich et al., 2018b). We compared the numbers of putative regulatory elements proximal to all genes, the top 500 neuronal effector genes (excluding transcription factors) and all transcription factor genes. We observed that mouse neuronal effector genes were associated with the largest number of accessible genomic regions (Figure 6B-6D and S6A). Remarkably, this number was even larger than the number of putative regulatory elements associated with transcription factors, a class of genes known to be controlled by complex regulatory regions located in "gene deserts" (Nelson et al., 2004; Ovcharenko et al., 2005). Neuronal genes expressed in fruit fly were associated with a fivefold fewer putative regulatory elements than mouse neuronal genes. Moreover, fruit fly accessible regions were predominantly located in intragenic domain (within introns or exons) (Figure 6C and 6D), as shown for pairs of neuronal genes with orthologous function in the two species: a glutamate receptor gene (*Grik2/GluR*), axon pathfinding gene (*Dcc/fra*) and a calcium sensor regulating neurotransmitter release (*Syt1*) (Figure 6D).

If the evolution of new neuronal cell types is facilitated by the introduction of new regulatory elements into the genome, then regulatory elements utilized by recently evolved neuronal cell types would be expected to exhibit a lower degree of conservation. To investigate this possibility, we compared putative regulatory elements in mouse neocortical pyramidal neurons, a more recently evolved neuronal structure, to spinal motor neurons, an ancient cell type present in all vertebrates and invertebrates. We assessed per base 60-way vertebrate evolutionary conservation using PhastCons and PhyloP to quantify conservation levels in 5401 accessible non-coding DNA elements specific to excitatory neurons and 3344 elements specific to motor neurons that are associated with the shared set of broadly expressed neuronal genes (Figure 6E). While the distribution of accessible regions in intergenic space is comparable for the two cell types with median distance of motor neuron and cortical pyramidal neuron elements from promoters being 232kb and 208kb, respectively (Figure S6B), motor neuron elements exhibited >2.5 times higher levels of conservation compared to neocortical excitatory neurons (PhastCons: MN = 0.198, EX = 0.076 , Random = 0.075, p< 2e-16 Figure 6E, 6F and 6G; PhyloP: MN = 0.277, EX = 0.096, p< 2e-16, Figure S6C, S6D and S6E). These data suggest that the evolution of new cell types in the nervous system is associated with the introduction of new cell type specific regulatory elements into the non-coding regions of neuronal effector genes, facilitating the implementation of multifaceted gene expression programs underlying the expansion in the cellular diversity of the vertebrate central nervous system.

## Discussion

Arguably, the most striking finding of the human genome project is that while the total number of genes remained relatively static, non-coding DNA underwent a dramatic expansion during evolution (Lander et al., 2001). This increase has been attributed to genome duplications, segmental duplications, expansion of repetitive elements and insertion of non-coding regulatory elements (Alexander et al., 2010; Consortium, 2012; Shen et al., 2012; Touceda-Suarez et al., 2020; Yue et al., 2014). Our findings suggest that expansion of the non-coding genome is in part driven by the increased demand for more complex gene regulatory programs in vertebrates, particularly in the nervous system where an extreme diversity of cell types have evolved.

Previously, it has been noted that some genes are located in gene poor regions termed gene deserts (Lander et al., 2001; Ovcharenko et al., 2005). Evolutionarily conserved gene deserts have been shown to be predominantly associated with genes coding for transcriptional regulators. Our genome-wide analysis revealed that vertebrate neuronal effector genes are on average associated with even larger intergenic regions. This expansion of intergenic space is not associated with a specific developmental timepoint or a region of the CNS. Genes expressed in the embryonic spinal cord, dorsal root ganglia, adult cortex and cerebellum exhibit a similar 2-3 fold increase in their associated intergenic space. The intergenic size surrounding neuronal genes appears to scale with the complexity of organisms (Figure 6A, 6D) and with the complexity of gene expression patterns (Figure 1C and S1E-F). Interestingly, the variable gene deserts in which many neuronal effector genes reside exhibit significantly lower degree of conservation compared to gene deserts associated with transcription factors (Ovcharenko et al., 2005). We show that the lower degree of conservation in neuronal intergenic regions is not due to a lower density of regulatory elements, instead it is indicative of the more recent evolution of vertebrate neuronal enhancers. This raises an interesting possibility of using the conservation of cell type specific regulatory elements as a measure of evolutionary age of individual nerve cells in vertebrate CNS, as demonstrated here on the example of cortical and spinal motor neurons.

While neuronal genes are associated with larger numbers of putative regulatory elements, individual elements were shown to function in a highly cell type and cell stage specific manner (Blankvoort et al., 2018; Lindtner et al., 2019; Nord et al., 2013; Patel et al., 2021; Rhee et al., 2016; Sandberg et al., 2016; Stefanakis et al., 2015; Stroud et al., 2020; Visel et al., 2013). Accordingly, high-resolution mapping of enhancer-promoter interactions in transcriptionally programmed motor neurons revealed that only a small fraction of the putative regulatory regions are actively utilized. Comparable analysis of accessible genomic regions in several purified primary neuronal cell types from distinct parts of the nervous system and distinct developmental stages revealed that the overwhelming majority of regulatory elements are not shared. This is the case even for broadly expressed effector genes, suggesting that neuronal genes are principally not controlled by a common transcription factor that would activate a cis-regulatory element shared across diverse neuronal populations. Instead, these genes appear to be predominantly controlled by temporally dynamic and cell-type specific regulatory programs acting through distinct regulatory elements.

Whether a similar degree of enhancer specificity extends to closely related neuronal subtypes (e.g. limb and axial muscle innervating motor neurons) and to what extent regulatory regions remain dynamic during neuronal maturation remains to be determined. Initial studies in *C. elegans* suggest that a shared activator program in motor neurons might be refined by expression of subtype specific repressors (Kerk et al., 2017). The remarkable advances in single cell genomics that provide both gene expression (scRNA-seq) and putative regulatory (scATAC-seq) information might provide a deeper understanding of the regulatory logic controlling neuronal gene expression programs in the highly diverse vertebrate nervous system. Indeed, single cell studies suggest that even closely related cortical excitatory neurons utilize subtype specific regulatory programs to control expression of their effector genes (Cusanovich et al., 2018a; Li et al., 2020) and that chromatin is dynamic during multiple stages of neuronal maturation (Di Bella et al., 2021). Thus, the ever-growing number of neuronal subtypes discovered in the vertebrate CNS (Macosko et al., 2015; Moffitt et al., 2018; Saunders et al., 2018; Shekhar et al., 2016; Tasic et al., 2016), together with the findings that neuronal expression programs and their corresponding regulatory elements remain highly dynamic into early adulthood (Patel et al., 2021; Stroud et al., 2020), place further demand on the genomic space required to accommodate the neuronal regulatory programs.

This is manifested in the observation that a typical neuronal gene and its associated regulatory system often spans the majority of a topologically associated chromatin domain. Although the size of annotated TADs harboring neuronal genes is not significantly increased compared to those populated by non-neuronal genes, there are fewer total genes per TAD and the size of insulated regulatory regions surrounding neuronal genes are expanded, as revealed by a greater distances between CTCF/cohesin bound insulator elements (Figure S1D). How these large regulatory domains evolved remains to be determined. New enhancers might emerge by DNA insertion into existing regulatory space surrounding neuronal genes, thus pushing boundary elements of insulated domains further apart. Alternatively, the neuronal regulatory space might expand by a loss of existing insulated boundary elements, leading to a co-option of the neighboring genomic space. The functional importance of the large non-coding genomic domains surrounding neuronal genes is supported by the low rates of chromosomal rearrangements observed in gene deserts during evolution (~10 times lower compared to the rest of the genome), suggesting that the presence of numerous distributed regulatory elements creates a selective pressure to maintain the relationship between these non-coding domains and their target genes (Ovcharenko et al., 2005).

We demonstrate that broadly distributed neuronal regulatory regions engage in long distance chromatin interactions with promoters, forming hubs in which individual enhancers independently contribute to the ultimate levels of the target gene transcription. In this context, it is interesting to consider why functionally related neuronal enhancers are dispersed throughout the intergenic space and not found in super-enhancer-like clusters, as has been reported for other tissues (Whyte et al., 2013). We hypothesize that the observed distributed regulatory system results from independent evolution of individual enhancers, rather than from local duplication of an ancestral regulatory region. Spaced enhancers would have a tendency to drift towards each other through the loss of intervening, evolutionarily neutral, genomic DNA, resulting in a convergence of regulatory elements into super-enhancer-like domains. However, such drift is not possible around neuronal genes, where intervening genomic regions are populated by regulatory elements active in other cell types or other developmental stages. Paradoxically, the distance between functionally related neuronal enhancers likely increased during evolution, due to the insertion of new regulatory elements controlling gene expression in more recently evolved neuronal cell types. This is supported by the comparable distance distribution of enhancers active in evolutionarily older motor neurons and more recently evolved neocortical neurons (Figure S6B).

Increase in regulatory complexity in terms of genic length (Gabel et al., 2015; Sugino et al., 2019; Sugino et al., 2014), alternative splicing (Barbosa-Morais et al., 2012), 3' untranslated region length (Hilgers et al., 2011; Miura et al., 2013) has been previously ascribed to neuronal genes. Here we expand on the theme by proposing that the evolutionary increase in the size of the intergenic space and the number of regulatory elements associated with neuronal genes facilitates diversification of gene expression programs in the increasingly complex vertebrate cytoarchitecture in space and time. Although a distributed regulatory system relying on hundreds of cell-type- and cell-stage-specific enhancers associated with neuronal genes might seem inefficient and poorly designed from an engineering perspective, we hypothesize that it can dramatically increase the evolutionary flexibility and robustness of the central nervous system by reducing interdependencies in regulatory networks that control cell identity. By shifting the weight of evolution from genes to enhancers (King and Wilson, 1975), expression programs can be tuned in a highly cell type-specific manner by insertions, deletions, or mutations of regulatory elements in the non-coding genome, facilitating rapid emergence of new functionalities in existing neuronal cell types and evolution of new neuronal subtypes.

## RESOURCE AVAILABILITY

### Lead Contact

Requests for additional information, resources and reagents should be directed to the Lead Contact, Hynek Wichterle (hw350@columbia.edu).

### Materials Availability

All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement.

### Data and Code Availability

- All raw data has been deposited into GEO under Accession number GSE149971.
- All code and data analysis pipelines have been previously reported and are cited.
- Any additional information required to reanalyze the data reported in this paper is available from the Lead Contact upon request.

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Animal models

Mouse experiments were performed on embryos derived from previously generated mouse lines (Bai et al., 2015; Wichterle et al., 2002). Embryos were not screened for specific sex. Male and female adult mice were maintained following standard housing conditions in accordance with NIH guidelines and the Institutional Animal Care and Use Committee of Columbia University (IACUC).

### Cell lines

All mouse embryonic stem cell lines used and generated in this study were derived from previously generated parental lines (Mazzoni et al., 2013; Wichterle et al., 2002). Mouse embryonic stem cells (mESCs) were cultured on a monolayer of Mitomycin-C–treated mouse embryonic fibroblasts (GlobalStem) in EmbryoMax D-MEM (Life Technologies) supplemented with 15% Knockout serum replacement (Life Technologies), 2 mM L-glutamine (Life Technologies), 0.1 mM β-mercaptoethanol and 100 U $ml^{-1}$ leukemia inhibitory factor (Millipore), 1.25uM GSK3i (Selleckchem), and 100nM FRI (Gift from Austin Smith). Cells were intermittently verified to be mycoplasma negative.

## METHOD DETAILS

### mESC culture and motor neuron differentiation

Mouse embryonic stem cells (mESCs) were cultured on a monolayer of Mitomycin-C–treated mouse embryonic fibroblasts (GlobalStem) in EmbryoMax D-MEM (Life Technologies) supplemented with 15% Knockout serum replacement (Life Technologies), 2 mM L-glutamine (Life Technologies), 0.1 mM β-mercaptoethanol and 100 U $ml^{-1}$ leukemia inhibitory factor (Millipore), 1.25uM GSK3i (Selleckchem), and 100nM FRI (Gift from Austin Smith).

Motor neuron differentiation was performed as previously described (Mazzoni et al., 2013; Wichterle et al., 2002). Briefly, mESCs were dissociated with .05% trypsin (Invitrogen) and seeded in single cell suspension at $5 \times 10^4$ cells per ml in ANDFK medium (Advanced DMEM/F12:Neurobasal (1:1) Medium, 10% Knockout Serum Replacement (vol/vol), Pen/Strep, 2 mM L-glutamine, and 0.1 mM 2-mercaptoethanol) to initiate formation of embryoid bodies (EBs) (day 0). Medium was exchanged on day 2 of differentiation and motor neuron identity was induced by supplementing media with 1 μM all-trans retinoic acid (RA) to caudalize and 500nM Sonic Hedgehog agonist (SAG) to ventralize EBs for directed differentiation or 1 μM RA and 1ug/ml doxycycline to induce programming factor expression for direct programming.

### Mouse genetics and primary neuron purification

Hb9::GFP heterozygous C57/B6 males (Mazzoni et al., 2013; Wichterle et al., 2002) were mated with wildtype C57/B6 females to generate GFP labeled motor neurons at e10.5. Chat-Cre homozygous mice were mated with ROSA-LSL-Sun1::GFP (Mo et al., 2015) to generate GFP labeled motor neuron nuclei at e13.5. TrkC-TdTomato heterozygous C57/B6 males (Bai et al., 2015) were mated with wildtype C57/B6 females to generate TrkC-TdT labeled proprioceptive sensory neurons at e14.5. Embryonic spinal cords were dissected from GFP positive embryos at e10.5 and e13.5 by dissecting from forelimb to hindlimb of the embryos. Embryonic dorsal root ganglia were dissected from forelimb level of e14.5 embryos. Spinal cords or DRGs were then dissociated in Accumax (Sigma Aldrich) at 37 degrees for 15 mins with agitation. Tissue was triterated ~20 times until single cell suspension was achieved. Primary neurons or nuclei were sorted by fluorescence activated cell sorting (FACS) on a BioRad SE3 or BD FACS Aria 2.

## RNA-Seq

For mESC and mESC-derived motor neurons total RNA was extracted from approximately 200k cells using TRIzol reagent (Life Technologies) followed by isopropanol precipitation. RNA was treated with DNase-I to ensure no DNA contamination followed by acidic-phenol:chloroform extraction and another precipitation. Following RNA quality control on BioAnalyzer ribosomal RNA was depleted using the Illumina Ribo-Zero rRNA Removal Kit (MRZH116). RNA-seq libraries were prepared for sequencing using total RNA following standard Illumina protocols for 150nt kit, and sequenced by 75bp paired end sequencing on Illumina HiSeq 2000 or NextSeq500. FACS sorted primary neuron samples were prepared similarly, except RNA was purified from approximately 50k-100k cells and ribosomal RNA was depleted using Kapa Ribo Erase. Sequencing data was aligned to build version NCBI38/mm10 assembly of the mouse genome using in the STAR (Spliced Transcripts Alignment to a Reference) alignment algorithm version2.3.1 (Dobin et al., 2013). The relative abundances of transcripts were computed using the Cufflinks package release 2.2.1 (Trapnell et al., 2012). Differential expression analysis was performed using the edgeR package (Robinson et al., 2010). For differential expression genes were required to have a minimum average gene expression value of FPKM>=1 in at least one condition, p-value <.01 and logFC > |2|.

## ChIP-Seq

Chromatin immunoprecipitation experiments were performed as previously described (Mazzoni et al., 2013). Briefly, on the appropriate day of differentiation approximately 20-30 million cells were dissociated with trypsin into single cell suspension from EBs and crosslinked with 1% formaldehyde for 15 min at room temperature. Cells were lysed, chromatin pellets were isolated then solubilized and fragmented by sonication to 200-500bp length. Fragmented chromatin was then subject to immunoprecipitation using Protein G coated Dynabeads (Life Technologies) and antibodies (~5 μg) against the protein of interest (Isl1, the Developmental Studies Hybridoma Bank, clone numbers 39F7 and 4D5, gifts from T. Jessell and Susan Morton; V5, Abcam, ab15828; RNA Pol2, Covance, 8WG16; H3K27ac, Abcam, ab4729; Med1, Bethyl Labs, A300-793A; Smc1a, Bethyl Labs, A300-055A;. CTCF, Millipore, 07-729). After washing the beads to remove un-bound proteins and DNA, ChIP-seq samples were eluted from the magnetic beads and crosslinks were reversed overnight. Finally, protein and RNA were digested and DNA was purified by phenol chloroform extraction followed by ethanol precipitation. ChIP-Seq libraries were prepared following Illumina Tru-Seq ChIP Library Preparation Kit and sequenced by 40bp single-end sequencing on either Illumina HiSeq 2000 or NextSeq 500. All sequencing data were aligned to build version NCBI38/mm10 assembly of the mouse genome using Bowtie2 (Langmead and Salzberg, 2012). Peak calling was performed using the GEM peak caller (Guo et al., 2012) for transcription factors and MACS2 (Zhang et al., 2008) for histone modifications.

## ChIA-PET

ChIA-PET experiments were performed similar to previously described (Fullwood et al., 2009; Li et al., 2012). Briefly, on the appropriate day of differentiation, embryoid bodies were dissociated in trypsin into single cell suspension. Cells were cross-linked using 2mM

DSG for 15 mins followed by 1% formaldehyde for 15 mins. Fixation was quenched with 2M Glycine (125mM final concentration) for 10 mins at room temp. Cross-linked chromatin was fragmented by sonication to a size of approximately 300bp. Chromatin complexes were immunoprecipitated with monoclonal anti-RNAPII (Covance, 8WG16) coated protein G Dynabeads (Life Technologies). To prepare ChIA-PET libraries DNA was end polished with T4 DNA polymerase and then ligated with biotinylated half linker fragments with T4 DNA Ligase (Life Technologies). Next ligation was performed under dilute conditions to promote intramolecular ligation events and avoid inter-ligation products from distinct chromatin complexes. Following ligation, DNA was purified by reverse crosslinking followed by phenol-chloroform extraction and ethanol precipitation. Ligation products were released via Mme1 restriction digest followed by purification on streptavidin beads. Finally adapters were ligated and PET constructs were amplified and subjected to Illumina 40bp PE sequencing.

We developed a method called Chromatin Interaction Discovery (CID) (Guo et al., 2019) that uses unbiased density-based clustering to predict interactions directly from paired-end reads. CID takes aligned paired-end reads from ChIA-PET datasets as inputs. The PETs are then segmented into independent genomic region pairs such that no PETs in different regions are within 5000bp of each other. The distance between two PETs is evaluated as the Chebyshev distance: Distance ($PET_1$, $PET_2$) = max ( $|read_{1,a} – read_{2,a}|$, $|read_{1,b} – read_{2,b}|$ ), where $read_{i,a}$ and $read_{i,b}$ are the left and right read positions of $PET_i$, respectively. A density-based clustering algorithm (Rodriguez and Laio, 2014) is then used to cluster the nearby PETs into candidate interactions. This clustering method recognizes clusters based on the density of the data points, automatically determines the number of clusters and automatically identifies outlier data points. The only parameter for this algorithm, $d_c$, represents the distance within which data points are considered as neighbors to each other. CID estimates $d_c$ based on the genomic span between the region pairs: $d_c$ = 100 + span/20. If the two region pairs overlap with each other, the span is defined as the maximum distance among all the start and end positions of the two regions. $d_c$ is capped at 5000bp because chromatin interactions rarely contains PETs that are 5000bp apart. After clustering, singleton PETs are considered as noise and are excluded. PET clusters with count>=2 are considered candidate interactions and their statistical significance is computed using a published model MICC (He et al., 2015). Interactions with a false discovery rate 0.05 are used for downstream analysis. Following interaction calling distal chromatin anchors were defined as regions >5kb from a TSS. Distal anchors were then intersected with distal enhancers (H3K27ac or Mediator) using intersectBED from bedtools (Quinlan and Hall, 2010) requiring at least a single base pair overlap to map enhancer-promoter interactions and link distal enhancers to their target gene.

## ATAC-Seq

ATAC-seq libraries were prepared as described previously (Buenrostro et al., 2013). Approximately 20,000-50,000 cells were lysed in hypotonic buffer (10mM Tris-HCl pH 7.4, 10mM NaCl, 3mM $MgCl_2$, 0.1% NP-40) to isolate nuclei and pelleted at 800g for 6 min. The nuclear pellet was resuspended in 50 µL of the transposase reaction mix (Illumina DNA Library Preparation Kit, FC-121-1030) containing 2.5 µL Tn5 transposase, 25 µL

2xTD buffer, 22.5 uL nuclease-free water, and then incubated for 30 min at 37°C. Following transposition reaction DNA was purified using DNA-5 purification kit (Zymo Research). The purified DNA was PCR amplified with sequencing adaptors, purified using AMPure XP beads (Beckman Coulter), and sequenced by 38bp paired-end sequencing on an Illumina NextSeq 500. ATAC-Seq data was aligned using Bowtie2 to build version NCBI38/mm10 assembly of the mouse genome. Peaks were called using MACS2. Cell specific and shared ATAC peaks were identified using intersectBED from bedtools.

### Motif enrichment analysis

Motif enrichment analysis was performed using AME from the MEME suite. Individual enhancer regions were defined as 200bp windows surrounding the peak call. Input regions were searched and compared to a comparable number of random genomic sequences of the same window size. For ChIA-PET experiments peak calls for Med1 were used as anchor points for motif searching.

### Conservation analysis of accessible elements

Conservation of distal accessible DNA elements was assessed by quantifying per base PhastCons and PhyloP values downloaded from UCSC genome browser. For plots in Figure 6E, 6F and Figure S6C, S6D conservation was quantified at single nucleotide resolution in 2kb windows surrounding the peak calls for cell type specific ATAC-seq peak calls in spinal motor neurons and cortical excitatory neurons. Random regions were used as a control to assess the background levels of conservation. For plots in Figure 6G and Figure S6E average conservation values in 200bp windows surrounding cell type specific peaks was quantified.

### Human HiC chromatin interaction data

Enhancer-promoter interactions obtained from HiC data were downloaded and filtered for progenitor zone or cortical plate specific DNA accessibility and interactions as previously reported (de la Torre-Ubieta et al., 2018; Won et al., 2016). Genes with distal interactions were then filtered based on gene expression (logFC >2) to define progenitor zone specific and cortical plate specific genes and total interaction distances were measured for each set of genes. Statistical significance was determined by Wilcox rank sum test (p <.001).

### Single cell ATAC-seq

Annotated genomes of *Drosophila melanogaster* (dm3) and *Mus musculus* (mm9) were split into “domains”: overlapping transcript regions were merged, and regions between two non-overlapping but neighboring transcript regions were split at the midpoint and assigned as part of either transcript to define each non-overlapping “domain". Single-cell ATAC datasets (Cusanovich et al., 2018a; Cusanovich et al., 2018b) were then used to count the total (regardless of cell type or peak height) number of detected discrete accessible sites in any given domain. Gene expression profiles of individual tissues were obtained from bulk RNA-seq data from modENCODE for *D. melanogaster* and ENCODE and this study for *Mus musculus*. For *D. melanogaster*, larval stage 3 CNS and digestive system bulk expression profiles were used to define “neuronal” genes (top 500 genes with largest fold change of expression between CNS and digestive system among genes with TPM > 5 in at least one of the two tissues). For mouse, a generic neuronal gene expression profile was obtained as the mean of: embryonic (E14.5-E18) brain tissue, embryonic motor neurons (this study), adult cerebellum, adult cortical plate and adult prefrontal cortex gene expression profiles. Likewise, a gut gene expression profile was obtained as the mean expression profile of adult colon, duodenum, stomach, small/large intestines. Neuronal genes were defined as the top 500 genes with the largest fold change of expression with TPM > 10 in at least one tissue.

## Immunocytochemistry and Imaging

EBs were fixed for 15 min at 4°C in 4% paraformaldehyde in phosphate-buffered saline (PBS) followed by three 10 min washes in PBS at 4°C. Washed EBs were passed through a sucrose gradient to 30% sucrose (vol/vol in PBS) and then embedded in OCT (Tissue-Tek) and sectioned for immunostaining. Embryonic spinal cords were treated similarly, except for fixation. Spinal cords were dissected from embryonic mouse or chicken spinal cords and immediately fixed for 1-1.5 hours at 4°C in 4% PFA in PBS (vol/vol). Following fixation spinal cords were washed three times for 20 mins in PBS at 4°C followed by overnight wash in PBS at 4°C.

Sections were cut at 15uM for EBs and 30uM for spinal cords. Primary antibody staining was carried out for 16-24 hours at 4°C and secondary antibodies for 1.5 hours at room temperature in the dark. After staining, samples were mounted with Aqua Poly Mount (Polyscience). Antibodies used for immunostaining in this study are as follows: GFP (rabbit, Life Technologies), V5 (mouse, Life Technologies), Hb9 (mouse and guinea pig, gifts from T. Jessell and Susan Morton), Isl1 (mouse and guinea pig; gifts from T. Jessell and Susan Morton), Olig2 (guinea pig, gifts from T. Jessell and Susan Morton); Alexa 488; Cy3; Cy5-conjugated secondary antibodies were used (Life Technologies or Jackson Immunoresearch). All images were acquired using confocal laser scanning microscope (LSM Zeiss Meta 510 or 780) with either 10X or 20X objective. Images and quantifications were analyzed using Image-J, Metamorph or manually counted.

## In ovo electroporation

An enhancer reporter construct was engineered to contain a minimal promoter upstream of destabilized GFP (via PEST domain) by sub-cloning the destabilized GFP sequence from the pZsGreen1-DR Vector (Clontech) into the pGL4.23 vector (Promega) while simultaneously excising and replacing the Luciferase gene. Enhancer sequences were amplified from genomic DNA and sub-cloned into the reporter plasmid upstream of the minimal promoter sequence. Enhancer reporter constructs were concentrated to 1ug/ml and resuspended in .05% fast green for visualization during DNA injection into the neural tube. Reporters were co-electroporated with a ubiquitously expressed pCAGGS-mCherry reporter as a positive control into the developing neural tube of the Hamburger Hamilton (HH) stage 13 chick embryos (Hamburger and Hamilton, 1951) at a 10:1 ratio (1ug/ul for enhancer-GFP and 100ng/ul for CAGGS-mCherry). Electroporation was performed with 6 pulses at 50mv using an ECM 830 Square Wave Electroporation System (BTX, 45-0002). To target both ventral motor neurons and dorsally derived sensory neurons, the negative electrode was placed slightly above and positive electrode was placed slightly below the embryo for the first three pulses of electroporation to target negatively charged DNA to motor neurons in the ventral domain and then the orientation was quickly switched for dorsal targeting. Chick embryos were incubated at 39°C in a humidified incubator, and analyzed at 48 hours after electroporation for reporter expression.

### CRISPR/Cas9 genome engineering

Genome engineering was performed as previously described. Briefly, mESCs were nucleofected with Cas9-p2A-mCherry (Jacko et al., 2018) and plasmids containing appropriate gRNAs (Mali et al., 2013) using the Mouse Neural Stem Cell Nucleofector Kit (setting A24) in conjunction with the Lonza nucleofector system. To make targeted deletions of enhancers, pairs of gRNAs were designed flanking an enhancer peak spanning approximately 1kb in the genome ensuring deletion of the entirety of highly conserved DNA to ensure complete loss of function. Approximately 36 hours post-electroporation, cells were harvested and FACS-sorted by positive selection to purify cells expressing the mCherry reporter. mCherry-positive cells were plated at low density (~5k cells in 6cm dish) and cultured 6-7 days on mito-C treated mouse embryonic fibroblasts for colony expansion. Individual clones were picked and genotyped by PCR using primers flanking the deletion site by ~200bp on either side in addition to an internal primer to the enhancer. Multiple clones were generated for each enhancer deletion line and characterized to ensure reproducibility.

### Analysis of genomic organization of intergenic domains and regulatory elements

All genomic analysis was performed on the NCBI38/mm10 build of the reference genome, using mRNA genes from gencodeVM3.ens76 transcript annotation. For all analysis regarding quantifications of differences in intergenic sizes, conserved elements or accessible DNA elements statistical significance was determined by Wilcox Rank Sum test where $p < .01$.

Intergenic distances for non-coding regions were computed by first determining the longest annotated transcript for a given gene. Then the linear genomic distance was computed for all consecutive transcription start site (TSS-TSS) pairs throughout the genome. Genes were then annotated to a specific tissue by first generating an average expression profile for all neuronal and non-neuronal tissues. Cell/tissue specific expression was then determined by taking the top 500 genes ranked by gene induction in neuronal tissue compared to non-neuronal tissue or vice versa. Gene induction was computed as $\mathrm{Log}_2((\mathrm{TPMa}+1)/(\mathrm{TPMb}+1))$ where a and b are distinct cell or tissue types. For plot in Figure S1A and Figure S4A intergenic sizes were calculated as a moving average (window size 1000 genes) across genes sorted based on differential expression in neuronal and non-neuronal tissues.

Putative regulatory elements with a high degree of conservation were identified by mapping regions of continuous conservation in non-coding regions throughout the genome. We masked all coding sequence and then scanned the non-coding genome with a 500 base pair sliding window. Regulatory elements were defined as 500bp windows with an average phastCons score greater than 0.5 consistent with previous studies reporting relative conservation of distal regulatory elements (Shen et al., 2012). To assign putative DNA regulatory elements to their putative target genes we assigned based on proximity and location within annotated topological domains with the assumption that genes are more likely to target genes within their respective topological domains (Bonev et al., 2017; Dixon et al., 2015; Dixon et al., 2012). First, the genome was split into topological domains using data downloaded from Dixon et. al. then individual conserved elements were assigned to putative target genes using proximity assignment that was restricted to the closest gene within a topological domain. If no annotated topological domains existed for a gene, then the element was assigned to the most proximal gene. DNA accessibility associated with tissue specific genes was determined using the master peak list of DNase1 Hypersensitve Sites (DHS) from the mouseENCODE project (Yue et al., 2014). Peaks were assigned in the same method as done for conserved elements previously described.

Genomic data including gene ID, gene start, gene end, and mouse orthologue % identity scores were downloaded from the Ensembl BioMartv99 mouse (Mus musculus, GRCm38.p6), human (Homo sapiens, GRCh38.p13), zebrafish (Danio rerio, GRCz11), fugu (Takifugu rubripes, fTakRub1.2), fruit fly (Drosophila melanogaster, BDGP6.28), worm (Caenorhabditis elegans, WBcel235) and sea squirt (Ciona intestinalis, KH). Intergenic space associated with each gene was calculated as a sum of the upstream and downstream distances to the nearest protein-coding gene. Genes were then orthologue matched from mouse to the other six species where the orthologue with the highest score was taken. Finally, we filtered for genes with orthologue matches in all species. Neuronal genes were compared with average gene expression for non-neuronal tissues (same genes set as above for comparison of mouse domains in Figure 1). For genomic analysis across different species, we calculated the intergenic sizes associated with top 500 mouse neuronal enriched genes compared to the top 500 non-neuronal genes and computed the relative fold change in intergenic domain size.

## QUANTIFICATION AND STATISTICAL ANALYSIS

Quantifications and statistical analysis of genomic data was performed to general standards of the field (see individual methods sections for details). Image analysis was performed in ImageJ or Metamorph to determine signal intensity. Data was not assumed to be normally distributed and Mann-Whitney tests were used to determine statistical significance.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGEMENTS

We would like to thank Oliver Hobert, Tom Maniatis, and Stavros Lomvardas for advice and comments on the work; Martin Jacko for sharing reagents and experimental advice; Tom Jessell and Susan Morton for sharing reagents. Funding: H.W. is a recipient of the Jerry and Emily Spiegel endowed chair. This work was supported by NIH R01NS109217 (H.W. and D.G.); R01NS116141 (H.W.) and Project ALS (H.W.). Support for this work was further provided by NIH Grant R01HG008363 (to D.K.G.).

## REFERENCES

Alexander RP, Fang G, Rozowsky J, Snyder M, and Gerstein MB (2010). Annotating non-coding regions of the genome. Nat Rev Genet 11, 559–571. [PubMed: 20628352]

Bai L, Lehnert BP, Liu J, Neubarth NL, Dickendesher TL, Nwe PH, Cassidy C, Woodbury CJ, and Ginty DD (2015). Genetic Identification of an Expansive Mechanoreceptor Sensitive to Skin Stroking. Cell 163, 1783–1795. [PubMed: 26687362]

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science 338, 1587–1593. [PubMed: 23258890]

Blankvoort S, Witter MP, Noonan J, Cotney J, and Kentros C (2018). Marked Diversity of Unique Cortical Enhancers Enables Neuron-Specific Tools by Enhancer-Driven Gene Expression. Curr Biol 28, 2103–2114 e2105. [PubMed: 30008330]

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171, 557–572 e524. [PubMed: 29053968]

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, and Greenleaf WJ (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218. [PubMed: 24097267]

Chen AI, de Nooij JC, and Jessell TM (2006). Graded activity of transcription factor Runx3 specifies the laminar termination pattern of sensory axons in the developing spinal cord. Neuron 49, 395–408. [PubMed: 16446143]

Consortium EP (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. [PubMed: 22955616]

Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al. (2018a). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174, 1309–1324 e1318. [PubMed: 30078704]

Cusanovich DA, Reddington JP, Garfield DA, Daza RM, Aghamirzaie D, Marco-Ferreres R, Pliner HA, Christiansen L, Qiu X, Steemers FJ, et al. (2018b). The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature 555, 538–542. [PubMed: 29539636]

Dasen JS, Liu JP, and Jessell TM (2003). Motor neuron columnar fate imposed by sequential phases of Hox-c activity. Nature 425, 926–933. [PubMed: 14586461]

Dasen JS, Tice BC, Brenner-Morton S, and Jessell TM (2005). A Hox regulatory network establishes motor neuron pool identity and target-muscle connectivity. Cell 123, 477–491. [PubMed: 16269338]

de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, and Geschwind DH (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell 172, 289–304 e218. [PubMed: 29307494]

Di Bella DJ, Habibi E, Stickels RR, Scalia G, Brown J, Yadollahpour P, Yang SM, Abbate C, Biancalani T, Macosko EZ, et al. (2021). Molecular logic of cellular diversification in the mouse cerebral cortex. Nature 595, 554–559. [PubMed: 34163074]

Dickel DE, Ypsilanti AR, Pla R, Zhu Y, Barozzi I, Mannion BJ, Khin YS, Fukuda-Yuzawa Y, Plajzer-Frick I, Pickle CS, et al. (2018). Ultraconserved Enhancers Are Required for Normal Development. Cell 172, 491–499 e415. [PubMed: 29358049]

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331–336. [PubMed: 25693564]

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, and Ren B (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380. [PubMed: 22495300]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. [PubMed: 23104886]

Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schujiers J, Lee TI, Zhao K, et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell 159, 374–387. [PubMed: 25303531]

Flames N, and Hobert O (2009). Gene regulatory logic of dopamine neuron differentiation. Nature 458, 885–889. [PubMed: 19287374]

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462, 58–64. [PubMed: 19890323]

Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, Hemberg M, Ebert DH, and Greenberg ME (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature 522, 89–93. [PubMed: 25762136]

Guo Y, Krismer K, Closser M, Wichterle H, and Gifford DK (2019). High resolution discovery of chromatin interactions. Nucleic Acids Res 47, e35. [PubMed: 30953075]

Guo Y, Mahony S, and Gifford DK (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput Biol 8, e1002638. [PubMed: 22912568]

Hamburger V, and Hamilton HL (1951). A series of normal stages in the development of the chick embryo. J Morphol 88, 49–92. [PubMed: 24539719]

Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen L, Kassouf MT, Marieke Oudelaar AM, Sharpe JA, Suciu MC, et al. (2016). Genetic dissection of the alpha-globin super-enhancer in vivo. Nat Genet 48, 895–903. [PubMed: 27376235]

He C, Zhang MQ, and Wang X (2015). MICC: an R package for identifying chromatin interactions from ChIA-PET data. Bioinformatics 31, 3832–3834. [PubMed: 26231426]

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459, 108–112. [PubMed: 19295514]

Hilgers V, Perry MW, Hendrix D, Stark A, Levine M, and Haley B (2011). Neural-specific elongation of 3' UTRs during Drosophila development. Proc Natl Acad Sci U S A 108, 15864–15869. [PubMed: 21896737]

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, and Young RA (2013). Super-enhancers in the control of cell identity and disease. Cell 155, 934–947. [PubMed: 24119843]

Hobert O (2008). Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. Proc Natl Acad Sci U S A 105, 20067–20071. [PubMed: 19104055]

Hobert O (2016). Terminal Selectors of Neuronal Identity. Curr Top Dev Biol 116, 455–475. [PubMed: 26970634]

Hobert O, Carrera I, and Stefanakis N (2010). The molecular and gene regulatory signature of a neuron. Trends Neurosci 33, 435–445. [PubMed: 20663572]

Jacko M, Weyn-Vanhentenryck SM, Smerdon JW, Yan R, Feng H, Williams DJ, Pai J, Xu K, Wichterle H, and Zhang C (2018). Rbfox Splicing Factors Promote Neuronal Maturation and Axon Initial Segment Assembly. Neuron 97, 853–868 e856. [PubMed: 29398366]

Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S, et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet 51, 1442–1449. [PubMed: 31501517]

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature 467, 430–435. [PubMed: 20720539]

Kerk SY, Kratsios P, Hart M, Mourao R, and Hobert O (2017). Diversification of C. elegans Motor Neuron Identity via Selective Effector Gene Repression. Neuron 93, 80–98. [PubMed: 28056346]

Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, Li G, Resch W, Baek S, Pruett N, Grontved L, et al. (2013). Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell 155, 1507–1520. [PubMed: 24360274]

Kim N, Park C, Jeong Y, and Song MR (2015). Functional Diversification of Motor Neuron-specific Isl1 Enhancers during Evolution. PLoS Genet 11, e1005560. [PubMed: 26447474]

King MC, and Wilson AC (1975). Evolution at two levels in humans and chimpanzees. Science 188, 107–116. [PubMed: 1090005]

Konstantinides N, Kapuralin K, Fadil C, Barboza L, Satija R, and Desplan C (2018). Phenotypic Convergence: Distinct Transcription Factors Regulate Common Terminal Features. Cell 174, 622–635 e613. [PubMed: 29909983]

Kramer I, Sigrist M, de Nooij JC, Taniuchi I, Jessell TM, and Arber S (2006). A role for Runx transcription factor signaling in dorsal root ganglion sensory neuron diversification. Neuron 49, 379–393. [PubMed: 16446142]

Kratsios P, Stolfi A, Levine M, and Hobert O (2011). Coordinated regulation of cholinergic motor neuron traits through a conserved terminal selector gene. Nat Neurosci 15, 205–214. [PubMed: 22119902]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921. [PubMed: 11237011]

Langmead B, and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. [PubMed: 22388286]

Lee SK, and Pfaff SL (2003). Synchronization of neurogenesis and motor neuron specification by direct coupling of bHLH and homeodomain transcription factors. Neuron 38, 731–745. [PubMed: 12797958]

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84–98. [PubMed: 22265404]

Li YE, Preissl S, Hou X, Zhang Z, Zhang K, Fang R, Qiu Y, Poirion O, Li B, Liu H, et al. (2020). An Atlas of Gene Regulatory Elements in Adult Mouse Cerebrum. bioRxiv, 2020.2005.2010.087585.

Lindtner S, Catta-Preta R, Tian H, Su-Feher L, Price JD, Dickel DE, Greiner V, Silberberg SN, McKinsey GL, McManus MT, et al. (2019). Genomic Resolution of DLX-Orchestrated Transcriptional Circuits Driving Development of Forebrain GABAergic Neurons. Cell Rep 28, 2048–2063 e2048. [PubMed: 31433982]

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214. [PubMed: 26000488]

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, and Church GM (2013). RNA-guided human genome engineering via Cas9. Science 339, 823–826. [PubMed: 23287722]

Mann RS, and Carroll SB (2002). Molecular mechanisms of selector gene function and evolution. Curr Opin Genet Dev 12, 592–600. [PubMed: 12200165]

Mazzoni EO, Mahony S, Closser M, Morrison CA, Nedelec S, Williams DJ, An D, Gifford DK, and Wichterle H (2013). Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. Nat Neurosci 16, 1219–1227. [PubMed: 23872598]

Mazzoni EO, Mahony S, Iacovino M, Morrison CA, Mountoufaris G, Closser M, Whyte WA, Young RA, Kyba M, Gifford DK, et al. (2011). Embryonic stem cell-based mapping of developmental transcriptional programs. Nat Methods 8, 1056–1058. [PubMed: 22081127]

Miura P, Shenker S, Andreu-Agullo C, Westholm JO, and Lai EC (2013). Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res 23, 812–825. [PubMed: 23520388]

Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, Urich MA, Nery JR, Sejnowski TJ, Lister R, et al. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron 86, 1369–1384. [PubMed: 26087164]

Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac C, et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science 362.

Nelson CE, Hersh BM, and Carroll SB (2004). The regulatory content of intergenic DNA shapes genome architecture. Genome Biol 5, R25. [PubMed: 15059258]

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. [PubMed: 24360275]

Osterwalder M, Barozzi I, Tissieres V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554, 239–243. [PubMed: 29420474]

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, and Stubbs L (2005). Evolution and functional classification of vertebrate gene deserts. Genome Res 15, 137–145. [PubMed: 15590943]

Patel T, Hammelman J, Closser M, Gifford DK, and Wichterle H (2021). General and cell-type-specific aspects of the motor neuron maturation transcriptional program. bioRxiv, 2021.2003.2005.434185.

Pfaff SL, Mendelsohn M, Stewart CL, Edlund T, and Jessell TM (1996). Requirement for LIM homeobox gene Isl1 in motor neuron generation reveals a motor neuron-dependent step in interneuron differentiation. Cell 84, 309–320. [PubMed: 8565076]

Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong CT, Hookway TA, Guo C, Sun Y, et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 153, 1281–1295. [PubMed: 23706625]

Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. [PubMed: 20110278]

Rhee HS, Closser M, Guo Y, Bashkirova EV, Tan GC, Gifford DK, and Wichterle H (2016). Expression of Terminal Effector Genes in Mammalian Neurons Is Maintained by a Dynamic Relay of Transient Enhancers. Neuron 92, 1252–1265. [PubMed: 27939581]

Robinson MD, McCarthy DJ, and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. [PubMed: 19910308]

Rodriguez A, and Laio A (2014). Machine learning. Clustering by fast search and find of density peaks. Science 344, 1492–1496. [PubMed: 24970081]

Sandberg M, Flandin P, Silberberg S, Su-Feher L, Price JD, Hu JS, Kim C, Visel A, Nord AS, and Rubenstein JLR (2016). Transcriptional Networks Controlled by NKX2-1 in the Development of Forebrain GABAergic Neurons. Neuron 91, 1260–1275. [PubMed: 27657450]

Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M, Bortolin L, Wang S, et al. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell 174, 1015–1030 e1016. [PubMed: 30096299]

Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, Adiconis X, Levin JZ, Nemesh J, Goldman M, et al. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell 166, 1308–1323 e1330. [PubMed: 27565351]

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120. [PubMed: 22763441]

Shim S, Kwan KY, Li M, Lefebvre V, and Sestan N (2012). Cis-regulatory control of corticospinal system development and evolution. Nature 486, 74–79. [PubMed: 22678282]

Stefanakis N, Carrera I, and Hobert O (2015). Regulatory Logic of Pan-Neuronal Gene Expression in C. elegans. Neuron 87, 733–750. [PubMed: 26291158]

Stroud H, Yang MG, Tsitohay YN, Davis CP, Sherman MA, Hrvatin S, Ling E, and Greenberg ME (2020). An Activity-Mediated Transition in Transcription in Early Postnatal Neurons. Neuron 107, 874–890 e878. [PubMed: 32589877]

Sugino K, Clark E, Schulmann A, Shima Y, Wang L, Hunt DL, Hooks BM, Trankner D, Chandrashekar J, Picard S, et al. (2019). Mapping the transcriptional diversity of genetically and anatomically defined cell populations in the mouse brain. Elife 8.

Sugino K, Hempel CM, Okaty BW, Arnson HA, Kato S, Dani VS, and Nelson SB (2014). Cell-type-specific repression by methyl-CpG-binding protein 2 is biased toward long genes. J Neurosci 34, 12877–12883. [PubMed: 25232122]

Sun Y, Dykes IM, Liang X, Eng SR, Evans SM, and Turner EE (2008). A central role for Islet1 in sensory neuron development linking sensory and spinal gene regulatory programs. Nat Neurosci 11, 1283–1293. [PubMed: 18849985]

Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci 19, 335–346. [PubMed: 26727548]

Touceda-Suarez M, Kita EM, Acemel RD, Firbas PN, Magri MS, Naranjo S, Tena JJ, Gomez-Skarmeta JL, Maeso I, and Irimia M (2020). Ancient Genomic Regulatory Blocks Are a Source for Regulatory Gene Deserts in Vertebrates after Whole-Genome Duplications. Mol Biol Evol 37, 2857–2864. [PubMed: 32421818]

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, and Pachter L (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578. [PubMed: 22383036]

Uemura O, Okada Y, Ando H, Guedj M, Higashijima S, Shimazaki T, Chino N, Okano H, and Okamoto H (2005). Comparative functional genomics revealed conservation and diversification of three enhancers of the isl1 gene for motor and sensory neuron-specific expression. Dev Biol 278, 587–606. [PubMed: 15680372]

Visel A, Taher L, Girgis H, May D, Golonzhka O, Hoch RV, McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ, et al. (2013). A high-resolution enhancer atlas of the developing telencephalon. Cell 152, 895–908. [PubMed: 23375746]

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, and Young RA (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319. [PubMed: 23582322]

Wichterle H, Lieberam I, Porter JA, and Jessell TM (2002). Directed differentiation of embryonic stem cells into motor neurons. Cell 110, 385–397. [PubMed: 12176325]

Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 538, 523–527. [PubMed: 27760116]

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364. [PubMed: 25409824]

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137. [PubMed: 18798982]

Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, Lim J, Tai E, Poh HM, Wong E, et al. (2013). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 504, 306–310. [PubMed: 24213634]

## Highlights

- Neuronal genes have expanded intergenic domain size and regulatory complexity
- Neuronal genes are controlled by distinct cell- and stage-specific enhancers
- Neuronal enhancers are not clustered but distributed across expanded non-coding DNA
- Neuronal gene regulation expanded during transition from invertebrates to vertebrates

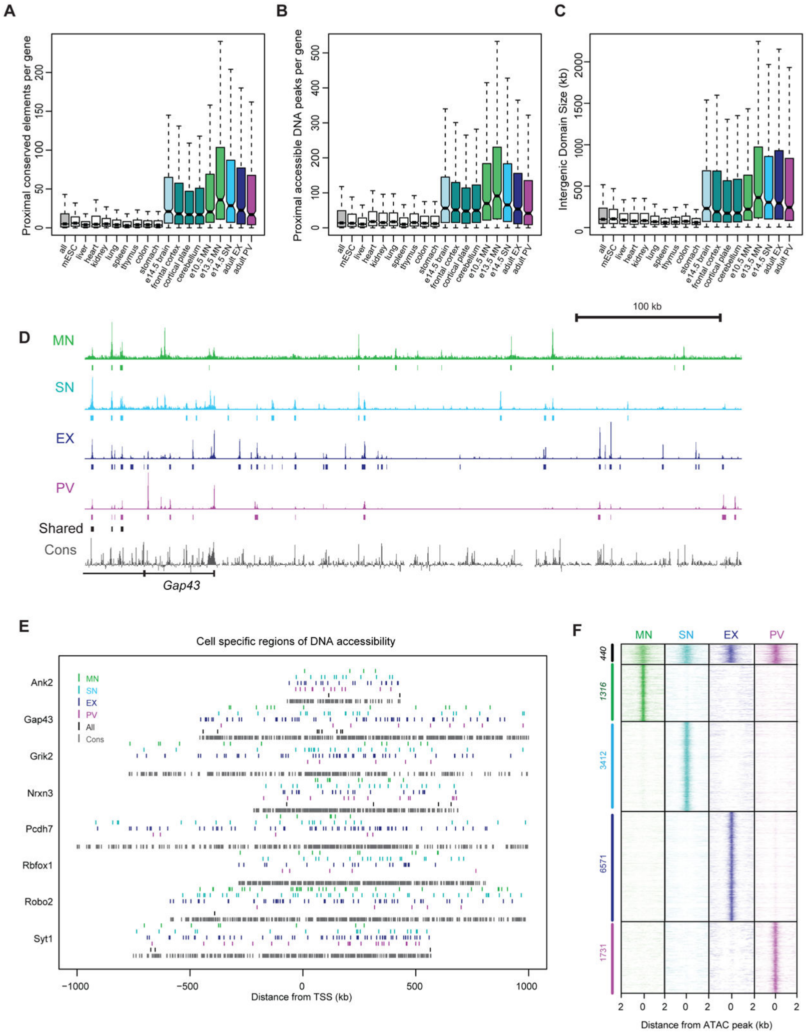

**Figure 1. An expansion in gene regulatory complexity associated with neuronal genes.**

**(A)** Quantifications of the total numbers of highly conserved (PhastCons > 0.5) putative regulatory elements in the proximity of top 500 genes highly induced in neuronal (motor neuron (green), sensory neuron (cyan), cortical excitatory neuron (navy), cortical PV interneuron (magenta), adult frontal cortex, cortical plate and cerebellum (teal) and embryonic brain (light blue)), non-neuronal tissues (white) and all genes (gray) in the mouse. **(B)** Quantifications of the total numbers of DNase accessible putative regulatory elements in the proximity of the top 500 genes highly induced in each tissue and cell type (colors are same as (A)). **(C)** Quantifications of the non-coding intergenic regulatory domain size associated with the top 500 genes highly induced in each tissue and cell type (colors are same as (A). **(D)** Cell and stage specific peaks of DNA accessibility in the proximity of the broadly expressed neuronal gene Gap43 from four distinct primary neuronal cell types (embryonic motor neuron (green); embryonic sensory neurons (cyan); adult excitatory neurons (blue); adult inhibitory neurons (magenta), shares (black) and conserved elements (gray)). **(E)** Complex patterns of cell specific DNA accessibility around broadly expressed neuronal genes encompassing many aspects of neuronal identity. Each tick mark represents a cell specific peak of DNA accessibility in the proximity of a given gene. **(F)** Genome-wide sets of all cell specific peaks associated broadly expressed neuronal genes.

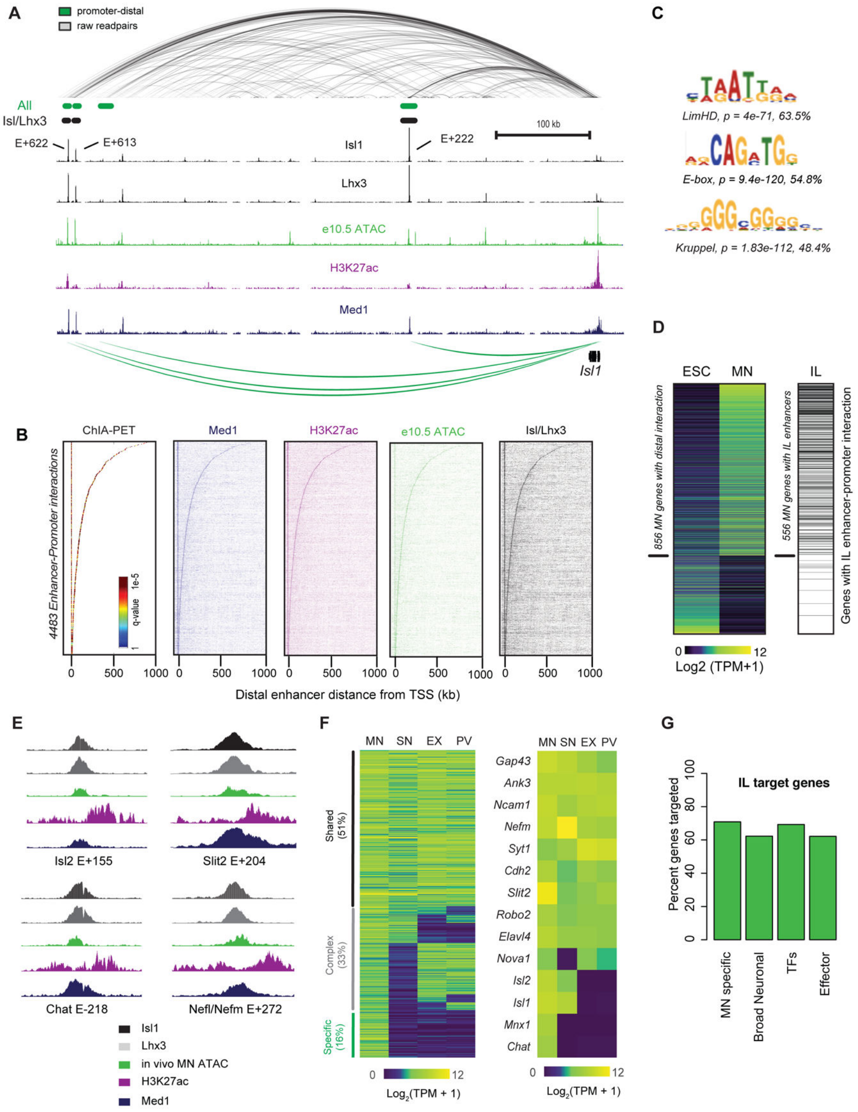

**Figure 2. Motor neuron selector transcription factors globally regulate motor neuron gene expression program.**

**(A)** Genome browser view of ~650kb window downstream of the *Isl1* locus in motor neurons. *Isl1* is engaged with 4 distal regions (green arcs and tick marks) distributed across the ~650kb region including three distinct enhancers (E+222, E+613, E+622) highly enriched for Mediator (blue) and H3K27ac (magenta), occupied by the motor neuron specific selector transcription factors Isl1 and Lhx3 (black tick marks) and exhibiting DNA accessibility in vivo (green). **(B)** Genome wide maps of all enhancer-promoter interactions ranked by interaction distance and corresponding enrichment of enhancer associated factors at two anchors (Promoter centered at x=0(left) and enhancer(right)). P-value of interaction call, raw sequencing reads for Med1(blue), H3K27ac(magenta), in vivo DNA accessibility (ATAC-seq) in e10.5 Hb9::GFP motor neurons (green) and for Isl1/Lhx3 binding at distal enhancers (black) from left to right. **(C)** Top three motifs enriched in engaged motor neuron enhancers. **(D)** Heat map displaying gene expression levels (Log2(TPM+1)) for differentially expressed genes during motor neuron specification compared to mESC (left). Binary plot showing which genes are interacting with distal Isl/Lhx3 (IL) bound enhancers (right). Black bar demarcates a gene which interacts with a distal IL bound enhancer. **(E)** High-resolution examples of ~2kb windows surrounding distal enhancers (colors same as 1a) bound by Isl1/Lhx3 regulating broad aspects of motor neuron identity (Isl2 (transcriptional identity), Chat (neurotransmitter identity), Slit2 (axon guidance), Nefl/m (structural/cytoskeleton)). **(F)** Heat map of gene expression levels for neuronal genes targeted by Isl1/Lhx3 enhancers (left, subset of Isl1/Lhx3 target genes from Figure 2D). Heat map of a selected set of Isl/Lhx3 target genes which exhibit either broad or motor neuron specific expression (right) **(G)** Bar plot showing broad targeting of Isl/Lhx3 enhancers across different gene categories.

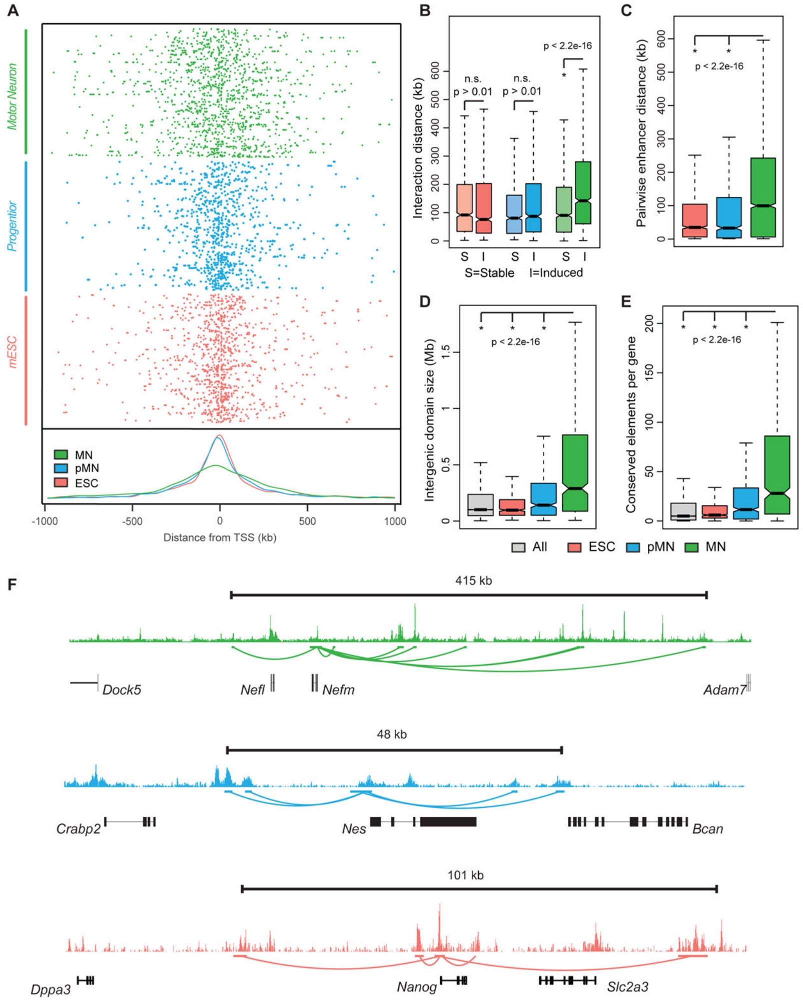

**Figure 3. Distributed gene regulatory architecture in postmitotic neurons.**

**(A)** Plots of all enhancer-promoter interactions associated with induced postmitotic motor neuron (MN) identity genes (green), motor neuron progenitor (pMN) identity genes (blue) and mouse embryonic stem (mES) cell identity genes (orange) showing motor neuron genes are regulated by distantly distributed enhancers while pMN and mES cell genes are regulated by proximally clustered enhancers. (top) Distributions for enhancer-promoter interaction distances for cell identity genes in MN (green), pMN (blue) and mES (orange) (bottom). **(B)** Boxplots of interaction distances for stably expressed (S, light shade) and induced (I, in one cell type relative to the two others, dark shade) motor neuron (green), pMN (blue) and mES (orange) genes. **(C)** Boxplots of pairwise enhancer-enhancer distances assessing distributed enhancer organization associated with motor neuron genes (green) compared to pMN (blue) and mES (orange) genes. **(D)** Boxplots of intergenic domain size for motor neuron (green), pMN (blue), mES (orange) and all (gray) genes. **(E)** Boxplots quantifying the numbers of conserved DNA elements in the proximity of motor neuron genes (green) compared to pMN (blue), mES (orange) and all (gray) genes. **(F)** Individual examples of expanded interaction domains around motor neuron genes. Enhancer-promoter interactions surrounding the postmitotic motor neuron gene, *Nefm* (green, top), neural progenitor gene *Nestin* (blue, middle) and embryonic stem cell gene *Nanog* (orange, bottom). Genome browser tracks represent enrichment of Med1 in each cell type. Black scale bars represent total interaction span for each gene.

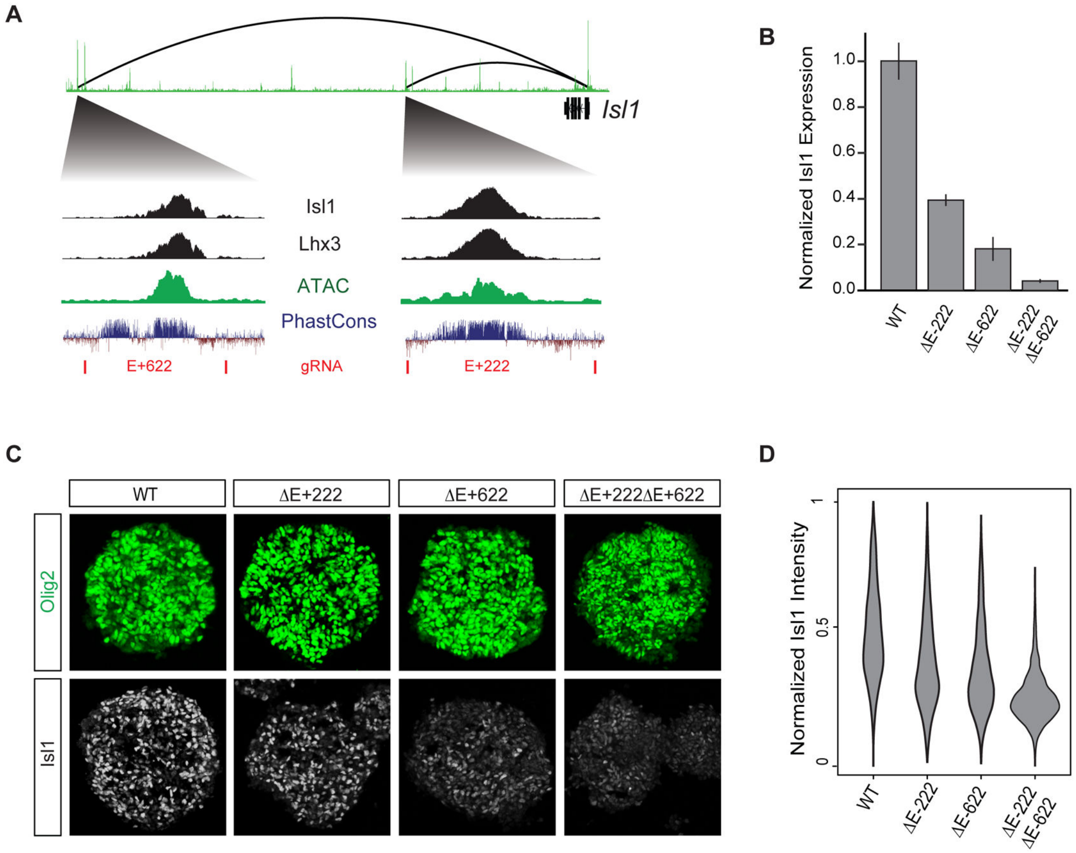

**Figure 4. Functional dissection of distributed enhancers in motor neurons.**

**(A)** Individual enhancers around the Isl1 gene targeted with pairs of gRNAs to selectively delete in isolation and combination all regions of enriched transcription factor binding (black), in vivo DNA accessibility (green) as well as highly conserved non-coding sequence (blue). **(B)** qPCR analysis of endogenous Isl1 gene expression in motor neurons upon enhancer deletion (n >=3, Data represented as the mean where error bars represent +/ − SEM). **(C)** Immunostaining analysis of Olig2 ((green) progenitor marker, top) and Isl1((gray) postmitotic marker, bottom) protein levels upon enhancer deletion during motor neuron differentiation with patterning signals retinoic acid and sonic hedgehog agonist. **(D)** Quantification of decrease in Isl1 intensity upon enhancer deletion in C (n >=3).

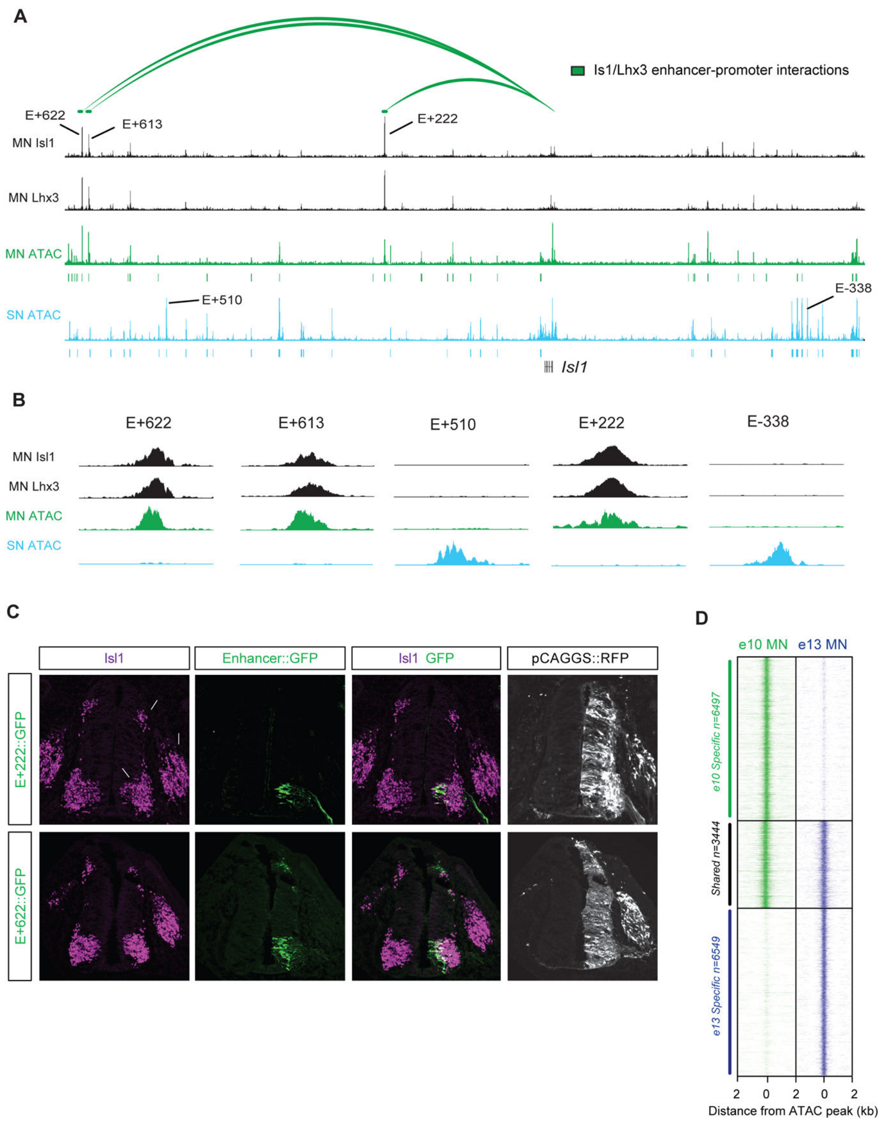

**Figure 5. Cell-type- and cell-stage-specific enhancer activity in vivo.**

**(A)** Genome browser view of chromatin interactions (green arcs), DNA accessibility and transcription factor binding in a ~1Mb genomic region surrounding the Isl1 gene displaying largely non-overlapping patterns of DNA accessibility in primary motor neurons (green track) and sensory neurons (cyan track) associated with motor neuron selector factors Isl1 and Lhx3 (black). **(B)** High resolution examples (~1-2kb) of neuron specific DNA accessibility. MN specific enhancers E+222, E+613 and E+622 are bound by Isl1 and Lhx3 and accessible specifically in primary motor neurons (green) but inaccessible in primary sensory neurons (cyan) while enhancers E-338 and E+510 exhibits SN specific DNA accessibility void of Isl1 binding in motor neurons. **(C)** Enhancer E+222 shows cell type specificity in spinal cord and dorsal root ganglia (DRG) with activity only in nascent and ventro-medial located Isl1+ motor neurons (magenta). Enhancer E+622 exhibits complex patterns of activity detected in ventro-medial located Isl1+ motor neurons (magenta) and nascent dorsal progenitors (n >=3). **(D)** Line plots of +/− 2kb window surrounding ATAC-

Seq peak calls showing dynamic reorganization of patterns of DNA accessibility around motor neuron genes in vivo between e10.5 and e13.5.

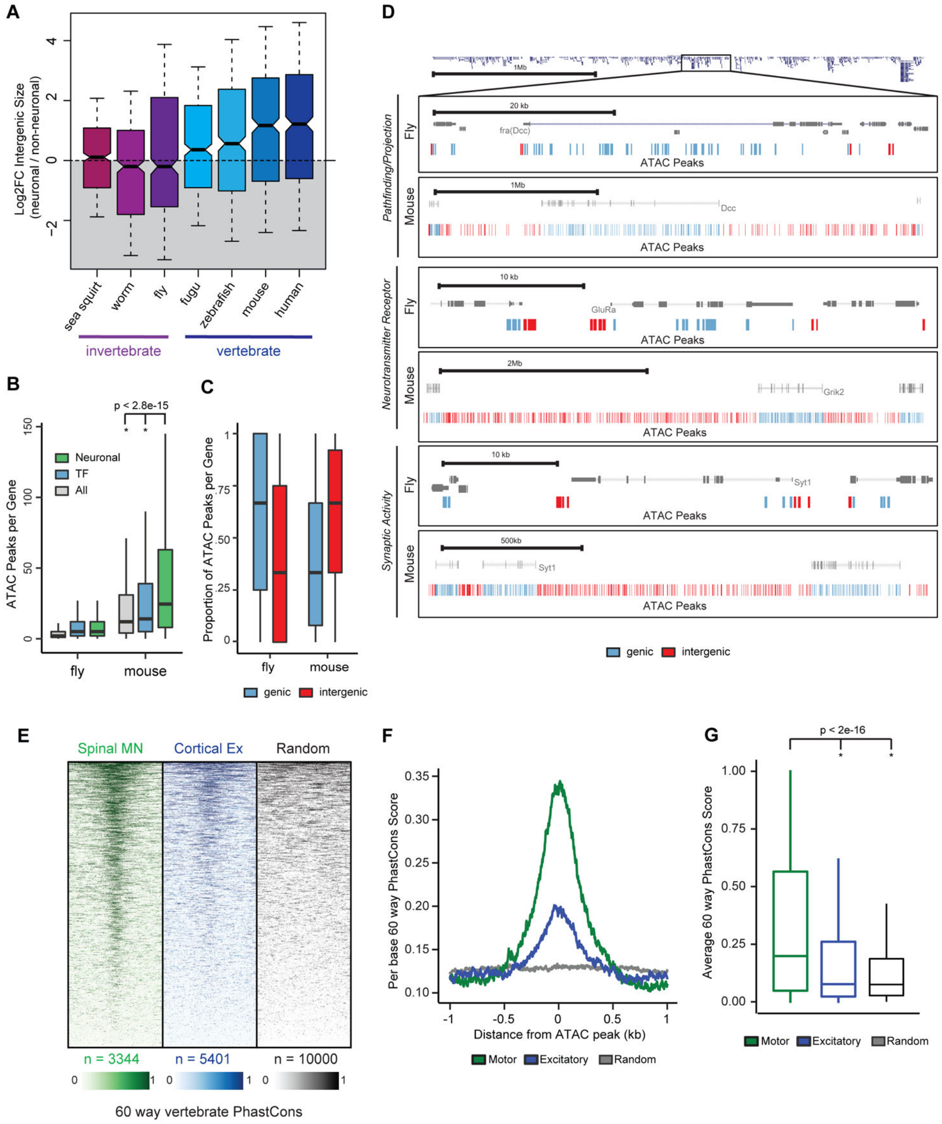

**Figure 6. Evolutionary dynamics of neuronal regulatory domains.**

**(A)** Quantifications showing fold change in relative intergenic size for orthologue matched neuronal genes compared to non-neuronal genes across seven species (vertebrates (shades of blue) and invertebrates (shades of purple)). Dashed line represents no change between the gene sets. Gray shading represents relative decrease in neuronal domain size. **(B)** Quantifications of increases in accessible DNA elements per gene associated with all genes (gray), transcription factors (blue) and neuronal effector genes (green) between fly (left) and mouse (right). **(C)** Proportions of accessible DNA elements per gene located in genic (light blue) and intergenic (red) non-coding DNA in fly (left) and mouse (right). **(D)** Expansion in intergenic domains and numbers of accessible DNA elements (genic elements (light blue) and intergenic elements (red)) around neuronal effector genes, Dcc/Fra (pathfinding), Grik2/GluRa (neurotransmitter receptor) and Syt1 (synaptic transmission) between mouse and fly. **(E)** Heatmaps representing per base 60 way vertebrate PhastCons scores in 2kb windows surrounding cell type specific peaks of DNA accessibility around broadly expressed neuronal genes in spinal motor neurons (green), cortical excitatory neurons (blue) and random genomic regions (black). **(F)** Composite plots representing average per base 60 way vertebrate PhastCons scores in 2kb windows surrounding cell type specific peaks of DNA accessibility around broadly expressed neuronal genes in spinal motor neurons (green), cortical excitatory neurons (blue) and random genomic regions (black). **(G)** Boxplots representing average per base 60 way vertebrate PhastCons scores per 200bp region surrounding cell type specific peaks of DNA accessibility around broadly expressed neuronal genes in spinal motor neurons (green), cortical excitatory neurons (blue) and random genomic regions (black).

## KEY RESOURCES TABLE

| **REAGENT or RESOURCE** | **SOURCE** | **IDENTIFIER** |
|---|---|---|
| **Antibodies** | | |
| Rabbit anti-Med1 | Bethyl Labs | A300-793A RRID:AB_577241 |
| Rabbit anti-H3K27ac | Abcam | ab4729 RRID:AB_2118291 |
| Mouse anti-Isl1/2 4D5 | Thomas Jessell/DHSB | 39.4D5 **RRID:AB_2314683** |
| Mouse anti-V5 | Thermo Fisher | R960-25<br>RRID:AB_2556564 |
| Rabbit anti-V5 | Abcam | ab9116 RRID:AB_307024 |
| Anti-RNA Pol2 | Covance | 8WG16 **RRID:AB_10013665** |
| Rabbit anti-Isl1 | Thomas Jessell | N/A |
| Guinea Pig anti-Olig2 | Thomas Jessell | N/A |
| Mouse anti-Hb9 | Thomas Jessell/DHSB | 81.5C10 RRID:AB_2145209 |
| Guinea Pig anti-Hb9 | Thomas Jessell | N/A |
| Rabbit anti-GFP | Thermo Fisher | A-6455 RRID:AB_221570 |
| Rabbit anti-CTCF | EMD Millipore | 07-729 RRID:AB_441965 |
| | | |
| Bacterial and virus strains | | |
| | | |
| Biological samples | | |
| | | |
| Chemicals, peptides, and recombinant proteins | | |
| Sonic Hedgehog Agonist (SAG) | Selleck Chemicals | S7779 |
| Retinoic Acid (RA) | Sigma-Aldrich | R2625 |
| Doxycycline | Thermo Fisher | NC0424034 |
| FRI (PD 173074) | Gift from Austin Smith | N/A |
| Gsk3I (CHIR99021) | Millipore-Sigma | 361559 |
| | | |
| Critical commercial assays | | |
| Nextera DNA Sample Preparation Kit | Illumina | FC-121-1030 |
| Lonza Neural Stem Cell Nucleofector Kit | Lonza | VPG-1004 |
| | | |
| Deposited data | | |
| ES-Derived Motor Neuron Pol2 ChIA-PET | This Study | GSE149971 |
| ES-Derived Motor Neuron RNA-Seq | This Study | GSE149971 |
| ES-Derived Motor Neuron Med1 ChIP-Seq | This Study | GSE149971 |
| ES-Derived Motor Neuron H3K27ac ChIP-Seq | This Study | GSE149971 |
| ES-Derived Motor Neuron Isl1/2 ChIP-Seq | This Study | GSE149971 |
| ES-Derived Motor Neuron Lhx3 ChIP-Seq | This Study | GSE149971 |
| E10.5 Motor Neuron RNA-Seq | This Study | GSE149971 |

| **REAGENT or RESOURCE** | **SOURCE** | **IDENTIFIER** |
|---|---|---|
| **Antibodies** | | |
| E10.5 Motor Neuron ATAC-Seq | This Study | GSE149971 |
| E13.5 Motor Neuron RNA-Seq | This Study | GSE149971 |
| E13.5 Motor Neuron ATAC-Seq | This Study | GSE149971 |
| E14.5 Sensory Neuron RNA-Seq | This Study | GSE149971 |
| E14.5 Sensory Neuron ATAC-Seq | This Study | GSE149971 |
| | | |
| Experimental models: Cell lines | | |
| mESC Hb9::GFP | Wichterle et al. 2002 | N/A |
| mESC iNgn2-Isl1-Lhx3 | Mazzoni et al. 2013 | N/A |
| | | |
| Experimental models: Organisms/strains | | |
| Mouse: C57BL/6 | Jackson Lbaoratories | N/A |
| Mouse: Hb9-GFP | Wichterle et al. 2002 | N/A |
| Mouse: TrkC-TdTomato | Bai et al. 2015 | N/A |
| Mouse: Chat-Cre x Rosa-LSL-Sun1-GFP | Mo et. al 2015 | N/A |
| Pathogen Free Fertile Sterile Chicken Eggs | Charles River | 10100326 |
| | | |
| Oligonucleotides | | |
| gRNA_E222_1 | This study | N/A |
| gRNA_E222_2 | This study | N/A |
| gRNA_E622_1 | This study | N/A |
| gRNA_E622_2 | This study | N/A |
| Isl1_qPCR_fw | This study | N/A |
| Isl1_qPCR_rv | This study | N/A |
| | | |
| Recombinant DNA | | |
| pCAGGS_Cas9_mCherry | Jacko et al. 2018 | N/A |
| gRNA cloning vector | Addgene | #41824 |
| minP_destabalized_GFP (minP_dGFP) | This study | N/A |
| minP_E222_dGFP | This study | N/A |
| minP_E622_dGFP | This study | N/A |
| | | |
| Software and algorithms | | |
| CID | Guo et al. 2019 | N/A |
| MACS2 | Zhang et al. 2008 | N/A |
| GEM | Guo et al. 2012 | N/A |
| EdgeR | Robinson et al. 2010 | N/A |
| Bedtools | Quinlan et al. 2010 | N/A |

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Antibodies** | | |
| | | |