bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Transcriptional Logic of Mammalian Neuronal Diversity

Ken Sugino[1*], Erin Clark[2], Anton Schulmann[1], Yasuyuki Shima[2], Lihua Wang[1], David L. Hunt[1], Bryan M. Hooks[1], Dimitri Tränkner[1], Jayaram Chandrashekar[1], Serge Picard[1], Andrew Lemire[1], Nelson Spruston[1], Adam Hantman[1], Sacha B. Nelson[2*]

***For correspondence:**
suginok@janelia.hhmi.org (FMS);
nelson@brandeis.edu (FS)

[1]Janelia Research Campus; [2]Brandeis Universiy

Abstract The mammalian nervous system is constructed of many cell types, but the principles underlying this diversity are poorly understood. To assess brain-wide transcriptional diversity, we sequenced the transcriptomes of the largest collection of genetically and anatomically identified neuronal classes. Using improved expression metrics that distinguish information content from signal-to-noise-ratio, we found that homeobox transcription factors contain the highest information about cell types and have the lowest noise. Genes that contribute the most to neuronal diversity tend to be long and enriched in factors specifically involved in neuronal function. Genome accessibility measurements reveal that long genes have more candidate regulatory elements arrayed in more distinct patterns. These elements frequently overlap interspersed repeats (mobile elements) and the pattern of repeats is predictive of gene expression. New regulatory sites resulting from elongation of neuronal genes by mobile elements may be an evolutionary force enhancing nervous system complexity.

Introduction

The extraordinary diversity of vertebrate neurons has been appreciated since the proposal of the neuron doctrine (***Cajal, 1888***). Typically, this diversity is characterized by neuronal morphology, physiology, molecular expression, and circuit connectivity. The exact number of neuronal cell types remains unknown, but estimates of 40-60 have been provided for the retina (***Macosko et al., 2015***; ***Masland, 2004***) and for mouse cortex (***Tasic et al., 2016***; ***Zeisel et al., 2015***). If similar numbers are discovered in most brain regions, the number could be in the thousands or more. Although neuronal diversity has long been recognized, the question of how this diversity arises is only beginning to be addressed (***Arendt, 2008***; ***Muotri and Gage, 2006***). Describing the cell types of the brain and understanding the principles governing their diversity are fundamental goals for neuroscience.

Currently two techniques dominate the efforts to profile the transcriptional diversity of cell types in the brain: one is RNA-seq from single neurons, (single-cell RNA-seq; SCRS), (e.g. ***Shapiro et al., 2013***) and the other is from genetically or anatomically marked pools of neurons (e.g. ***Okaty et al., 2015***; ***Cembrowski et al., 2016***). An obvious advantage of the SCRS approach is that, by definition, each measurement comes from only a single cell type. However, SCRS measurements can be noisy and, depending on the approach, can have limited depth and sensitivity (***Parekh et al., 2016***; ***Svensson et al., 2017***). So far, the field attempts to generate accurate and precise transcriptional profiles of cell types by clustering and then averaging the profiles of single cells. But the process of clustering itself can add noise (***Ntranos et al., 2016***), and the unbiased nature of the measurement complicates the assessment of reproducibility. Pooling reduces noise, but

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

can suffer from unknowingly lumping together more than one cell type. In the end, performing both methods will allow for a more confident assessment of the cell types of the brain. While large, unbiased single cell efforts have been completed or are underway, similar large scale efforts for genetically identified neurons have yet to be reported. We performed RNA-seq on the largest set to date of genetically identified and fluorescently labeled pooled neurons from micro-dissected brain regions. In total, we profiled 179 neuronal cell types and 15 non-neuronal cell types and quantitatively compared our cortical profiles to those obtained in SCRS studies. (A more precise description of our use of the term "cell type" is provided in the Methods). The comparison reveals a comparable level of homogeneity, but a much lower level of noise in the bulk sorted profiles. We have curated these reproducible and precise expression profiles to serve as a look-up table for linking single cell and cell type expression profiles to genetic strains in which they can be repeatedly accessed.

Cell types are typically identified by performing differential expression analyses. Standard differential expression methods focus on signal variance but are influenced by both information content and robustness of differential expression. We introduced two simple metrics to separate out these features of the data. Signal contrast (SC) is a signal-to-noise ratio that (unlike ANOVA) is not sensitive to differences in information content. Differentiation index (DI) is a measure of information content closely related to mutual information. Using these metrics, we identify homeobox transcription factors (TF) as the gene family with the lowest noise and highest ability to distinguish cell types and use these and other TFs to construct a compact “code” for profiled neuronal cell types. We find that the effector genes carrying the most information about cell types are synaptic genes like receptors, ion channels and cell adhesion molecules. Interestingly, a common feature of these genes is their long genomic length, reflecting the increased number and length of their introns. Our ATAC-seq results indicate that long genes contain a larger number of candidate regulatory regions which are arrayed in more diverse patterns than found in short genes, suggesting the longer length of the genes may permit increased regulatory complexity. Moreover, these long genes are elongated during evolution by insertions of mobile elements and a large portion of the candidate regulatory regions identified by ATAC-seq overlap with these mobile elements. Thus, the increased length of neuronal genes may provide a platform for evolution to fine-tune gene expression and thus diversify the cell types of the nervous system.

Results

A dataset of cell type-specific neuronal transcriptomes

To begin exploring the diversity in the nervous system, we collected transcriptomes from 166 types of neurons and 15 types of non-neuronal genetically/retrogradely labeled cell populations (Table 1; Figure 1 Supplement 1; Supplementary Table 1,2). Data from 9 previously published hippocampal cell types (***Cembrowski et al., 2016***), 2 hypothalamic cell types (***Henry et al., 2015***), and 2 neocortical cell types (***Shima et al., 2016***), harvested and processed in the same way as other samples, were also included in our analyses. Each neuron type collected represents a group of fluorescently labeled cells dissociated and sorted from a specific micro-dissected region of the mouse brain or other tissue. In most cases, the fluorescent label was genetically expressed in a mouse driver line, but retrograde labeling was used in some cases. The pipeline for cell type-specific transcriptome collection is depicted in Figure 1A (see Methods for additional details). Mouse lines were first characterized by generating a high resolution atlas of reporter expression (Figure 1B), then regions containing labeled cells with uniform morphology were chosen for sorting and RNA-seq. This effort constitutes the largest and most diverse single collection of genetically identified cell types profiled by RNA-seq. The processed data, including anatomical atlases, RNASeq coverage, and TPM are available at http://neuroseq.janelia.org (Figure 1C).

To determine the sensitivity of our transcriptional profiling, we used ERCC spike-ins. Amplified RNA libraries had an average sensitivity (50% detection) of 23 copy*kbp of ERCC spike-ins across all

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

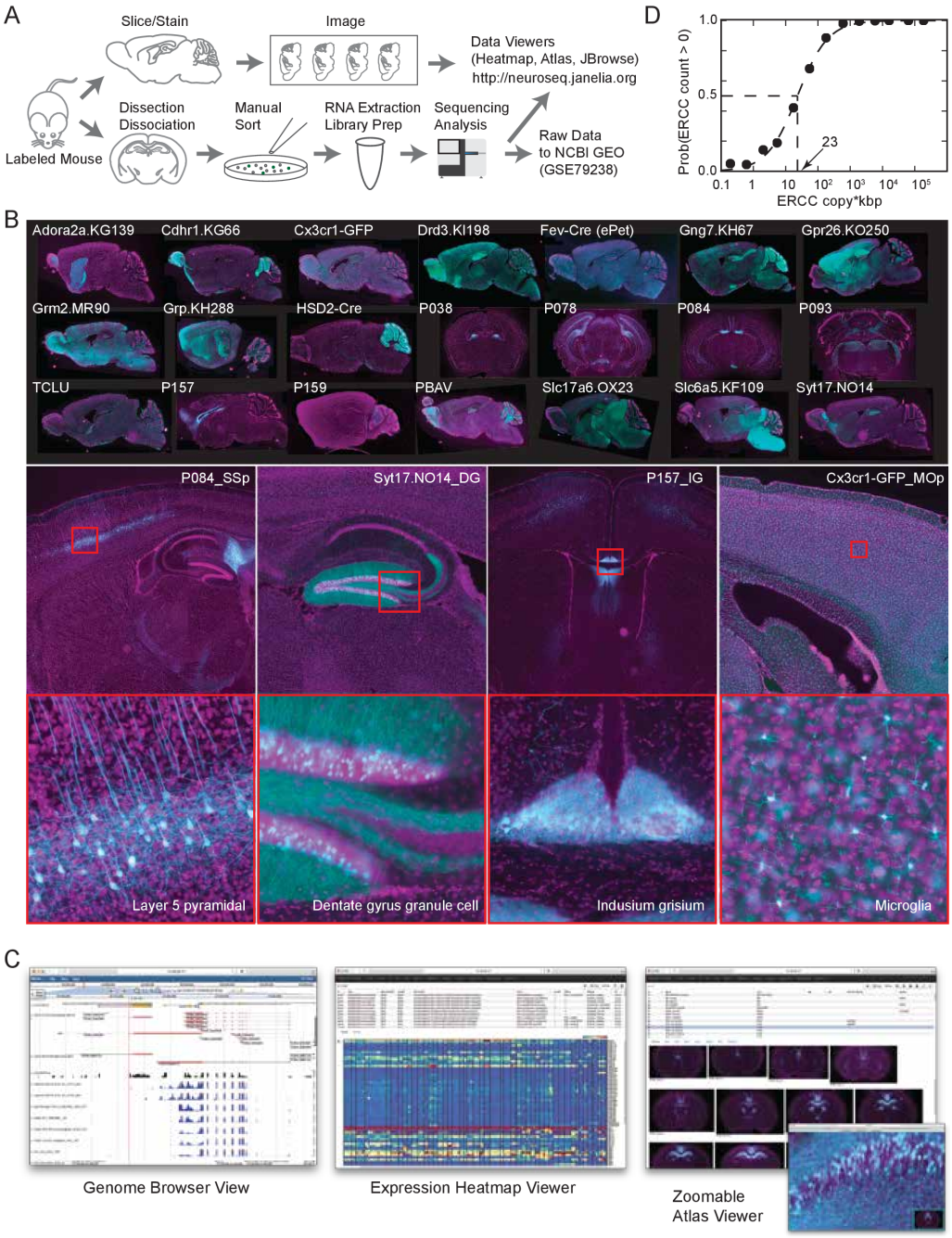


Figure 1. The NeuroSeq dataset. (A) Schema of pipeline for anatomical and genomic data collection. **(B)** Example sections from atlases at low (top), medium (middle) and high (bottom) magnifications. **(C)** Web tools available at http://neuroseq.janelia.org

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

libraries (Figure 1D). Since manually sorted samples had 132±16 cells (mean± sem, all following as well), this indicates our pipeline had the sensitivity to detect a single copy of a transcript per cell 80% of the time. In total we sequenced 2.3 trillion bp in 565 libraries. Total reads per library was 41±0.5M reads (Figure 1 Supplement 2A top). Using the aligner STAR (***Dobin et al., 2012***), 68.9±0.37% of the reads mapped uniquely to the mm10 genome, 2.8±0.06% mapped to multiple loci, 5.6±0.14% did not map to mm10, and 22.7±0.36% contained abundant sequences such as ribosomal RNA or mitochondrial sequences (Figure 1 Supplement 2A bottom) and 0.06%± 0.004% contained short reads (less than 30bp after removing adaptor sequences). Sequenced library data were deposited in NCBI GEO (accession number:GSE79238). This high sensitivity allowed for deep transcriptional profiling in our diverse set of cell types.

To assess the extent of contamination in the dataset, we checked expression levels of marker genes for several non-neuronal cell types (Figure 1 Supplement 2B). As previously shown (***Okaty et al., 2011***), manual sorting produced, in general, extremely clean data.

To demonstrate the utility of the dataset, made possible by its broad sampling of cell types, we extracted pan-neuronal genes (genes expressed commonly in all neuronal cell types but expressed at lower levels or not at all in non-neuronal cell types; Figure 1 Supplement 3). Broad sampling is essential to avoid false positives (***Zhang et al., 2014b***; ***Mo et al., 2015***; ***Stefanakis et al., 2015***). Extracted pan-neuronal genes contain well known genes such as *Eno2* (Enolase2), which is the neuronal form of Enolase required for the Krebs cycle, *Slc2a3* (chloride transporter) required for inhibitory transmission, and *Atp1a3* (ATPase Na+/K+ transporting subunit alpha 3) which belongs to the complex responsible for maintaining electrochemical gradients across the membrane, as well as genes not previously known to be pan-neuronal, such as *2900011O08Rik* (now called Migration Inhibitory Protein;***Zhang et al.*** (***2014a***)). Synaptic genes are often differentially expressed among neurons, but some included in this pan-neuronal list such as *Syn1, Stx1b, Stxbp1, Sv2a*, and *Vamp2* appear to be common components required in all neurons, highlighting essential parts of these complexes. Thus, this pan-neuronal gene list reveals components necessary for any neuron. The dataset should also be useful for many other applications, especially those requiring comparisons across a wide variety of neuronal cell types.

Comparison to single cell datasets

Pools of sorted neurons may be heterogeneous if multiple neuronal subtypes are labeled in the same brain region of the same strain. SCRS has recently emerged as a viable method for profiling cellular diversity that does not suffer from this limitation. However, since profiles of cell types in SCRS studies are obtained by clustering individual, often noisy, cellular profiles, inaccuracies can arise from misclustering or overclustering. In order to assess the relative cellular homogeneity of our sorted samples, we compared the current dataset to the cluster profiles from SCRS studies. We focused on neuronal and non-neuronal cell types in the neocortex, profiled in two recent studies (***Tasic et al., 2016***; ***Zeisel et al., 2015***). Assuming each sorted population corresponds to a linear combination of one or more SCRS profiles, we assessed homogeneity by linear decomposition using non-negative least squares (NNLS). We performed multiple checks on the validity of the procedure (see Figure 2 Supplements 1-3 and Methods) and found that it is able to fairly accurately decompose mixtures of component expression profiles when those components are well separated.

For each sorted cell type, the procedure identifies the weights (coefficients) of component clusters (cell types) from the SCRS datasets (Figure 2A). As expected, cell types present in the SCRS studies, but not profiled in NeuroSeq, (e.g. L4 neurons, VIP interneurons and oligodendrocytes), were not matched (purely blue columns in Figure 2A). Other cell types matched perfectly to a single SCRS cell type (e.g., microglia, astrocytes, ependyma) or matched to more than one, implying heterogeneity in the sorted profiles or poor separation of the SCRS profiles. Profiles with imperfect matches usually matched closely related cell types. For example, the NeuroSeq Pvalb interneuron group matched one or two of the SCRS Pvalb-positive interneuron clusters, and layer 2/3 (L2/3) pyramidal neurons matched SCRS L2/3 clusters, or an adjacent cluster in L4 (Tasic: L4 Arf5). The

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 1. Summary of Profiled Samples.

	region/type	transmitter	#groups	subregions	#samples
CNS neurons	Olfactory (OLF)	glu	10	AOBmi,MOBgl,PIR,AOB,COAp	30
		GABA	4	AOBgr,MOBgr,MOBmi	11
	Isocortex	glu	22	VISp,AI,MOp5,MO,VISp6a,SSp,SSs,ECT,ORBm,RSPv	68
		GABA	3	Isocortex,SSp (Sst+, Pvalb+)	7
		glu,GABA	1	RSPv	3
	Subplate (CTXsp)	glu	1	CLA	4
	Hippocampus (HPF)	glu	24	CA1,CA1sp,CA2,CA3,CA3sp,DG,DG-sg,SUBd-sp,IG	65
		GABA	4	CA3,CA,CA1 (Sst+, Pvalb+)	12
	Striatum (STR)	GABA	12	ACB,OT,CEAm,CEAl,islm,isl,CP	33
	Pallidum (PAL)	GABA	1	BST	4
	Thalamus (TH)	glu	11	PVT,CL,AMd,LGd,PCN,AV,VPM,AD	29
	Hypothalamus (HY)	glu	11	LHA,MM,PVHd,SO,DMHp,PVH,PVHp	36
		GABA	4	ARH,MPN,SCH	15
		glu,GABA	2	SFO	3
	Midbrain (MB)	DA	2	SNc,VTA	5
		glu	2	SCm,IC	6
		5HT	2	DR	10
		GABA	1	PAG	4
		glu,DA	1	VTA	3
	Pons (P)	glu	7	PBl,PG	22
		NE	1	LC	2
		5HT	2	CSm	7
	Medulla (MY)	GABA	7	AP,NTS,MV,NTSge,DCO	18
		glu	6	NTSm,IO,ECU,LRNm	20
		ACh	2	DMX,VII	6
		5HT	1	RPA	3
		GABA,5HT	1	RPA	4
		glu,GABA	1	PRP	3
	Cerebellum (CB)	GABA	10	CUL4, 5mo,CUL4, 5pu,CUL4, 5gr,PYRpu	25
		glu	4	CUL4, 5gr,NODgr	10
	Retina	glu	5	ganglion cells (MTN,LGN,SC projecting)	14
	Spinal Cord	glu	1	Lumbar (L1-L5) dorsal part	3
		GABA	4	Lumbar (L1-L5) dorsal part, central part	12
PNS	Jugular	glu	2	(TrpV1+)	7
	Dorsal root ganglion (DRG)	glu	2	(TrpV1+, Pvalb+)	5
	Olfactory sensory neurons (OE)	glu	4	MOE,VNO	9
non-neuron	Microglia		2	MOp5(Isocortex),UVU(CB) (Cx3cr1+)	6
	Astrocytes		1	Isocortex (GFAP+)	4
	Ependyma		1	Choroid Plexus	2
	Ependyma		2	Lateral ventricle (Rarres2+)	6
	Epithelial		1	Blood vessel (Isocortex) (Apod+,Bgn+)	3
	Epithelial		1	olfactory epithelium	2
	Progenitor		1	DG (POMC+)	3
	Pituitary		1	(POMC+)	3
non brain	Pancreas		2	Acinar cell, beta cell	7
	Myofiber		2	Extensor digitorum longus muscle	7
	Brown adipose cell		1	Brown adipose cell from neck.	4
		total	194		565

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

spread of coefficients repeatedly involved the same few SCRS cell clusters (e.g. columns L5b Tph2 and L5b Cdh13 in Tasic; and S1PyrL5, S1PyrL6 in Zeisel), which could occur if these clusters are not well separated, which we confirmed by a cross-validation procedure (Figure 2 Supplement 3). We measured the "purity" of the decomposition as the fractional match to the highest coefficient. The purity scores for the decomposition of NeuroSeq cell types by the two SCRS datasets were higher than those obtained for SCRS cell clusters decomposed by the other SCRS data set (Figure 2B,C). This implies that although sorted sample heterogeneity may exist in some of our sorted samples, it is comparable (or smaller) than the inaccuracies introduced by clustering single cell profiles. We also compared the separability of cell types assayed in the sorted and SCRS datasets (Figure 2 Supplement 4) by calculating the gene expression distances between each cell type within each dataset. NeuroSeq profiles were far more separable than clusters in either SCRS dataset, likely because of the noise reduction achieved by averaging across cells and because of the larger numbers of cells and reads comprising each profile. Hence sorted and single cell techniques have complimentary strengths and cross referencing both data modalities may provide the most accurate assessment of cell type specific expression.

Improved metrics to quantify differential expression

Analysis of expression differences between individual groups is the basis of most profiling efforts. Variance-based metrics, such as Analysis of Variance (ANOVA) F-Value or coefficient of variation (CV) are commonly used for this purpose. These metrics are jointly affected by the information content of the differential expression (pattern) and the robustness of the differences (effect size) and so cannot readily separate these two parameters. As a complement to traditional metrics and to begin mining our extensive and complex dataset for novel insights, we developed two easily calculated metrics that better separate the information content and the robustness of expression differences.

First, in order to extract the transcriptional signals related to cell type identity, we quantified each gene's ability to differentiate each pair of profiled cell types. Based on expression levels and variability (Figure 3A; Methods) we compiled a Differentiation Matrix (DM) with elements equal to one or zero depending on whether or not the gene is differentially expressed between each pair of profiles (see Methods). The Differentiation Index (DI) is simply the fraction of pairs distinguished, excluding self-comparisons; and ranges from 0 to 1. The maximum observed value of 0.65 indicates that the gene distinguishes 65% of the pairs, while a value of 0 indicates that the gene distinguishes none (i.e., expressed at similar levels in all cell types).

The ability to detect transcriptional differences between cell types depends on both magnitude of difference and associated noise. To quantify this in our second metric, we defined the Signal Contrast (SC), which closely reflects Signal-to-Noise-Ratio (SNR). Since the signals we are interested in are the gene expression differences distinguishing cell types, we used a noise estimate derived from all undistinguished pairs from the same gene. SC, which indicates how robustly pairs are distinguished, is the ratio of the average effect size for distinguished and undistinguished pairs. High SC genes robustly distinguish cell populations and are therefore suitable as "marker genes".

Our metrics outperform existing metrics such as ANOVA, CV, and Fano factor in distinguishing the information content and robustness of differential expression. To illustrate the properties of DI and SC relative to existing metrics, we calculated these metrics against various simulated expression patterns with added noise (Figure 3 Supplement 1A). The results (Figure 3 Supplement 1A, lower part) demonstrate that DI (blue) is highly correlated with mutual information (MI; green), yet much easier to calculate. This makes intuitive sense, since the division of cell types into those that can and cannot be distinguished (DM; Figure 3A) corresponds to a unit of information about cell types provided by a gene expression pattern (for more details of the relationship between DI and MI, see Figure 3 Supplement 1C and 2). The simulations also show that DI is fairly independent from SNR. For example, both high and low SNR binary patterns yield similar DIs. In contrast, SC (orange) is independent from MI, but is highly correlated to SNR. Thus, DI provides an estimate of the information content of expression patterns across cell types, whereas SC provides an estimate

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

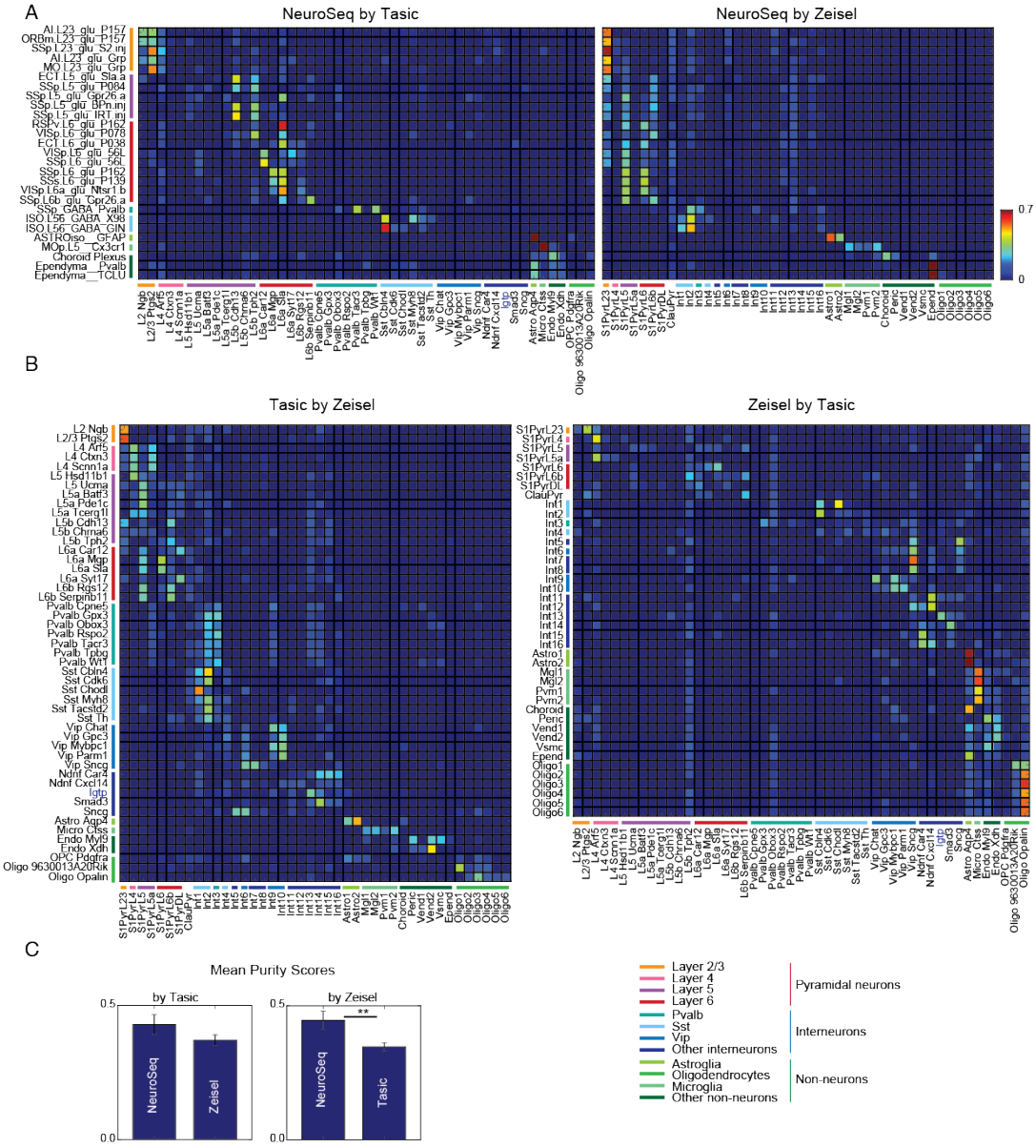


Figure 2. Decomposition by NNLS. (A) NNLS coefficients of NeuroSeq cell types by two SCRS datasets. **(B)** (Left) Tasic et al. clusters decomposed by Zeisel et al. clusters. (Right) Zeisel et al. clusters decomposed by Tasic et al. clusters. There are few perfect matches. **(C)** Mean purity scores for NeuroSeq and SCRS datasets. The purity score for a sample is defined as the ratio of the highest coefficient to the sum of all coefficients. (**:p<0.01, t-test.)

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

of SNR.

Unlike DI and SC, traditional variance-based methods like ANOVA F-values and CV are either affected by both MI and SNR (ANOVA) or by neither (CV). These differences between metrics are summarized in Figure 3 Supplement 1B. The fact that ANOVA does not distinguish between information content and SNR is also apparent in the data. As shown in Figure 3B, high-ANOVA genes include both high DI and high SC genes. Therefore, SC and DI are useful because they provide independent measures of the robustness and magnitude of differential expression between cell types.

Genes with the highest information regarding cell types

To determine the types of genes most differentially expressed (highest DI) and most robustly different (highest SC) between cell types, we used the PANTHER (***Thomas, 2003***) gene families (Figure 3D). As expected, high DI genes are enriched for neuronal effector genes including receptors, ion channels and cell adhesion molecules (Figure 3D top). The highest signal-to-noise expression differences (highest SC) were those of homeobox transcription factors (TFs) and the more inclusive categories (TFs, DNA binding proteins) that encompass them (Figure 3D bottom). Hence DI and SC respectively emphasize the information content of genes mediating the distinctive neuronal phenotypes that distinguish cell types, and the robust, low-noise expression of genes involved in shaping these cell types unique transcriptional programs.

Genes may also contribute to cell type differences through differential splicing. We analyzed splicing events by computing the relative likelihood (branch probabilities) of each donor site in a transcript being spliced to multiple acceptor sites, and of each acceptor site being spliced to multiple donors (Figure 3C). Interestingly, when these branch probabilities are computed separately for each cell type, they are highly bimodal, reflecting virtually all-or-none splicing at each alternatively spliced site. This pattern has previously been observed for individual cells in some systems (***Shalek et al., 2013***). The present observations suggest that these splicing decisions are made at the level of cell types, rather than independently for individual cells of the same type. We applied a variant of the DM/DI method to alternative splicing (Figure 3C,E,F; for details see Methods) and found that voltage-gated calcium and sodium channels are highly alternatively spliced, consistent with previously known results (e.g. ***Lipscombe et al., 2013***). We also found that G-protein modulators, especially guanyl-nucleotide exchange factors (GEFs), are highly alternatively spliced. Hence, differential splicing of multi-exon genes also contributes to transcriptome diversity across neuronal cell types.

SC, like SNR, is a ratio between signal and noise, and so can reflect high expression levels in ON cell types (high signal), low expression levels in OFF cell types (low noise), or both. Homeobox genes are not among the most abundantly expressed genes. Their average expression levels (~30 FPKM) are significantly lower than, for example, those of neuropeptides (~90 FPKM). This suggests that the high SC of homeobox TFs depend more on low noise than on their high signal. In fact, most homeobox TFs have uniformly low expression in OFF cell types (e.g. Figure 4A). We quantified this "OFF noise" for all genes and found that homeobox genes are enriched among genes that have both low OFF noise and at least moderate ON expression levels (red dashed region in Figure 4B).

Since tight control of expression may reflect closed chromatin, we measured chromatin accessibility using ATAC-seq (***Buenrostro et al., 2013***) on 7 different neuronal cell types (see Methods). As expected, compared to high-noise genes (Figure 4C bottom), genes with low OFF noise were more likely to have fewer, smaller peaks within their transcription start site (TSS) and gene body (Figure 4C top, Figure 4D), consistent with the idea that their expression is controlled at the level of chromatin accessibility.

Functionally, the tight control of homeobox TF expression levels may reflect their known importance as determinants of cell identity, and the fact that establishing and maintaining robust differences between cell types may require tight ON/OFF regulation rather than graded regulation. If they are, in fact, important "drivers" of cell type-specific differences, their expression pattern should be highly informative about cell types. However, the homeobox family was not identified

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

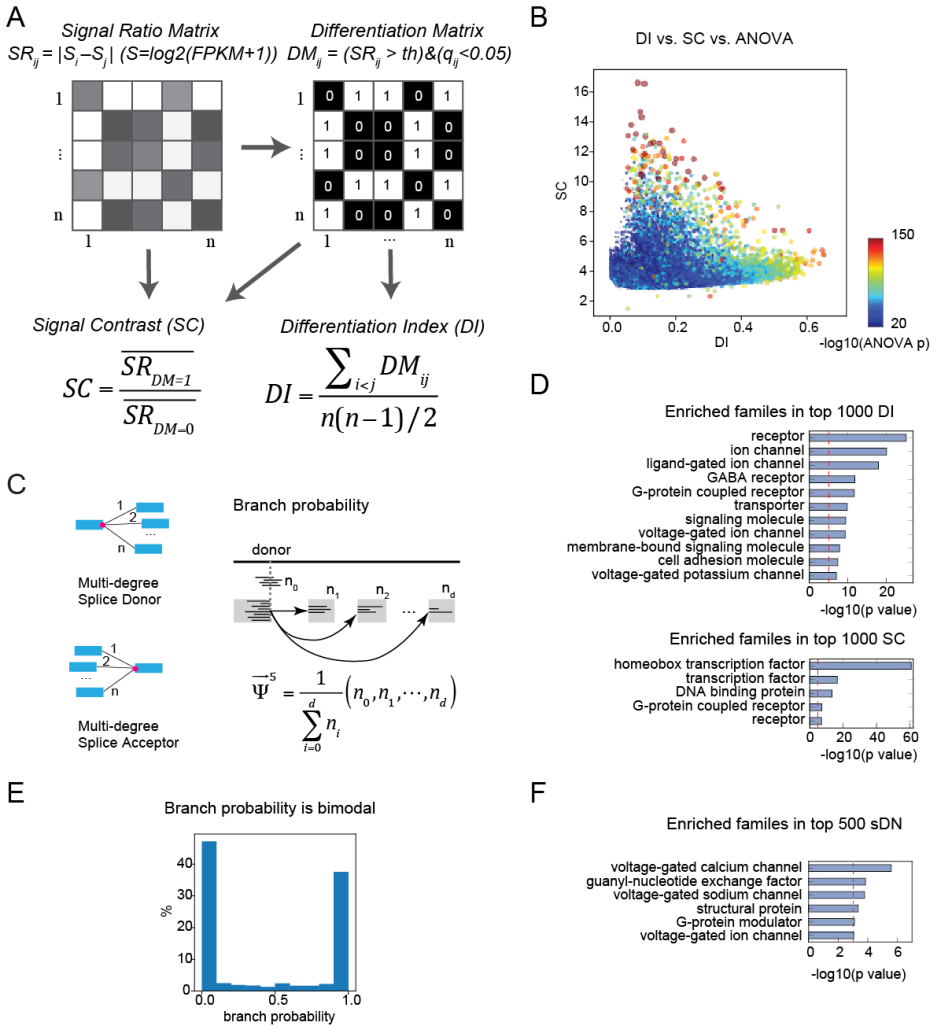


Figure 3. Gene expression metrics related to information content and robustness (A) Expression differences between cell types are compiled into a signal ratio matrix (SR) and binarized into a differentiation matrix (DM) reflecting whether each pair of cell types is distinguished (1) or not (0). The Differentiation Index (DI) is the fraction of nonzero values. The Signal Contrast (SC) is the average expression difference between distinguished pairs divided by the average expression difference between undistinguished pairs. **(B)** Highly significant ANOVA genes (warm colored dots) include a mixture of genes with high SC and low DI and genes with low SC and high DI. **(C)** Definition of generalized PSI (percent spliced in). For a splice donor, a generalized form of PSI (donor branch probability) can be defined as the joint distribution of transition probabilities from the donor to each acceptor. Acceptor branch probability can be defined conversely. **(D)** PANTHER (***Thomas, 2003***) gene families enriched in the top 1000 DI and the top 1000 SC genes. Red lines indicate the $p = 10^{-5}$ threshold used to judge significance. **(E)** Histogram of all donor branch probabilities from alternatively spliced sites. The distribution is highly bimodal, indicating that alternative splicing is "all or none" for each site in each cell type (though often varying between cell types). **(F)** PANTHER gene families enriched in the top 500 DN genes. The number of cell types distinguished by a gene's splice variants (sDN; see Methods for calculation) rather than the ratio (DI) is used since the denominator of DI (total number of cell types potentially distinguished) varies for each gene. This is because genes not expressed in a cell type can contribute to distinctions based on expression, but not to those based on splicing. Red lines indicate the $p = 10^{-3}$ threshold used to judge significance.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on the basis of a particularly high DI (Figure 3C and Figure 4 Supplement 1B; mean DI=0.21; rank 16th) compared to, for example, cyclic nucleotide-gated ion channels (mean 0.31, highest) or GABA receptors (0.29, 2nd). We infer that this is due to the fact that graded expression differences also contribute to DI. Since binary ON/OFF expression patterns may be more critical for cell type specification than graded expression patterns, we calculated a binary version of DI (bDI; see Methods). With this metric, the homeobox TF family is the most enriched PANTHER family among the top 1000 bDI genes (Figure 4 Supplement 1A) and had the 2nd highest average bDI (0.07) among PANTHER families after neuropeptides (0.08) (Figure 4 Supplement 1B). Among TF subfamilies, the LIM domain subfamily of homeobox genes had the highest mean bDI (Figure 4 Supplement 1C), consistent with its known role in specifying spinal cord and brainstem cell types (***Tsuchida et al., 1994***; ***Philippidou and Dasen, 2013***).

The ability of gene families to provide information about cell types is determined by both how informative individual family members are, and the relationships between them. If the information across family members is independent, the overall information is increased relative to the case in which multiple members contain redundant information (Figure 4 Supplement 1D). This aspect of "family-wise" information is not captured by "gene-wise" metrics like mean bDI, or by enrichment analysis (Figure 3C, Figure 4 Supplement 1A-C). One way of capturing the additive, non-redundant information within a gene family is to measure its ability to separate cell types using a distance metric. This analysis (Figure 4E) reveals that homeobox TFs yield the largest distances between cell types. Thus, homeobox TFs provide the best separation of profiled cell types both individually (Figure 4 supplement 1A,B) and as a family (Figure 4E). It has long been known that a subset of homeobox TFs, the HOX genes, play an evolutionarily conserved role in specifying cell types in invertebrates (***Kratsios et al., 2017***; ***Zheng et al., 2015***) and in the vertebrate spinal cord and brainstem (***Dasen and Jessell, 2009***; ***Philippidou and Dasen, 2013***). Our current analyses suggest that the larger family of homeobox TFs play a broader role in transcriptional diversity of cell types across the mammalian nervous system.

In summary, by defining novel metrics DI and SC, we identify homeobox TFs as the most robustly distinguishing family of genes as well as synaptic and signaling genes as the most differentially expressed genes. These two categories of genes drive neuronal diversity by orchestrating cell type-specific patterns of transcription and by endowing neuronal cell types with specialized signaling and connectivity phenotypes.

A compact TF code for neuronal identity

In addition to identifying the most informative transcription factors across the entire set of cell types studied, we also identify the most informative TFs for individual cell types. To accomplish this, we extracted the most compact set of “ON” or “OFF” TFs needed to specify each cell type generating a hierarchy of TFs constituting a decision tree that efficiently classifies cell types (***Gabitto et al., 2016***). At each level of the tree, TFs were chosen to optimally bisect (by their expression level) the set of cell types into two groups that differed maximally from each other in terms of their overall expression profile (assessed within the full transcriptome). To generate a classifier operating at each level of anatomical organization, we favored TFs whose bisected groups are consistent with anatomical divisions (see Methods for details).

The selected TFs included many genes previously implicated as key transcriptional regulators (KTRs) in the development or maintenance of the distinguished cell types. For example, *Foxg1*, which split forebrain from other cell types, is known to be critically required for normal development of the telencephalon (***Xuan et al., 1995***; ***Danesin and Houart, 2012***) and is known to function cell autonomously within the olfactory placode for the production of olfactory sensory neurons, as well as for all other cells in the olfactory lineage (***Duggan et al., 2008***). Similarly, at the next levels, *Tbr1* (***Bedogni et al., 2010***), *Satb2* ***Leone et al.*** (***2014***), *Egr3* (***Chandra et al., 2015***), *Isl1* (***Lu et al., 2013***) and *Emx2* (***Zhang et al., 2016***), are known as KTRs involved in the development and/or maintenance of the relevant cell types, providing significant validation of this method.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

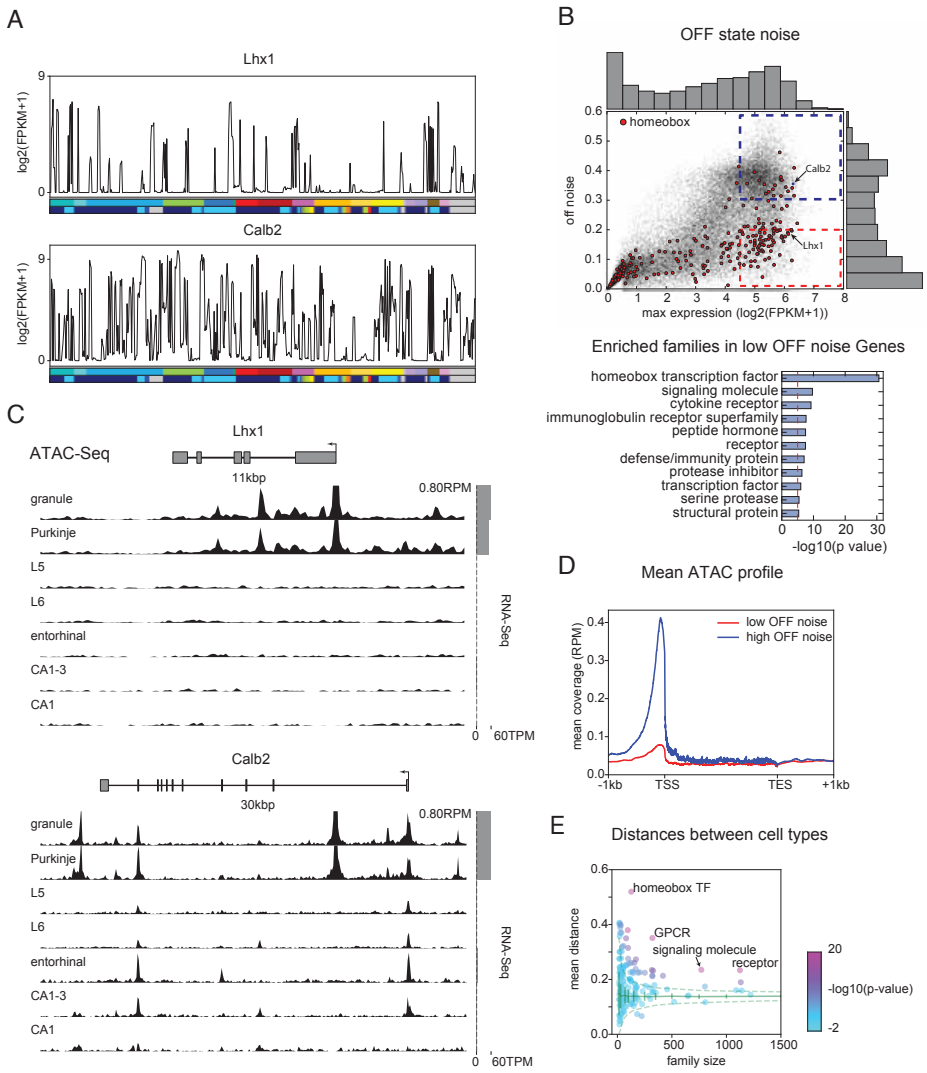


Figure 4. Mechanisms contributing to high information content and low noise of Homeobox TFs. (A) Example expression patterns of a LIM class homeobox TF (*Lhx1*) and a calcium binding protein (*Calb2*) with similar overall expression levels. Cell type legend is given in Figure 1 Supplement 1. **(B)** (upper) OFF state noise (defined as std. dev. of samples with FPKM<1) plotted against maximum expression. (lower) PANTHER families enriched in the region indicated by red dashed lines in the upper panel. **(C)** Average (replicate N=2) ATAC-seq profiles for the genes shown in A. Some peaks are truncated. Expression levels are plotted at right (grey bars). **(D)** Length-normalized ATAC profile for genes with high (> 0.3, blue dashed box in B, n=853) and low (< 0.2, red dashed box in B, n=1643) off state expression noise. **(E)** Mean separability of cell types for PANTHER families. Separability is a measure of gene expression distance (defined as the average of 1- Pearson's corr. coef.) calculated across a set of genes. Since dispersion of separability decreases with family size, results are compared to separability calculated from randomly sampled groups of genes (green solid lines: mean and std. dev.; green dashed lines: 99% confidence interval). Z-scores: homeobox TF: 17.4, GPCR: 16.1, receptor: 13.1 and signaling molecule: 11.2.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The TF code identified for each cell type is not unique. First, there are additional TFs that are consistent with the tree (see Supplementary Table 3). Second, past the first level (*Foxg1*), TFs may be expressed outside of the cell types shown and so could contribute to encoding other expression differences. More generally, the details of the tree may depend on the precise procedure used to extract it. We explored variant procedures that better preserved the known anatomical and developmental relationships between cell types (Figure 5 Supplement 1) as well as procedures that made no assumptions about these relationship whatsoever (Figure 5 Supplement 2). Interestingly, in each case, the majority of the same genes were identified, suggesting they encode cell type information that is robust to the precise methods used to extract them.

Although the decision tree classifier identifies many known KTRs, it also suggests hypotheses about less studied genes. For example, *Tox2* has received little prior study in the CNS, although it has recently been identified and replicated as a locus of heritability for Major Depressive Disorder (***Zeng et al., 2016***). Based on its position in the tree, we hypothesize that *Tox2* is a KTR of midbrain, hypothalamic and hindbrain cell types, including dopaminergic and serotonergic cell types in these regions, although its expression in other cell types may also contribute. Hence the tree of identified TFs is a robust and rich source of novel hypotheses about transcriptional regulation in genetically identified cell types. Known and hypothesized KTRs identified by the decision tree classifier are tabulated in Supplementary Table 3.

Long genes contribute disproportionately to neuronal diversity

We found that neuronal effector genes such as ion channels, receptors and cell adhesion molecules have the greatest ability to distinguish cell types (highest DI; Figure 3C). Previously, these categories of genes have been found to be selectively enriched in neurons and to share the physical characteristic of being long (***Sugino et al., 2014***; ***Gabel et al., 2015***; ***Zylka et al., 2015***). Consistent with this, DI is strongly biased toward long gene length (Figure 6A). Interestingly, the expression of long genes is not uniform across brain regions, but is highest in the evolutionary newer forebrain and is lower in the older brainstem and hypothalamus (Figure 6B). Non-neuronal cell types expressed only 1/2 to 1/5 as many long genes as neuronal cell types (blue bars in Figure 6B). This was true even for non-dividing cell types like myocytes and largely non-dividing tissues like the heart (separate data not shown). Hence long genes, which are preferentially expressed in neurons, also contribute most to the differential expression between neuronal cell types.

REST is an important zinc-finger transcription factor that represses expression of neuronal genes in non-neurons (***Chong et al., 1995***; ***Schoenherr and Anderson, 1995***). We wondered if REST preferentially targets long genes. To assess the magnitude of this effect and its influence on the length distribution of neuronal genes (Figure 6 Supplement 1A), we plotted the length-dependence of genes containing RE1/NRSE elements (Figure 6 Supplement 1B) and observed that they are indeed biased toward long genes. When these REST targets are removed from neuronally expressed genes, the length distribution of expressed genes looks similar to that of non-neurons (Figure 6 Supplement 1C). However, consistent with the fact that only 8.6% of neuronally expressed genes are REST targets (contain an NRSE), the removal of these genes has only a modest effect on the length distribution of DI (Figure 6 Supplement 1D). Therefore, although REST targets are long, many other long genes also contribute to neuronal diversity.

Long genes differ from more compact genes primarily in the number and length of their introns, which, for the longest genes, comprise all but a few percent of their length (Figure 6 Supplement 1E). Introns often contain *cis* regulatory elements that regulate transcription, splicing and other aspects of gene expression. Could these longer introns increase the regulatory capacity of long genes? In order to determine whether or not the introns of long genes have enhanced regulatory capacity, we identified candidate regulatory elements as sites of enhanced genome accessibility using our ATAC-seq data. As expected, long genes had more candidate regulatory elements (ATAC peaks; Figure 6 Supplement 1F) and these peaks were present in a greater number of distinct patterns per gene across cell types (Figure 6C,D). Consistent with the hypothesized role in differential expression,

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

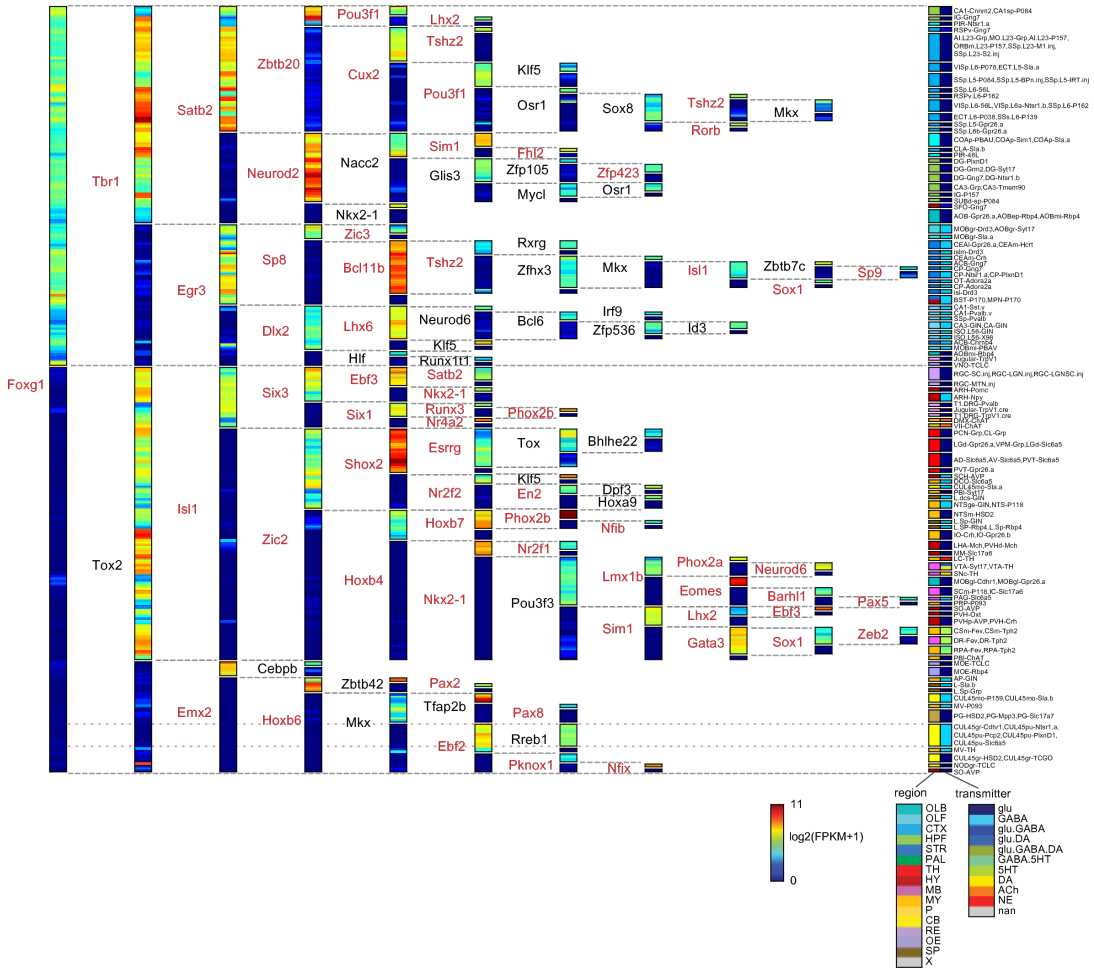


Figure 5. A compact TF code. A decision tree classifier constructed from the most informative TFs for profiled cell types. Cell types are bisected at each node by TF expression level, (color scale). Each cell type can be specified by the "ON" (warm colors) or "OFF" (cool colors) expression of 4 to 11 TFs as indicated. For example, Purkinje cells (yellow-light blue group near the right bottom corner, consisting of CUL4,5gr-Cdhr1, CUL4,5pu-Pcp2, etc.) have a code which can be read from left to right within the red dotted lines, consisting of: Foxg1(OFF)-Tox2(OFF)-Emx2(OFF)-Hoxb6(OFF)-Mkx(OFF)-Ebf2(ON)-Rreb1(ON). Blue dashed lines mark positions of ON/Off transitions for each TF.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the number of unique patterns correlated well with the degree of differential expression across cell types (Figure 6E). Hence long genes have enhanced regulatory capacity that correlates with their enhanced contribution to neuronal diversity.

To compare candidate regulatory elements in long genes between neurons and non-neurons, we used publicly available DNase-seq data from the ENCODE project (***Dunham et al., 2012***). We found a significantly higher number of open chromatin sites in brain compared to non-brain tissue. This bias was particularly pronounced in forebrain, and was stronger in human than in mouse tissue (Figure 6 Supplement 1G-J). Together, these data support the hypothesis that neuronal genes may have increased in length over evolutionary time in part to support more complex and nuanced regulatory regimes.

To assess the relative contribution of long (≥100kbp) and short (<100kbp) genes, we first calculated averages of "gene-wise" metrics (Figure 6F). Signal contrast is comparable between these two groups of genes, but, for all other metrics (DI, bDI, sDN; for sDN see Figure 3 D,E and Methods), averages for long genes are about twice that of short genes. Enhanced alternative splicing of long genes (high sDN) is readily understandable from the increased number of alternative splice sites in long genes (Figure 6 Supplement 1K). Although there is no significant difference in SC between long and short genes, low OFF noise genes (Figure 4B-D) are significantly shorter than high OFF noise genes (Figure 6 Supplement 1L).

To assess the "group-wise" contribution (akin to the "family-wise" analysis of Figures 3C,F and 4E), we first observed that both groups are fairly decorrelated between member genes (Figure 6 Supplement 1M). Despite similar decorrelation, the distances between cell types based on long gene expression are larger than those obtained from expression of short genes (Figure 6G). Thus, long genes, as a group, contribute more than short genes to neuronal diversity.

TE insertions elongate genes and carry regulatory information

The above results indicate that gene length is an important contributor to gene expression diversity across cell types. Gene lengths differ widely across species (Figure 7A and Figure 7 Supplement1A), suggesting genes are elongated during evolution. In fact, evolutionary older genes are longer (Figure 7 Supplement 1B; ***Grishkevich and Yanai*** (***2014***)). To better understand mechanisms of gene elongation over mammalian evolution, we examined segments inserted into the human and mouse genomes by comparing them to closely related species (Figure 7B). Plotted in Figure 7B (left) is a histogram of the lengths of the segments inserted into human (see also ***Mikkelsen et al.*** (***2005***). Two clear peaks are recognizable, corresponding to Alu and L1 repeats. Moreover, around 92% of the base pairs of the inserted segments overlap with known repeats (Figure 7B inset; ***Bao et al.*** (***2015***)). Similar results are observed in the mouse genome (Figure 7B right; see also ***Pozzoli et al.*** (***2007***)). These comparisons indicate that genes are elongated by transposable element (TE) insertions.

Since long genes have a greater number of candidate regulatory elements, as indicated by more ATAC-peaks, we asked whether these can originate from mobile elements. As shown in Figure 7C, 56% of the ATAC peaks overlap known repeats and this number increases to 75% when only newly inserted segments are considered, indicating that TEs may carry regulatory functions. To explore the possibility that TE/repeats contribute to global regulation of neuronal gene expression, we fit gene expression levels with counts of individual repeats within and surrounding each gene (Figure 7D). The R^2 values for each cell type calculated using test genes (20%) not used for fitting (Figure 7E, blue) are much larger than expected by chance (Figure 7E, green/red/orange). If counts and genes are shuffled (green) cross validated R^2 values drop below 0. However, if the length of the gene is retained in the shuffling control (orange, red) the R^2 values drop to about 1/3 of those in the original fitting. This reflects the fact that gene length is highly correlated with expression (Figure 7 Suplement 1C; c=0.418: mean Pearson's r between log gene length and expression rank) and some repeats, such as SINEs, are highly correlated with both gene length (c=0.841) and expression (Figure 7 Supplement 1C; mean c=0.454). We also varied the size and position of the regions used to count

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

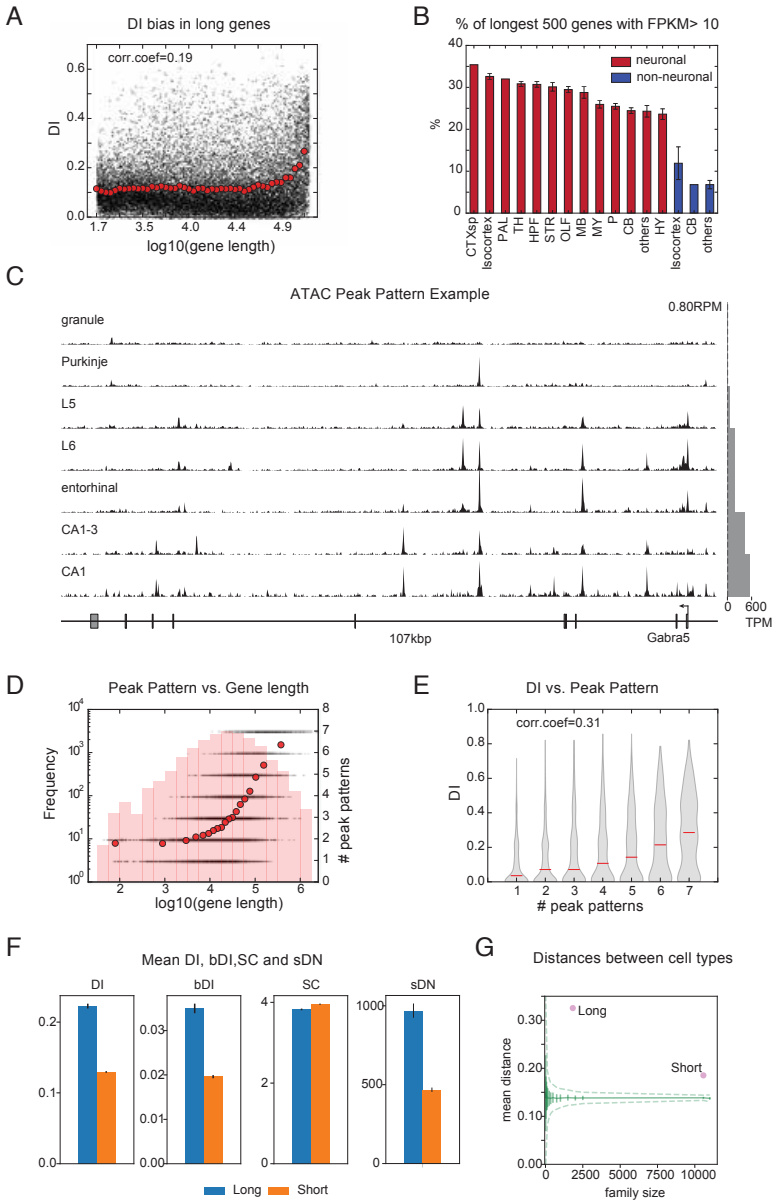


Figure 6. Long genes have a greater capacity for differential expression. (A) Black dots: DI of each gene is plotted against sorted gene length. Red dots: binned average of DIs (1000 genes per bin, sorted by length). **(B)** Fraction of the longest 500 genes expressed within each brain region profiled for neuronal (red bars) and non-neuronal cell types (blue bars). **(C)** ATAC-seq peaks for *Gabra5* showing different patterns of peaks for each of 7 cell types. Scale (top right) in reads per million. Expression levels for each cell type are shown at right (gray bars). **(D)** Black dots: number of distinct peak patterns observed across 7 ATAC-seq profiled cell types plotted against the gene length for each gene; 7 corresponds to a distinct pattern for each profiled cell type. Red dots: binned averages of black dots as in panel A. Background histograms show numbers of genes in each length bin. **(E)** Violin plot showing the relationship between DI and the number of different patterns of ATAC-seq peaks. Corr.coef. (0.31) is greater than that between DI and gene length (0.19; panel A). **(F)** Average metrics for long (≥100kbp) and short (<100kbp) neuronal genes (reproducibly expressed in neuronal cell types). **(G)** Separability of cell types calculated as in Figure 4E, but using long neuronal genes and short neuronal genes rather than functionally defined gene families. Z-score is 33.2 for long and 22.1 for short neuronal genes. Both are highly different from randomly sampled genes (green solid lines mean and Std. dev.; dashed lines = 99% confidence interval), but long genes provide greater separation.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

repeats and found that predictions about expression (R^2) were best when including the gene body and the adjacent $10 \sim 50$kbp. (Figure7 Supplement 1D,E). There are several prior examples relating TE regulation of gene expression (e.g. ***Han et al.*** (***2004***); ***Chuong et al.*** (***2016***)), however, this is the first to show the existence of a global network of TEs affecting gene expression in neurons.

In summary, genes are elongated by insertions of TEs which overlap candidate regulatory elements, and are predictive of relative gene expression levels, suggesting they may increase the capacity of long genes to be differentially expressed.

Discussion

A Resource of Neuronal Cell type specific Transcriptomes

The dataset presented here is the largest collection of cell type-specific neuronal transcriptomes obtained by RNA-seq (Table 1) and so offers the broadest view to date of the transcriptional basis of neuronal diversity. Prior RNA-seq data from sorted cells have been focused primarily on what distinguishes neurons as a class from other brain cell types (***Zhang et al., 2014b***), or have focused on a limited number of brain regions, such as the somatosensory cortex, hippocampus (***Zeisel et al., 2015***; ***Cembrowski et al., 2016***; ***Tasic et al., 2016***) and retina (***Macosko et al., 2015***). Our strategy of profiling labeled populations of ~ 100 cells is intermediate between single cell profiling, which can be limited by the noisiness of single cell assays (***Marinov et al., 2013***) and tissue profiling, which cannot resolve the heterogeneity of component cell types (***Nelson et al., 2006***). This approach enabled us to obtain highly sensitive and reproducible transcriptomes from genetically accessible target populations. The wide range of cell types in the dataset is suitable for addressing general questions regarding neuronal identity and diversity, but at the same time, the fact that each transcriptome corresponds to a genetically (or retrogradely) labeled population, allows investigation of the same population of the cells across time and labs in order to address more specific questions about those cell types.

We developed a quantitative approach for comparing cell type profiles across multiple studies using NNLS decomposition. The results reveal multiple cases in which pooled cell profiles mapped to more than one SCRS profile. It is likely that at least some of these cases represent biologically distinct cell types that share a genetic marker (like subtypes of Pvalb interneurons). However, in most of these cases, the SCRS clusters were barely separable, and the two SCRS studies available for comparison did not agree. Given the complimentary advantages of improved reproducibility, separability and deeper depth of sequencing afforded by the pooling approach, and of reduced heterogeneity afforded by the SCRS approach, it is likely that further integration of these approaches with other modalities, such as FISH (***Moffitt et al., 2016***) will be needed to accurately catalog the full census of brain cell types.

A transcriptional code for neuronal diversity

We developed novel, easily calculated metrics that capture essential features of the robustness and information content of transcriptome diversity. These measures are not cleanly captured by traditional variance-based metrics like ANOVA and CV (Figure 3 Supplement 1). We found that the homeobox family of TFs exhibited the most robust (high SC) expression differences across cell types (Figure 3D bottom). These ON/OFF differences were characterized by extremely low expression in the OFF state (Figure 4A-D). Mechanistically, the low expression was associated with reduced genome accessibility measured by ATAC-seq (Figure 4C,D), presumably reflecting epigenetic regulation, known to occur for example at the clustered Hox genes via Polycomb group (PcG) proteins (***Montavon and Soshnikova, 2014***). Although this regulation has been studied most extensively at Hox genes, genome-wide ChIP studies reveal that PcG proteins are bound to over 100 homeobox TFs in ES cells (***Boyer et al., 2006***). Our results indicate that strong cell type-specific repression persists in the adult brain. Presumably this represents the continued functional importance of preventing even partial activation of inappropriate programs of neuronal identity.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

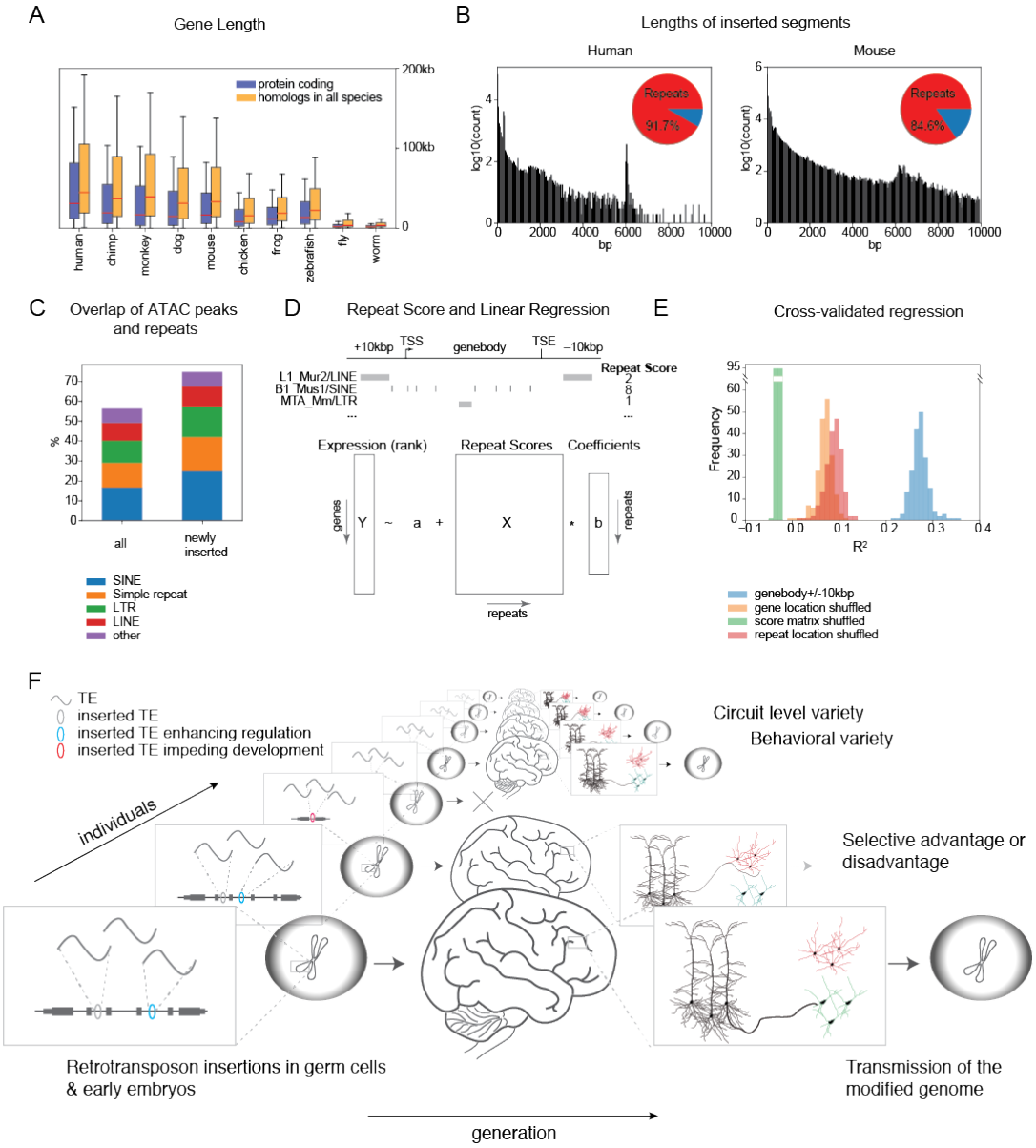


Figure 7. Genes are elongated by TE insertions and TEs contain information for gene expression (A) Distribution of gene length for various well annotated species. Red lines indicate means and whiskers indicate inter-quartile range. Blue bars are all protein coding genes and yellow bars are the subset of genes with homologs in all species. (human: *Homo sapiens*; chimp: *Pan troglodytes*; monkey: *Macaca mulatta*, mouse: *Mus musculus*; dog: *Canis lupus familiaris*; chicken: *Gallus gallus*; frog: *Xenopus tropicalis*; zebrafish: *Danio rerio*; fly: *Drosophila melanogaster*; worm: *Caenorhabditis elegans*) **(B)** Histograms of lengths of segments inserted into the human genome compared to chimp (left) and mouse genome compared to rat (right). Peak near 300bp (more visible in human) corresponds to Alu, and near 6000bp corresponds to LINE. Pie charts (insets) indicate fraction of inserted bp overlapping transposable elements (TE) and other types of repeats. Gorilla and Guinea pig are used as surrogates of common ancestors of human and chimp, and mouse and rat, respectively (see Methods). **(C)** Percentage of ATAC peaks overlapping major categories of repeat elements. Left side: all ATAC peaks, right side: ATAC peaks overlapping recently inserted segments calculated in (B). **(D)** Schema describing repeat score and regression model. Repeat scores (upper panel) are calculated separately for each type of repeat element and for each gene as the count of that element in the specified interval determined by the gene. Regressions (lower panel) are calculated separately for each cell type by fitting coefficients (b) to ranked expression levels (Y) using intercept(a) and repeat score (X). **(E)** Fits to 80% of the genes are cross validated using the remaining 20%. Histograms show cross validated R^2 for each cell type (blue), and for controls shuffling the relationship between repeat scores and genes(score matrix; green) or changing the repeat score by randomly changing the location of repeats (red) or by calculating the repeat score over a randomly selected genomic interval of the same length as the gene (orange). The latter two shuffling methods retain some predictive value compared to shuffling the repeat score matrix (green) since they maintain the correlation between gene length and expression (See Figure 7 Supplement 1C). **(F)** A model of how neuronal genes become elongated over evolutionary time scales.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As a group, homeobox TFs distinguished 98% of neuronal cell types profiled. Historically, homebox TFs are well known to combinatorially regulate neuronal identity in Drosophila and C. elegans (***Kratsios et al., 2017***) and the vertebrate brainstem and spinal cord (***Dasen and Jessell, 2009***; ***Philippidou and Dasen, 2013***). The continued expression of homeobox TFs throughout the adult mammalian nervous system suggests that they likely also contribute to the maintenance of neuronal identity.

In order to reveal the relationship between specific cell types and TFs, we constructed a TF decision tree for classifying profiled cell types. As expected from their high information content, homeobox TFs figured prominently in this list (49/127). Many of the identified factors are known to be key transcriptional regulators of the cell types in which they continue to be expressed (Supplemental Table 3). In most cases it is not known whether or not these roles occur only in development, or are also important for the maintenance of neuronal identity. Lists of expressed TFs and the genetically accessible cell types in which they are expressed provide a ready source of testable hypotheses about how cell type specific transcriptional identity is maintained in the adult nervous system.

Long genes shape neuronal diversity

Our study suggests that long genes contribute disproportionately to neuronal diversity (Figure 6A,F,G). Increases in the number of alternative start and splice sites present in longer genes increase neuronal diversity (Figure 6F), but in addition, we hypothesize that longer genes have a larger number of regulatory elements that alter expression and enhance differential usage of these alternative sites. Long genes likely elongate during evolution, via insertions of TEs in their introns (Figure 7A,B; ***Sela et al., 2007***; ***Grishkevich and Yanai, 2014***). Long neuronal genes, such as ion channels and cell adhesion molecules, may be expressed primarily late in development (***Okaty et al., 2009***). Developmentally later and more spatially and cell-type restricted expression of neuronal genes may make mammalian genomes more tolerant to mutations caused by the insertion of TEs in these genes. Conversely, genes such as Hox genes, which are critical for early development, and are often expressed in progenitors giving rise to many cell types, are remarkably TE impoverished (***Chinwalla et al., 2002***; ***Simons, 2005***). TE insertions occurring randomly are expected to happen more frequently in long genes (Figure 7F, Figure 7 Supplement 1F,G), thereby accelerating their elongation over time.

Here we provide evidence supporting the hypothesis that evolution of the vertebrate nervous system may have taken advantage of TE insertions and subsequent exaptations to diversify neuronal cell types, increasing the complexity of brain circuits. Long genes are enriched in the signaling molecules, receptors and ion channels responsible for input/output transformations in neurons, and the cell adhesion molecules that specify neuronal connectivity. Thus, changes in their expression could lead to changes in circuit level function. Specifically, elongation of long genes through TE insertions, occurring in the early embryo or in germ cells, likely creates a reservoir of genetic elements providing fodder for regulatory innovation. Subsequent exaptation of a fraction of these elements may have enhanced cell type-, and hence, behavioral- diversity, in turn, increasing the ability of populations to adapt to their environment (Figure 7F). This evolutionary advantage of lengthening neuronal genes may help to explain the paradox of why long genes should be abundantly expressed in CNS neurons despite the fact that these genes are sites of genome instability associated with genetic lesions leading to autism and other developmental disorders (***Wei et al., 2016***). This hypothesis also shifts focus away from short, developmental time scales considered in other hypotheses linking TE insertion to neuronal function (***Muotri et al., 2005***; ***Richardson et al., 2014***; ***Perrat et al., 2013***). Instead of DNA rearrangements in neuronal progenitors producing neuronal diversity, we consider the time scales of evolution and thus also shift focus to the germ line, where natural selection has its influence.

In summary, the elongation of neuronal effector genes may have endowed them with increased capacity for differential expression, permitting enhanced neuronal diversity. This diversity can also

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

be characterized in terms of expression patterns of homeobox and other TFs. The maintenance of diverse neuronal identities must require interactions between expressed TFs and accessible *cis* regulatory elements within target effector genes. Identifying these interactions will require manipulating them within genetically identified cell types.

Methods and Materials

Cell Types and Mouse Lines

Cell types are defined operationally by the intersection of a transgenic mouse strain (or in some cases anatomical projection target) and a brain region. These "operational cell types" may or may not correspond to "atomic" cell types, but as shown in Figure 2 have comparable purity to clusters of single cells. Mouse lines profiled in this study are summarized in Supplementary Table 1. Most were obtained from GENSAT (***Gong et al., 2007***) or from the Brandeis Enhancer Trap Collection (***Shima et al., 2016***). For Cre-driver lines, the Ai3, Ai9 or Ai14 reporter (***Madisen et al., 2009***) was crossed and offspring hemizygous for Cre and the reporter gene were used for profiling. All experiments were conducted in accordance with the requirements of the Institutional Animal Care and Use Committees at Janelia Research Campus and Brandeis University.

Atlas

Animals were anesthetized and perfused with 4% paraformaldehyde and brains were sectioned at 50μm thickness. Every fourth section was mounted on slides and imaged with a slide scanner equipped with a 20x objective lens (3DHISTECH; Budapest, Hungary). In house programs were used to adjust contrast and remove shading caused by uneven lighting. Images were converted to a zoomify compatible format for web delivery and are available at http://neuroseq.janelia.org.

Cell Sorting

Manual cell sorting was performed as described (***Hempel et al., 2007***; ***Sugino et al., 2014***). Briefly, animals were sacrificed following isoflurane anesthesia, and 300μm slices were digested with pronase E (1mg/ml, P5147; Sigma-Aldrich) for 1 hour at room temperature, in artificial cerebrospinal fluid (ACSF) containing 6,7-dinitroquinoxaline-2,3-dione (20μM; Sigma-Aldrich), D-(–)-2-amino-5-phosphonovaleric acid (50μM; Sigma-Aldrich), and tetrodotoxin (0.1μM; Alomone Labs). Desired brain regions were micro-dissected and triturated with Pasteur pipettes of decreasing tip size. Dissociated cell suspensions were diluted 5-20 fold with filtered ACSF containing fetal bovine serum (1%; HyClone) and poured over Petri dishes coated with Sylgard (Dow Corning). For dim cells, Petri dishes with glass bottoms were used. Fluorescent cells were aspirated into a micropipette (tip diameter 30-50μm) under a fluorescent stereomicroscope (M165FC; Leica), and were washed 3 times by transferring to clean dishes. After the final wash, pure samples were aspirated in a small volume (1~3μl) and lysed in 47μl XB lysis buffer (Picopure Kit, KIT0204; ThermoFisher) in a 200μl PCR tube (Axygen), incubated for 30min at 40°C on a thermal cycler and then stored at -80°C. Detailed information on profiled samples are provided in Supplementary Table 2.

RNA-seq

Total RNA was extracted using the Picopure kit (KIT0204; ThermoFisher). Either 1μl of 10^{-5} dilution of ERCC spike-in control (#4456740; Life Technologies) or (number of sorted cells/50) * (1μl of 10^{-5} dilution of ERCC) was added to the purified RNA and speed-vacuum concentrated down to 5μl and immediately processed for reverse transcription using the NuGEN Ovation RNA-Seq System V2 (#7102; NuGEN) which yielded 4~8μg of amplified DNA. Amplified DNA was fragmented (Covaris E220) to an average of ~200bp and ligated to Illumina sequencing adaptors with the Encore Rapid Kit (0314; NuGEN). Libraries were quantified with a KAPA Library Quant Kit (KAPA Biosystems) and sequenced on an Illumina HiSeq 2500 with 4 to 32-fold multiplexing (single end, usually 100bp read length, see Supplemental Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RNA-seq analysis

Adaptor sequences (AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC for Illumina sequencing and CTTTGTGTTTGA for NuGEN SPIA) were removed from de-multiplexed FASTQ data using cutadapt v1.7.1 (http://dx.doi.org/10.14806/ej.17.1.200) with parameters "–overlap=7 –minimum-length=30". Abundant sequences (ribosomal RNA, mitochondrial, Illumina phiX and low complexity sequences) were detected using bowtie2 (***Langmead and Salzberg, 2012***) v2.1.0 with default parameters. The remaining reads were mapped to the UCSC mm10 genome using STAR (***Dobin et al., 2012***) v2.4.0i with parameters "–chimSegmentMin 15 –outFilterMismatchNmax 3". Mapped reads are quantified with HTSeq (***Anders et al., 2014***) using Gencode.vM13 (***Harrow et al., 2012***).

Pan-neuronal genes

Pan-neuronal genes are extracted as satisfying the following conditions: 1) mean neuronal expression level (NE)> 20 FPKM, 2) minimum NE > 5 FPKM, 3) mean NE > maximum non-neuronal expression level (NNE), 4) minimum NE > mean NNE, 5) mean NE > 4x mean NNE, 6) mean NE > mean NNE + 2x standard deviation of NNE, 7) mean NE − 2x standard deviation of NE > mean NNE.

DI/SC/DN calculation

To calculate DI, the following criteria were used to assign a "1" or "0" to each element in the difference matrix (DM): log fold change > 2 and q-value <0.05. Q-values were calculated using the limma package including the voom method (***Law et al., 2014***). To adjust the power to be similar across cell types, two replicates (the most recent two) are used for all cell types with more than two replicates. We have tried the same calculations with 3 replicates (using a fewer number of cell types) and obtained similar results (data not shown).

To calculate binary DI (bDI), the following DM criteria were used: expression levels of all the replicates in one of the cell types in the pair < 1FPKM and expression levels of all the replicates in the other cell type in the pair > 15FPKM, in addition to q-value <0.05.

To assess the extent of differentiation by alternative splicing, we calculate differentiation at the level of each splice branch. See Figure 3D for the definitions of a splice branch and of branch probability. For each branch, at each alternative splice site, we define each pair of cell types as "different" when 1) branch probabilities for all replicates in a group are less than 0.3 or greater than 0.7, and 2) both cell types in the pair have > 10 reads reads at the alternative site. Condition 1) is justified by the bimodal distribution of branch probabilities shown in Figure 3E. Accumulating over all pairs creates a DM for each branch. We then combine all the branches using a logical "OR" to create a gene-level DM for each gene. If any branch distinguishes a pair of cell types, that pair is called "different" at the gene level. The gene-level DM has a value of "1" for pairs of cell types distinguished by any of the branches belonging to that gene, and has a value of "0" for pairs of cell types not distinguished by any branch belonging to the gene. The number of pairs compared can differ, depending on the expression pattern of the gene, since branch probabilities can only be calculated for cell types that express the gene. This situation differs from that for DI or bDI (based on expression levels rather than splicing) since pairs of cell types can be distinguished even if one does not express the gene. Therefore, unlike DI and bDI which assume a fixed number of total pairs, we use DN (total number of pairs distinguished), rather than the fraction of pairs distinguished, to rank genes.

NNLS/Random forest decomposition

SCRS datasets deposited in NCBI GEO (GSE71585, ***Tasic et al.*** (***2016***); and GSE60361, ***Zeisel et al.*** (***2015***)) were used for NNLS decomposition. Specifically the deposited count data were converted to TPM and used for comparison. The NeuroSeq dataset was quantified using RefSeq and featurecount (***Liao et al., 2013***) and converted into TPM. Subsets of genes common to all three datasets are then used for all further analyses. Since distributions of TPM values differed between datasets, they were quantile normalized to an average profile generated from the NeuroSeq dataset. Since most genes

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in the SCRS profiles exhibited noisy expression patterns, using the entire gene set for decomposition is not feasible. Therefore, we selected for decomposition the genes deemed most informative for distinguishing cell classes based on ANOVA across cell classes. However, simply taking the top ANOVA genes lead to highly biased gene selection since some cell types exhibited much larger transcriptional differences than others (e.g. many ANOVE selected genes were specific to microglia). We therefore selected genes so as to minimize the overlap between the cell types distinguished. Beginning with the highest ANOVA gene (highest ANOVA F-value), genes were selected only if their DM (Differentiation Matrix defined in Figure 3) differed from those previously selected, defined with a Jaccard index threshold of 0.5. We chose 300 genes from each dataset, yielding a total of 563 genes when all three sets were combined. This gene set was then used for all decompositions. Decompositions were performed on average profiles created by summing NeuroSeq replicates or by summing single-cell profiles using cluster assignments provided by the authors. NNLS was implemented using the Python scipy library (http://www.scipy.org).

For Random forest, implementation in the Python scikit-learn library (***Pedregosa et al., 2011***) was used.

ATAC-seq

7 cell types, Purkinje and granule cells from cerebellum, excitatory layer 5, 6 and entorhinal pyramidal cells from cortex, excitatory CA1, or CA1-3 pyramidal cells from hippocampus, labeled in mouse lines P036, P033, P078, 56L, P038, P064, and P036 respectively (all from ***Shima et al., 2016***) were profiled with ATAC-seq. They were FACS sorted to obtain ∼20,000 labeled neurons. ATAC libraries for Illumina next-generation sequencing were prepared in accordance with a published protocol (***Buenrostro et al., 2013***). Briefly, collected cells were lysed in buffer containing 0.1% IGEPAL CA-630 (I8896, Sigma-Aldrich) and nuclei pelleted for resuspension in tagmentation DNA buffer with Tn5 (FC-121-1030, Illumina). Nuclei were incubated for 20-30 min at 37°C. Library amplification was monitored by real-time PCR and stopped prior to saturation (typically 8-10 cycles). Library quality was assessed prior to sequencing using BioAnalyzer estimates of fragment size distributions looking for a ladder pattern indicative of fragmentation at nucleosome intervals as well as qPCR to determine relative enrichment at two housekeeping genes compared to background (specifically the TSS of *Gapdh* and *Actb* were assessed relative to the average of three intergenic regions). For sequencing, Illumina HiSeq 2500 with 2 to 4-fold multiplexing and paired end 100bp read length was used. In addition to ATAC-seq, RNA-seq was performed on replicate samples of ∼2,000 cells collected in a similar way, and library prepared using the same method described above.

ATAC-seq analysis

Nextera adaptors (CTGTCTCTTATACACATCT) were trimmed from both ends from de-multiplexed FASTQ files using cutadapt with parameters "-n 3 -q 30,30 -m 36". Reads were then mapped to UCSC mm10 genome using bowtie2 (***Langmead and Salzberg, 2012***) with parameters "-X2000 –no-mixed –no-discordant". PCR duplicates were removed using Picard tools (http://broadinstitute.github.io/picard, v2.8.1) and reads mapping to mitochondrial DNA, scaffolds, and alternate loci were discarded. BigWig genomic coverage files were generated using bedtools (***Quinlan and Hall, 2010***) and scaled by the total number of reads per million. For reproducible peaks, liberal peaks were called using HOMER (v4.8.3) (***Heinz et al., 2010***) with parameters "-style factor -region -size 90 -fragLength 90 -minDist 50 -tbp 0 -L 2 -localSize 5000 -fdr 0.5" and filtered using the Irreproducibility Discovery Rate (IDR) in homer-idr (http://github.com/karmel/homer-idr.git) with parameters "–threshold 0.05 –pooled-threshold 0.0125". Peak counts and peak patterns were then quantified using bedtools.

TF Tree

The set of mouse TFs was constructed by combining 4 curated TF lists: genes annotated in 1) PANTHER (***Thomas, 2003***) PC00218 (transcription factor), 2) Riken Transcription Factor Database (***Kanamori et al., 2004***), 3) HUGO (***Gray et al., 2014***) families with TF functions and 4) Gene Ontology

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

(***Ashburner et al., 2000***) GO:0006355 (regulation of transcription). Genes appearing reproducibly in these list (i.e. in more than 1 list) were used as TFs. Anatomical regions used as constraints are defined in a hierarchical manner (see Supplementary Table 5).

The TF tree is constructed recursively using the following algorithm:

```
preparation:
    0. calculate bDIs for all subsets of samples defined by anatomical regions
function bisect(list of samples):
    1. if the list of samples consists of only one cell type, exit
    2. calculate bDI,SC within this group of samples for all TFs
    3. if there is no TF with bDI>0,  exit
    4. find the appropriate level in the hierarchy of anatomical regions
    5. penalize bDIs (from 2.) with bDIs of containing anatomical regions (from 0.)
    6. sort TFs by their penalized bDI and SC in descending order
    7. set candidates as TFs with penalized bDI>0.2, if there are none, take the top 5
    8. for each candidate, calculate divisions of samples according to expression level
       - at sample level, assign ON/OFF using FPKM=3 as threshold
       - at cell type level, assign ON/OFF according to dominant ON/OFF of samples
       - divide all cell types into ON or OFF groups
       - optionally constrain division to anatomical boundary
    9. if there is no division, exit
    10. if there is more than one division then
       - calculate "division strength" for all divisions:
           - a0 = mean number of binary distinctions of all genes between ON and OFF groups
           - a1 = mean number of binary distinctions of all genes within ON or OFF groups
           - division strength = a0/a1
       - then choose the division with the highest division strength
    11. output ON/OFF groups and corresponding TF(s) for the chosen division
    12. call bisect on ON group samples
    13. call bisect on OFF group samples
```

Inserted segments

The multiz alignments downloaded from the UCSC genome browser (***Kent et al., 2002***) was used to calculate inserted segments in human or mouse. By comparing closely related species (human vs. chimp or mouse vs. rat), candidate segments inserted into human (or mouse) are extracted. By using another closely related species as a common ancestor (gorilla, guinea pig respectively for human/chimp and mouse/rat), segments absent in chimp and gorilla (or absent in rat/guinea pig) are called insertion in human (or mouse), and segments absent in chimp but present in gorilla (or absent in rat but present in guinea pig) are called deletion in chimp (or rat).

TE fitting

Repeat annotations for mouse mm10 genome as detected by RepeatMasker (***Smit et al., 2013-2015***) with Repbase (ver. 20140131 ***Bao et al., 2015***) were used. Only repeat families with number of instances>200 are included. For individual repeats, only those with number of instances>50 are included. For repeats in the "Simple repeat" class, only those with number of instances>1000 are included. Repeat scores are calculated as described in Figure 7D using Gencode.vM13. Only genes with non-zero repeat scores are used for fitting. For fitting expression level (rank) by repeat score, a regularized version of linear regression, Ridge regression, was implemented in the Python scikit-learn library (***Pedregosa et al., 2011***).

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tissue data

In addition to cell type-specific data obtained in this study, we analyzed publicly available RNA-seq and DNase-seq data using tissue samples. Information on these samples are described in Supplementary Table 4.

Annotations

For reference annotations we used Gencode.vM13 (***Harrow et al., 2012***) downloaded from http://www.gencod NCBI RefSeq (***Pruitt et al., 2013***) downloaded from the UCSC genome browser.

Anatomical Region Abbreviations

Region abbreviations: AOBmi, Accessory olfactory bulb, mitral layer; MOBgl, Main olfactory bulb, glomerular layer; PIR, Piriform area; COAp, Cortical amygdalar area, posterior part; AOBgr, Accessory olfactory bulb, granular layer; MOBgr, Main olfactory bulb, granular layer; MOBmi, Main olfactory bulb, mitral layer; VISp, Primary visual area; AI, Agranular insular area; MOp5, Primary motor area, layer5; VISp6a, Primary visual area, layer 6a; SSp, Primary somatosensory area; SSs, Supplemental somatosensory area; ECT, Ectorhinal area; ORBm, Orbital area, medial part; RSPv, Retrosplenial area, ventral part; ACB, Nucleus accumbens; OT, Olfactory tubercle; CEAm, Central amygdalar nucleus, medial part; CEAl, Central amygdalar nucleus, lateral part; islm, Major island of Calleja; isl, Islands of Calleja; CP, Caudoputamen; CA3, Hippocampus field CA3; DG, Hippocampus dentate gyrus; CA1, Hippocampus field CA1; CA1sp, Hippocampus field CA1, pyramidal layer; SUBd-sp, Subiculum, dorsal part, pyramidal layer; IG, Induseum griseum; CA, Hippocampus Ammon's horn; PVT, Paraventricular nucleus of the thalamus; CL, Central lateral nucleus of the thalamus; AMd, Anteromedial nucleus, dorsal part; LGd, Dorsal part of the lateral geniculate complex; PCN, Paracentral nucleus; AV, Anteroventral nucleus of thalamus; VPM, Ventral posteromedial nucleus of the thalamus; AD, Anterodorsal nucleus; RT, Reticular nucleus of the thalamus; MM, Medial mammillary nucleus; PVH, Paraventricular hypothalamic nucleus; PVHp, Paraventricular hypothalamic nucleus, parvicellular division; SO, Supraoptic nucleus; DMHp, Dorsomedial nucleus of the hypothalamus, posterior part; ARH, Arcuate hypothalamic nucleus; PVHd, Paraventricular hypothalamic nucleus, descending division; SCH, Suprachiasmatic nucleus; LHA, Lateral hypothalamic area; SFO, Subfornical organ; VTA, Ventral tegmental area; SNc, Substantia nigra, compact part; SCm, Superior colliculus, motor related; IC, Ingerior colliculus; DR, Dorsal nucleus raphe; PAG, Periaqueductal gray; PBl, Parabrachial nucleus, lateral division; PG, Pontine gray; LC, Locus ceruleus; CSm, Superior central nucleus raphe, medial part; AP, Area postrema; NTS, Nucleus of the solitary tract; MV, Medial vestibular nucleus; NTSge, Nucleus of the solitary tract, gelatinous part; DCO, Dorsal cochlear nucleus; NTSm, Nucleus of the solitary tract, medial part; IO, Inferior olivary complex; VII, Facial motor nucleus; DMX, Dorsal motor nucleus of the vagus nerve; RPA, Nucleus raphe pallidus; PRP, Nucleus prepositus; CUL4,5mo, Cerebellum lobules IV-V, molecular layer; CUL4,5pu, Cerebellum lobules IV-V, Purkinje layer; PYRpu, Cerebellum Pyramus (VIII), Purkinje layer; CUL4,5gr, Cerebellum lobules IV-V, granular layer; MOE, main olfactory epithelium; VNO, vemoronasal organ.

Acknowledgments

We thank Jody Clements and Charlotte Weaver for help in preparing web site, Erina Hara, Asish Gulati, Xiaotang Jing and Zhe Meng for technical help, Keven McGowan for assistance in sequencing, Jim Cox, Amanda Zeladonis and Amanda Wardlaw for help in animal maintenance, Gabe Murphy for help in retinal sample collection, and Rosa Miyares for comments for the manuscript.

Competing Interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

References

Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput sequencing data. bioRxiv. 2014 feb; https://doi.org/10.1101%2F002824, doi: 10.1101/002824.

Arendt D. The evolution of cell types in animals: emerging principles from molecular studies. Nature Reviews Genetics. 2008 nov; 9(11):868–882. https://doi.org/10.1038%2Fnrg2416, doi: 10.1038/nrg2416.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000 may; 25(1):25–29. https://doi.org/10.1038%2F75556, doi: 10.1038/75556.

Bao W, Kojima KK, Kohany O. Repbase Update a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015 jun; 6(1). https://doi.org/10.1186%2Fs13100-015-0041-9, doi: 10.1186/s13100-015-0041-9.

Bedogni F, Hodge RD, Elsen GE, Nelson BR, Daza RAM, Beyer RP, Bammler TK, Rubenstein JLR, Hevner RF. Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. Proceedings of the National Academy of Sciences. 2010 jul; 107(29):13129–13134. https://doi.org/10.1073%2Fpnas.1002285107, doi: 10.1073/pnas.1002285107.

Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, Bell GW, Otte AP, Vidal M, Gifford DK, Young RA, Jaenisch R. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature. 2006 apr; 441(7091):349–353. https://doi.org/10.1038%2Fnature04733, doi: 10.1038/nature04733.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin DNA-binding proteins and nucleosome position. Nature Methods. 2013 oct; 10(12):1213–1218. https://doi.org/10.1038%2Fnmeth.2688, doi: 10.1038/nmeth.2688.

Cajal SR. Estructura de los centros nerviosos de las aves. Revista trimestral de histología normal y patológica. 1888 may; 1:1–36.

Cembrowski MS, Wang L, Sugino K, Shields BC, Spruston N. Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. eLife. 2016 apr; 5. https://doi.org/10.7554%2Felife.14997, doi: 10.7554/elife.14997.

Chandra R, Francis TC, Konkalmatt P, Amgalan A, Gancarz AM, Dietz DM, Lobo MK. Opposing Role for Egr3 in Nucleus Accumbens Cell Subtypes in Cocaine Action. Journal of Neuroscience. 2015 may; 35(20):7927–7937. https://doi.org/10.1523%2Fjneurosci.0548-15.2015, doi: 10.1523/jneurosci.0548-15.2015.

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, Miner TL, Nash WE, Nelson JO, Nhan MN, Pepin KH, Pohl CS, Ponce TC, Schultz B, Thompson J, Trevaskis E, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 dec; 420(6915):520–562. https://doi.org/10.1038%2Fnature01262, doi: 10.1038/nature01262.

Chong JA, Tapia-Ramirez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. Cell. 1995 mar; 80(6):949–57. https://www.ncbi.nlm.nih.gov/pubmed/7697725.

Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016 mar; 351(6277):1083–1087. https://doi.org/10.1126%2Fscience.aad5497, doi: 10.1126/science.aad5497.

Danesin C, Houart C. A Fox stops the Wnt: implications for forebrain development and diseases. Current Opinion in Genetics & Development. 2012 aug; 22(4):323–330. https://doi.org/10.1016%2Fj.gde.2012.05.001, doi: 10.1016/j.gde.2012.05.001.

Dasen JS, Jessell TM. Chapter Six Hox Networks and the Origins of Motor Neuron Diversity. In: *Current Topics in Developmental Biology* Elsevier; 2009.p. 169–200. https://doi.org/10.1016%2Fs0070-2153%2809%2988006-x, doi: 10.1016/s0070-2153(09)88006-x.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012 oct; 29(1):15–21. https://doi.org/10.1093%2Fbioinformatics%2Fbts635, doi: 10.1093/bioinformatics/bts635.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Duggan CD, DeMaria S, Baudhuin A, Stafford D, Ngai J. Foxg1 Is Required for Development of the Vertebrate Olfactory System. Journal of Neuroscience. 2008 may; 28(20):5229–5239. https://doi.org/10.1523%2Fjneurosci.1134-08.2008, doi: 10.1523/jneurosci.1134-08.2008.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 sep; 489(7414):57–74. https://doi.org/10.1038%2Fnature11247, doi: 10.1038/nature11247.

Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, Hemberg M, Ebert DH, Greenberg ME. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature. 2015 mar; 522(7554):89–93. https://doi.org/10.1038%2Fnature14319, doi: 10.1038/nature14319.

Gabitto MI, Pakman A, Bikoff JB, Abbott LF, Jessell TM, Paninski L. Bayesian Sparse Regression Analysis Documents the Diversity of Spinal Inhibitory Interneurons. Cell. 2016 mar; 165(1):220–233. https://doi.org/10.1016%2Fj.cell.2016.01.026, doi: 10.1016/j.cell.2016.01.026.

Gong S, Doughty M, Harbaugh CR, Cummins A, Hatten ME, Heintz N, Gerfen CR. Targeting Cre Recombinase to Specific Neuron Populations with Bacterial Artificial Chromosome Constructs. Journal of Neuroscience. 2007 sep; 27(37):9817–9823. https://doi.org/10.1523%2Fjneurosci.2707-07.2007, doi: 10.1523/jneurosci.2707-07.2007.

Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Research. 2014 oct; 43(D1):D1079–D1085. https://doi.org/10.1093%2Fnar%2Fgku1071, doi: 10.1093/nar/gku1071.

Grishkevich V, Yanai I. Gene length and expression level shape genomic novelties. Genome Research. 2014 jul; 24(9):1497–1503. https://doi.org/10.1101%2Fgr.169722.113, doi: 10.1101/gr.169722.113.

Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature. 2004 may; 429(6989):268–274. https://doi.org/10.1038%2Fnature02536, doi: 10.1038/nature02536.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, et al SS. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research. 2012 sep; 22(9):1760–1774. http://dx.doi.org/10.1101/gr.135350.111, doi: 10.1101/gr.135350.111.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell. 2010 may; 38(4):576–589. https://doi.org/10.1016%2Fj.molcel.2010.05.004, doi: 10.1016/j.molcel.2010.05.004.

Hempel CM, Sugino K, Nelson SB. A manual method for the purification of fluorescently labeled neurons from the mammalian brain. Nat Protoc. 2007 nov; 2(11):2924–2929. http://dx.doi.org/10.1038/nprot.2007.416, doi: 10.1038/nprot.2007.416.

Henry FE, Sugino K, Tozer A, Branco T, Sternson SM. Cell type-specific transcriptomics of hypothalamic energy-sensing neuron responses to weight-loss. eLife. 2015 sep; 4. https://doi.org/10.7554%2Felife.09800, doi: 10.7554/elife.09800.

Kanamori M, Konno H, Osato N, Kawai J, Hayashizaki Y, Suzuki H. A genome-wide and nonredundant mouse transcription factor database. Biochemical and Biophysical Research Communications. 2004 sep; 322(3):787–793. https://doi.org/10.1016%2Fj.bbrc.2004.07.179, doi: 10.1016/j.bbrc.2004.07.179.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, a D Haussler. The Human Genome Browser at UCSC. Genome Research. 2002 may; 12(6):996–1006. https://doi.org/10.1101%2Fgr.229102, doi: 10.1101/gr.229102.

Kratsios P, Kerk SY, Catela C, Liang J, Vidal B, Bayer EA, Feng W, Cruz EDDL, Croci L, Consalez GG, Mizumoto K, Hobert O. An intersectional gene regulatory strategy defines subclass diversity of C. elegans motor neurons. eLife. 2017 jul; 6. https://doi.org/10.7554%2Felife.25751, doi: 10.7554/elife.25751.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012 mar; 9(4):357–359. http://dx.doi.org/10.1038/nmeth.1923, doi: 10.1038/nmeth.1923.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014; 15(2):R29. https://doi.org/10.1186%2Fgb-2014-15-2-r29, doi: 10.1186/gb-2014-15-2-r29.

Leone DP, Heavner WE, Ferenczi EA, Dobreva G, Huguenard JR, Grosschedl R, McConnell SK. Satb2 Regulates the Differentiation of Both Callosal and Subcerebral Projection Neurons in the Developing Cerebral Cortex. Cerebral Cortex. 2014 jul; 25(10):3406–3419. https://doi.org/10.1093%2Fcercor%2Fbhu156, doi: 10.1093/cercor/bhu156.

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013 nov; 30(7):923–930. https://doi.org/10.1093%2Fbioinformatics%2Fbtt656, doi: 10.1093/bioinformatics/btt656.

Lipscombe D, Andrade A, Allen SE. Alternative splicing: Functional diversity among voltage-gated calcium channels and behavioral consequences. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2013 jul; 1828(7):1522–1529. https://doi.org/10.1016%2Fj.bbamem.2012.09.018, doi: 10.1016/j.bbamem.2012.09.018.

Lu KM, Evans SM, Hirano S, Liu FC. Dual role for Islet-1 in promoting striatonigral and repressing striatopallidal genetic programs to specify striatonigral cell identity. Proceedings of the National Academy of Sciences. 2013 dec; 111(1):E168–E177. https://doi.org/10.1073%2Fpnas.1319138111, doi: 10.1073/pnas.1319138111.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 may; 161(5):1202–1214. https://doi.org/10.1016%2Fj.cell.2015.05.002, doi: 10.1016/j.cell.2015.05.002.

Madisen L, Zwingman TA, Sunkin SM, Oh SW, Zariwala HA, Gu H, Ng LL, Palmiter RD, Hawrylycz MJ, Jones AR, Lein ES, Zeng H. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. Nature Neuroscience. 2009 dec; 13(1):133–140. https://doi.org/10.1038%2Fnn.2467, doi: 10.1038/nn.2467.

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Research. 2013 dec; 24(3):496–510. https://doi.org/10.1101%2Fgr.161034.113, doi: 10.1101/gr.161034.113.

Masland RH. Neuronal cell types. Current Biology. 2004 jul; 14(13):R497–R500. https://doi.org/10.1016%2Fj.cub.2004.06.035, doi: 10.1016/j.cub.2004.06.035.

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang SP, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK, Archidiacono N, Bork P, Butler J, Chang JL, Cheng Z, Chinwalla AT, deJong P, Delehaunty KD, Fronick CC, Fulton LL, et al. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005 sep; 437(7055):69–87. https://doi.org/10.1038%2Fnature04072, doi: 10.1038/nature04072.

Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, Urich MA, Nery JR, Sejnowski TJ, Lister R, Eddy SR, Ecker JR, Nathans J. Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron. 2015 jun; 86(6):1369–1384. https://doi.org/10.1016%2Fj.neuron.2015.05.018, doi: 10.1016/j.neuron.2015.05.018.

Moffitt JR, Hao J, Bambah-Mukku D, Lu T, Dulac C, Zhuang X. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. Proceedings of the National Academy of Sciences. 2016 nov; 113(50):14456–14461. https://doi.org/10.1073%2Fpnas.1617699113, doi: 10.1073/pnas.1617699113.

Montavon T, Soshnikova N. Hox gene regulation and timing in embryogenesis. Seminars in Cell & Developmental Biology. 2014 oct; 34:76–84. https://doi.org/10.1016%2Fj.semcdb.2014.06.005, doi: 10.1016/j.semcdb.2014.06.005.

Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005 jun; 435(7044):903–910. https://doi.org/10.1038%2Fnature03663, doi: 10.1038/nature03663.

Muotri AR, Gage FH. Generation of neuronal variability and complexity. Nature. 2006 jun; 441(7097):1087–1093. https://doi.org/10.1038%2Fnature04959, doi: 10.1038/nature04959.

Nelson SB, Sugino K, Hempel CM. The problem of neuronal cell types: a physiological genomics approach. Trends in Neurosciences. 2006 jun; 29(6):339–345. https://doi.org/10.1016%2Fj.tins.2006.05.004, doi: 10.1016/j.tins.2006.05.004.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ntranos V, Kamath GM, Zhang JM, Pachter L, Tse DN. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. Genome Biology. 2016 may; 17(1). https://doi.org/10.1186%2Fs13059-016-0970-8, doi: 10.1186/s13059-016-0970-8.

Okaty BW, Miller MN, Sugino K, Hempel CM, Nelson SB. Transcriptional and Electrophysiological Maturation of Neocortical Fast-Spiking GABAergic Interneurons. Journal of Neuroscience. 2009 may; 29(21):7040–7052. https://doi.org/10.1523%2Fjneurosci.0105-09.2009, doi: 10.1523/jneurosci.0105-09.2009.

Okaty BW, Freret ME, Rood BD, Brust RD, Hennessy ML, deBairos D, Kim JC, Cook MN, Dymecki SM. Multi-Scale Molecular Deconstruction of the Serotonin Neuron System. Neuron. 2015 nov; 88(4):774–791. https://doi.org/10.1016%2Fj.neuron.2015.10.007, doi: 10.1016/j.neuron.2015.10.007.

Okaty BW, Sugino K, Nelson SB. A Quantitative Comparison of Cell-Type-Specific Microarray Gene Expression Profiling Methods in the Mouse Brain. PLoS ONE. 2011 jan; 6(1):e16493. https://doi.org/10.1371%2Fjournal.pone.0016493, doi: 10.1371/journal.pone.0016493.

Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Scientific Reports. 2016 may; 6(1). https://doi.org/10.1038%2Fsrep25533, doi: 10.1038/srep25533.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M, Waddell S. Transposition-Driven Genomic Heterogeneity in the Drosophila Brain. Science. 2013 apr; 340(6128):91–95. https://doi.org/10.1126%2Fscience.1231965, doi: 10.1126/science.1231965.

Philippidou P, Dasen JS. Hox Genes: Choreographers in Neural Development Architects of Circuit Organization. Neuron. 2013 oct; 80(1):12–34. https://doi.org/10.1016%2Fj.neuron.2013.09.020, doi: 10.1016/j.neuron.2013.09.020.

Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, Sironi M. Intron size in mammals: complexity comes to terms with economy. Trends in Genetics. 2007 jan; 23(1):20–24. https://doi.org/10.1016%2Fj.tig.2006.10.003, doi: 10.1016/j.tig.2006.10.003.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Research. 2013 nov; 42(D1):D756–D763. http://dx.doi.org/10.1093/nar/gkt1114, doi: 10.1093/nar/gkt1114.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 jan; 26(6):841–842. http://dx.doi.org/10.1093/bioinformatics/btq033, doi: 10.1093/bioinformatics/btq033.

Richardson SR, Morell S, Faulkner GJ. L1 Retrotransposons and Somatic Mosaicism in the Brain. Annual Review of Genetics. 2014 nov; 48(1):1–27. https://doi.org/10.1146%2Fannurev-genet-120213-092412, doi: 10.1146/annurev-genet-120213-092412.

Schoenherr CJ, Anderson DJ. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. Science. 1995 mar; 267(5202):1360–3. https://www.ncbi.nlm.nih.gov/pubmed/7871435.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biology. 2007; 8(6):R127. https://doi.org/10.1186%2Fgb-2007-8-6-r127, doi: 10.1186/gb-2007-8-6-r127.

Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ, Gennert D, Gnirke A, Goren A, Hacohen N, Levin JZ, Park H, Regev A. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013 may; 498(7453):236–240. https://doi.org/10.1038%2Fnature12172, doi: 10.1038/nature12172.

Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics. 2013 jul; 14(9):618–630. https://doi.org/10.1038%2Fnrg3542, doi: 10.1038/nrg3542.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Shima Y, Sugino K, Hempel CM, Shima M, Taneja P, Bullis JB, Mehta S, Lois C, Nelson SB. A Mammalian enhancer trap resource for discovering and manipulating neuronal cell types. eLife. 2016 mar; 5. https://doi.org/10.7554%2Felife.13503, doi: 10.7554/elife.13503.

Simons C. Transposon-free regions in mammalian genomes. Genome Research. 2005 dec; 16(2):164–172. https://doi.org/10.1101%2Fgr.4624306, doi: 10.1101/gr.4624306.

Smit A, Hubley R, Green P, RepeatMasker Open-4.0; 2013-2015. http://www.repeatmasker.org.

Stefanakis N, Carrera I, Hobert O. Regulatory Logic of Pan-Neuronal Gene Expression in C. elegans. Neuron. 2015 aug; 87(4):733–750. https://doi.org/10.1016%2Fj.neuron.2015.07.031, doi: 10.1016/j.neuron.2015.07.031.

Sugino K, Hempel CM, Okaty BW, Arnson HA, Kato S, Dani VS, Nelson SB. Cell-Type-Specific Repression by Methyl-CpG-Binding Protein 2 Is Biased toward Long Genes. Journal of Neuroscience. 2014 sep; 34(38):12877–12883. https://doi.org/10.1523%2Fjneurosci.2674-14.2014, doi: 10.1523/jneurosci.2674-14.2014.

Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA. Power analysis of single-cell RNA-sequencing experiments. Nature Methods. 2017 mar; 14(4):381–387. https://doi.org/10.1038%2Fnmeth.4220, doi: 10.1038/nmeth.4220.

Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, Bertagnolli D, Goldy J, Shapovalova N, Parry S, Lee C, Smith K, Bernard A, Madisen L, Sunkin SM, Hawrylycz M, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature Neuroscience. 2016 jan; 19(2):335–346. https://doi.org/10.1038%2Fnn.4216, doi: 10.1038/nn.4216.

Thomas PD. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Genome Research. 2003 sep; 13(9):2129–2141. https://doi.org/10.1101%2Fgr.772403, doi: 10.1101/gr.772403.

Tsuchida T, Ensini M, Morton SB, Baldassare M, Edlund T, Jessell TM, Pfaff SL. Topographic organization of embryonic motor neurons defined by expression of LIM homeobox genes. Cell. 1994 dec; 79(6):957–970. https://doi.org/10.1016%2F0092-8674%2894%2990027-2, doi: 10.1016/0092-8674(94)90027-2.

Wei PC, Chang AN, Kao J, Du Z, Meyers RM, Alt FW, Schwer B. Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. Cell. 2016 feb; 164(4):644–655. https://doi.org/10.1016%2Fj.cell.2015.12.039, doi: 10.1016/j.cell.2015.12.039.

Xuan S, Baptista CA, Balas G, Tao W, Soares VC, Lai E. Winged helix transcription factor BF-1 is essential for the development of the cerebral hemispheres. Neuron. 1995 jun; 14(6):1141–1152. https://doi.org/10.1016%2F0896-6273%2895%2990262-7, doi: 10.1016/0896-6273(95)90262-7.

Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, Manno GL, Jureus A, Marques S, Munguba H, He L, Betsholtz C, Rolny C, Castelo-Branco G, Hjerling-Leffler J, Linnarsson S. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015 feb; 347(6226):1138–1142. https://doi.org/10.1126%2Fscience.aaa1934, doi: 10.1126/science.aaa1934.

Zeng Y, Navarro P, Shirali M, Howard DM, Adams MJ, Hall LS, Clarke TK, Thomson PA, Smith BH, Murray A, Padmanabhan S, Hayward C, Boutin T, MacIntyre DJ, Lewis CM, Wray NR, Mehta D, Penninx BWJH, Milaneschi Y, Baune BT, et al. Genome-wide Regional Heritability Mapping Identifies a Locus Within the TOX2 Gene Associated With Major Depressive Disorder. Biological Psychiatry. 2016 dec; https://doi.org/10.1016%2Fj.biopsych.2016.12.012, doi: 10.1016/j.biopsych.2016.12.012.

Zhang G, Titlow WB, Biecker SM, Stromberg AJ, McClintock TS. Lhx2 Determines Odorant Receptor Expression Frequency in Mature Olfactory Sensory Neurons. eNeuro. 2016 oct; 3(5). https://doi.org/10.1523%2Feneuro.0230-16.2016, doi: 10.1523/eneuro.0230-16.2016.

Zhang S, Kanemitsu Y, Fujitani M, Yamashita T. The newly identified migration inhibitory protein regulates the radial migration in the developing neocortex. Scientific Reports. 2014 aug; 4(1). https://doi.org/10.1038%2Fsrep05984, doi: 10.1038/srep05984.

Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, Phatnani HP, Guarnieri P, Caneda C, Ruderisch N, Deng S, Liddelow SA, Zhang C, Daneman R, Maniatis T, Barres BA, Wu JQ. An RNA-Sequencing Transcriptome and Splicing Database of Glia Neurons, and Vascular Cells of the Cerebral Cortex. Journal of Neuroscience. 2014 sep; 34(36):11929–11947. https://doi.org/10.1523%2Fjneurosci.1860-14.2014, doi: 10.1523/jneurosci.1860-14.2014.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Zheng C, Diaz-Cuadros M, Chalfie M. Hox Genes Promote Neuronal Subtype Diversification through Posterior Induction in Caenorhabditis elegans. Neuron. 2015 nov; 88(3):514–527. https://doi.org/10.1016%2Fj.neuron.2015.09.049, doi: 10.1016/j.neuron.2015.09.049.

Zylka MJ, Simon JM, Philpot BD. Gene Length Matters in Neurons. Neuron. 2015 apr; 86(2):353–355. https://doi.org/10.1016%2Fj.neuron.2015.03.059, doi: 10.1016/j.neuron.2015.03.059.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

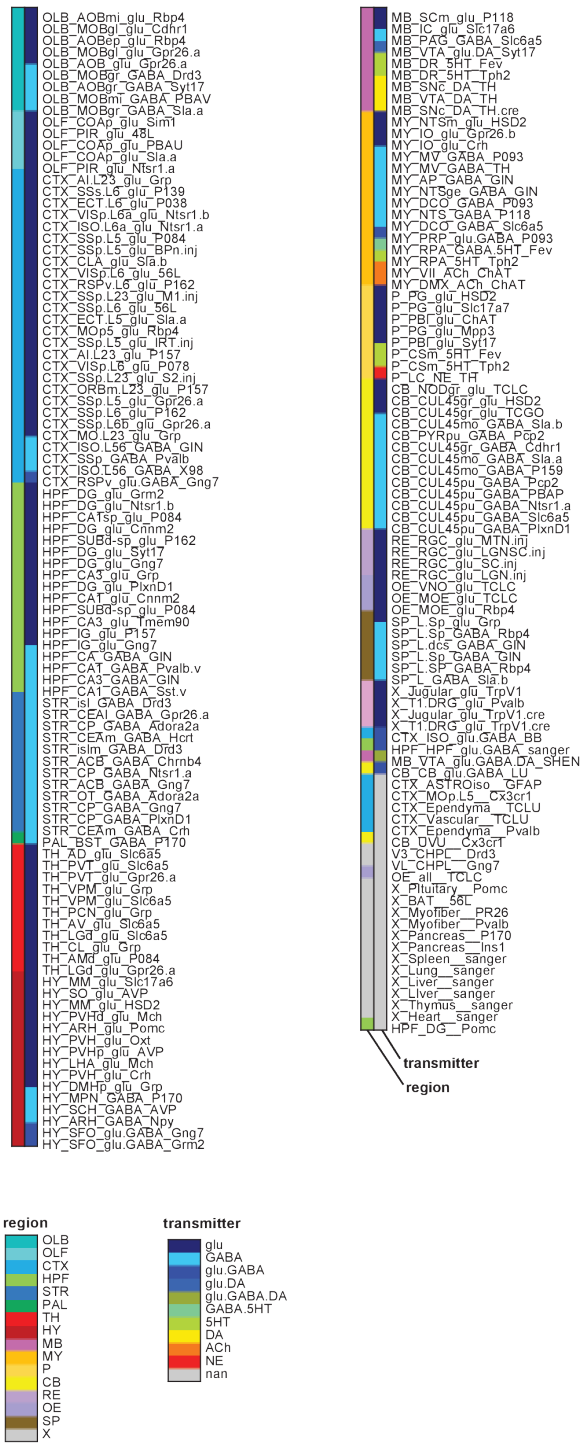


Figure 1–Supplement 1.

Cell type-specific samples. Sample groups color coded by region (left color bar) and transmitter phenotype (right color bar). Transmitter phenotype was determined from transmitter synthesis and storage enzyme expression. Abbreviations: OLB: olfactory bulb; OLF: olfactory regions (excluding bulb); CTX: Isocortex and Claustrum; HPF: hippocampal formation; STR: Striatum and related ventral forebrain structures; PAL: pallidum; TH: thalamus; HY: hypothalamus; MB: midbrain; MY: medulla; P: pons; CB: cerebellum; RE: retina; OE: olfactory epithelium; SP: spinal cord; X: peripheral nervous system or non-neural tissue. For additional abbreviations see Methods.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

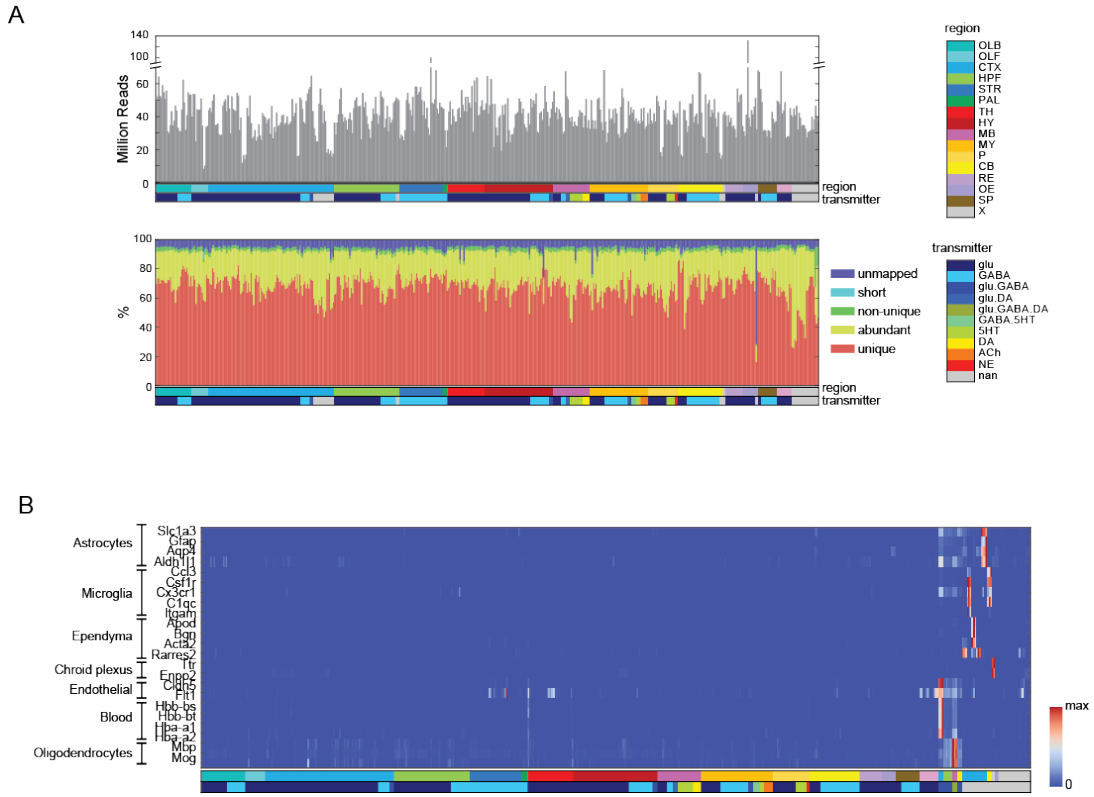


Figure 1–Supplement 2.

Quality Control measures. (A) (Top) Total reads for each of the libraries. Samples are color coded by region and transmitter, as shown in Figure 1 Supplement 1. (Bottom) Categories of reads in each library: unmapped: reads that did not map to the mm10 genome including chimeric and back-spliced reads; short: reads less than 30bp in length after removing adaptor sequences; non-unique: reads mapping to multiple locations; abundant: reads containing ribosomal RNA polyA, polyC and phiX sequences, and unique: uniquely mapped reads. For further analyses, abundant, short and unmapped reads were not used. **(B)** Contaminating transcripts from non-neuronal cell types. Samples with significant expression of these transcripts (at right) include tissue samples and non-neuronal samples. Each row is normalized by the maximum value.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

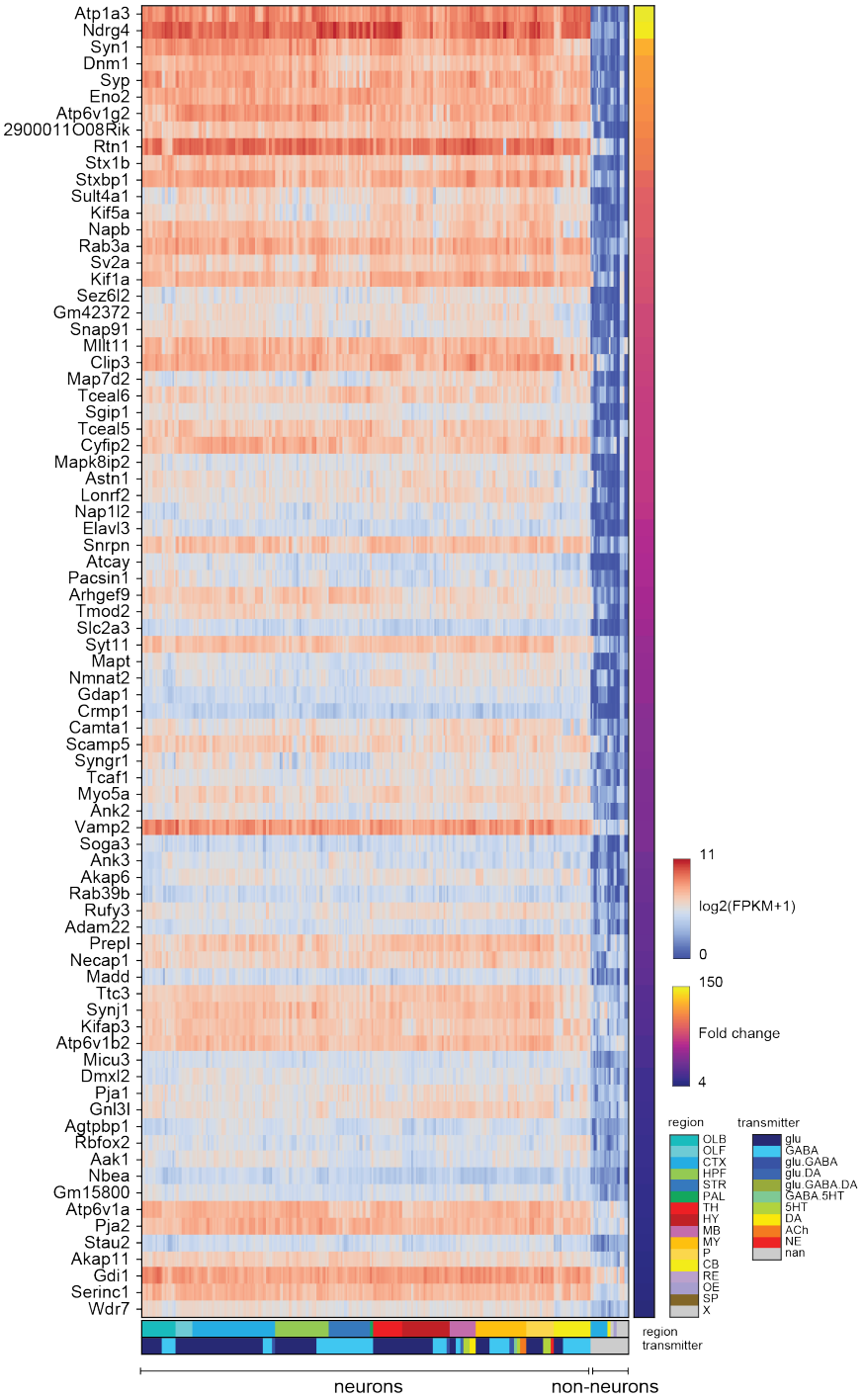


Figure 1–Supplement 3.

Pan-neuronal genes. Genes expressed in all neuronal cell types, but not (or at much lower levels) in non-neurons within the dataset. Heat-map shows log expression levels and the color at the right side indicates fold-change of the expression level between neurons and non-neurons. Criteria for extracting these genes are listed in the Methods.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

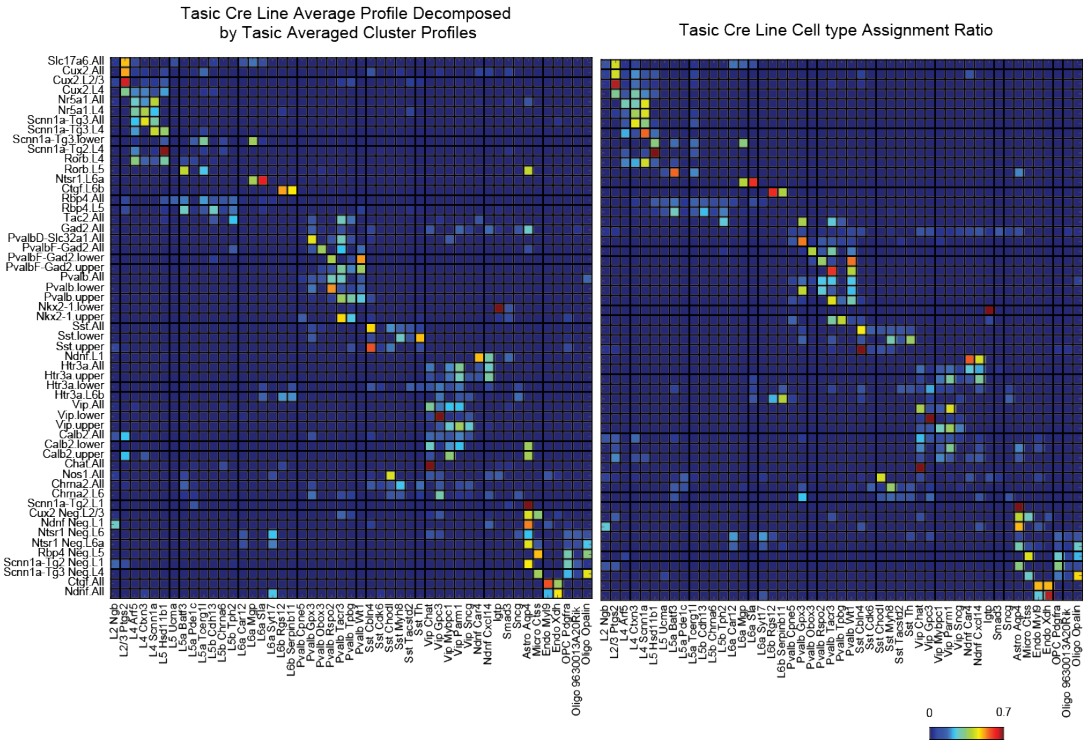


Figure 2–Supplement 1.

A test of NNLS decomposition. (Left) Single cell profiles from ***Tasic et al.*** (***2016***) were merged according to which of the 17 transgenic strains and sub-dissected layers they originated from (row labels). Merged profiles were then decomposed using NNLS by the same individual cluster profiles used in Figure 2 (column labels). **(Right)** The reported proportion of single cell profiles according to the author's classification. The close similarity between left and right matrices indicates an accurate NNLS decomposition of the merged clusters. Note that information about which and how many individual cell types were sorted from each line and set of layers was not explicitly provided to the decomposition algorithm, but were accurately deduced from the merged expression profiles.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

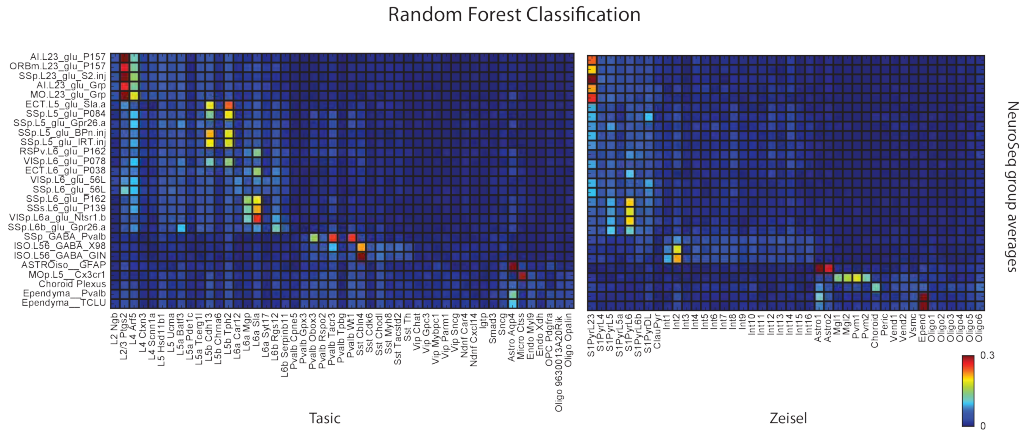


Figure 2–Supplement 2.

Random forest decomposition. A random forest classifier (500 decision trees) was trained from single cell profiles and their cluster assignment (column labels) and then used to decompose NeuroSeq cell types (row labels). Coefficients are the ratio of the votes from the 500 trees (coefficient ranges from 0 to 1 and 1 indicates all trees vote for a single class). The pattern of coefficients is similar to that obtained by NNLS (Figure 2A) suggesting the decomposition is relatively robust and does not reflect a peculiarity of the NNLS algorithm.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

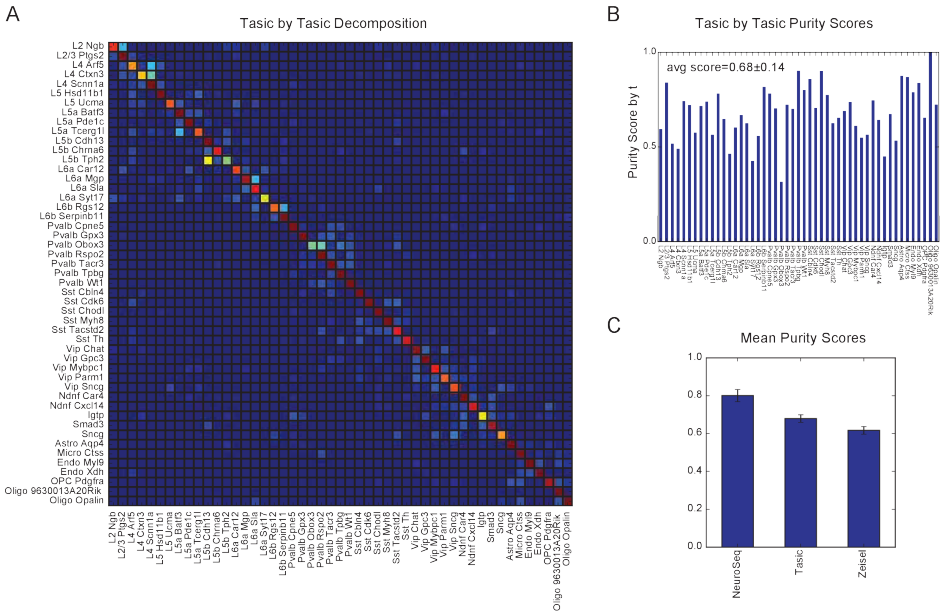


Figure 2–Supplement 3.

Cross validation of NNLS decompositions (A) Each of Tasic et al. cluster is randomly divided into two groups and one is used to decompose the other. Some cluster pairs share significant coefficients, suggesting they are too similar to each other to separate well. For example, pairs of clusters L2 Ngb and L2/3 Ptgs2, L4 Arf5 and L4 Scnn1a, L4 Ctxn3 and L4 Scnn1a, and L5 Cdh13 and L5 Tph2 are hard to distinguish. This is consistent with the observation of intermediate cells between each of these clusters in the original study (their Figure 4).**(B)** Purity scores (similar to Figure 2C) for the cross-validated NNLS decomposition of each Tasic et al. cluster. **(C)** Mean purity scores obtained from the same cross-validation procedure applied to each of the three datasets.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

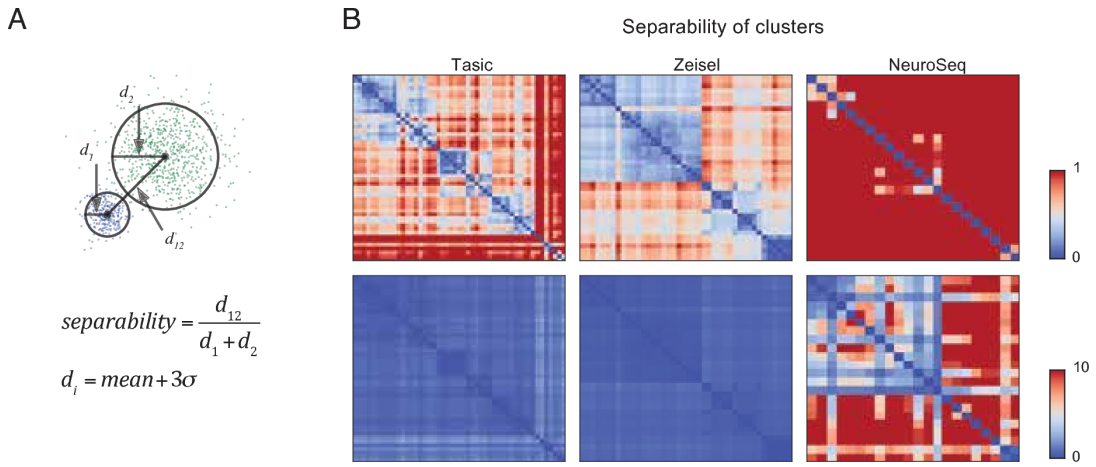
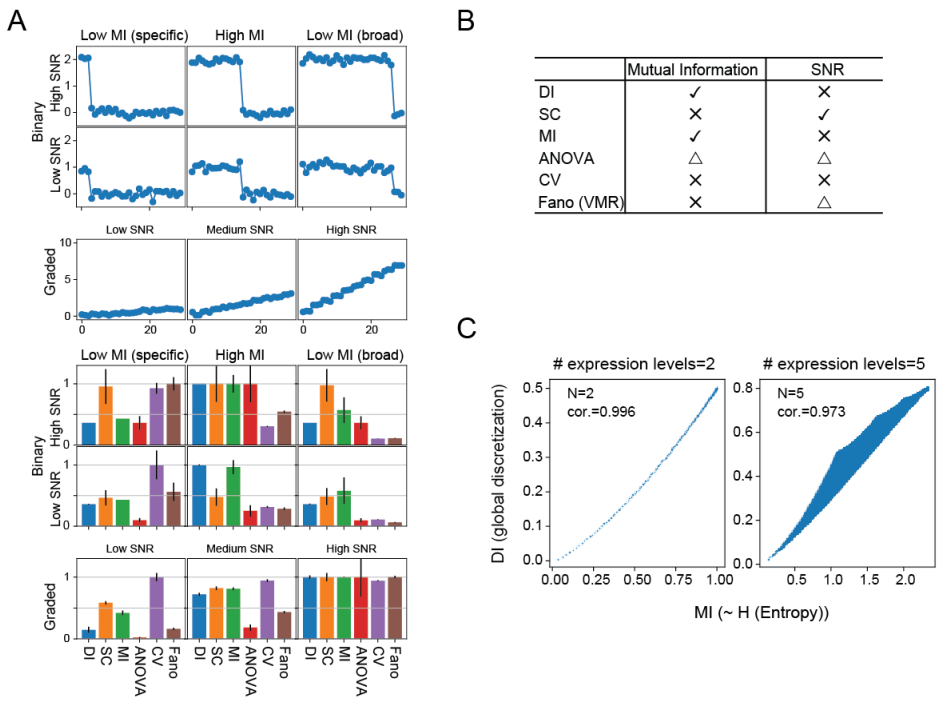


Figure 2–Supplement 4.

Separability of cell type clusters (A) Definition of separability. Cartoon represents two different single cell clusters as distributions of points. The separability is the ratio of the distance between the centroids to the sum of the "diameter" of each cluster. Here, we calculate the diameter of a cluster using the distances from the centroid of the cluster as the mean distance + 3 times the standard deviation of the distribution of the distances. With this definition, two clusters are "touching" when separability =1, overlapping when <1, and separate when >1. The multi-dimensional distance is computed as 1- Pearson's corr.coef. **(B)** Separabilities between cell type clusters for three datasets shown with two different dynamic ranges (color scale; 0-1 for upper row and 0-10 for lower row). The order of cell type clusters are the same as in Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



	Mutual Information	SNR
DI	✓	×
SC	×	✓
MI	✓	×
ANOVA	△	△
CV	×	×
Fano (VMR)	×	△

Figure 3–Supplement 1.
Simulated data reveal features of expression metrics
(A) (Upper) An example of simulated binary and graded expression patterns with added noise. X-axis indicates sample/groups. (Lower) Various average metrics calculated from the simulated expression patterns (100 individual simulations; error bars are standard deviations). Values are normalized within each metric across binary expression group or graded expression group. **(B)** Summary of each metric's correlation with Mutual Information and SNR: check mark–correlated, X–uncorrelated, triangle–partially correlated. **(C)** DI and MI are highly correlated. The relationship between DI, calculated without considering replicates, and MI with expression levels discretized into 2 levels (left) and 5 levels (right). Although increasing the number of discrete expression levels decreases the degree of correlation, they remain monotonically and closely related.

Figure 3–Supplement 2.
Relationship between DI and MI Here we explore more detailed relationship between mutual information and differentiation index. To calculate mutual information between expression levels and cell types, we discretize expression levels into N_e levels. Let N_s be number of samples. Let n_{ij} be counts in the contingency table where $i = 1, ..., N_e$ and $j = 1, ..., N_s$. Then the joint probability distribution and the marginal probability distribution can be written as:

$$p(i,j) = \frac{n_{ij}}{N_s} \tag{1}$$

$$p(i) = \frac{\sum_j n_{ij}}{N_s} = \frac{n_i}{N_s} \tag{2}$$

$$p(j) = \frac{\sum_i n_{ij}}{N_s} = \frac{n_j}{N_s} \tag{3}$$

$$\tag{4}$$

Where $n_i = \sum_j n_{ij}$ and $n_j = \sum_i n_{ij}$. n_j is number of replicates in cell type j. The mutual information

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

between expression level (E) and samples (S) is:

$$I(E;S) = \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p(i)p(j)} \tag{5}$$

$$= \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p(j)} - \sum_{i,j} p(i,j) \log p(i) \tag{6}$$

$$= \sum_{i,j} p(j)p(i|j) \log p(i|j) - \sum_{i,j} p(i,j) \log p(i) \tag{7}$$

$$= \sum_{j} p(j) \sum_{i} p(i|j) \log p(i|j) - \sum_{i} \log p(i) \sum_{j} p(i,j) \tag{8}$$

$$= -\sum_{j} p(j) H(E|S=j) - \sum_{i} p(i) \log p(i) \tag{9}$$

$$= -H(E|S) + H(E) \tag{10}$$

$H(E|S=j)$ is the entropy of expression levels in cell type j, which represents the expression noise in cell type j, and $H(E|S)$ is the average of these across all cell types. When there is no replicates $H(E|S)$ is zero. When there are replicates, $H(E|S=j)$ represents how noisy the expression is. This may depends on expression level, and $H(E|S)$, the average of $H(E|S=j)$ may depends on expression prevalence (i.e., how widely the gene is expressed), but in any case, the first term $-H(E|S)$ represents reduction of the mutual information by noise.

The second term $H(E)$ is the entropy of marginal distribution $p(i)$ and represents the main information content of cell types encoded in expression levels. This can be rewritten using counts in the contingency table as:

$$H(E) = -\sum_{i} p(i) \log p(i) \tag{11}$$

$$= -\sum_{i} \frac{n_i}{N_s} \log \frac{n_i}{N_s} \tag{12}$$

$$= -\sum_{i} \frac{n_i}{N_s} \log n_i + \sum_{i} \frac{n_i}{N_s} \log N_s \tag{13}$$

$$= -\frac{1}{N_s} \sum_{i} n_i \log n_i + \log N_s \tag{14}$$

Thus, it takes maximum when all n_i's are 0 or 1, which corresponds to the case where one expression level corresponds to one cell type, making all cell types distinguishable by the expression levels. This is when the discretization levels are larger than number of samples. When the number of discretization levels (N_e) is smaller than the number of samples (N_s), $H(E)$ takes the maximum value of $\log N_e$ when all the samples are distributed equally to each bin.

To explore the relationship between $H(E)$ and DI, the $\log n_i$ in the first term is replaced (approximated) by $(n_i - 1)$ (first two terms in the Taylor expansion of $\log n_i$ around $n_i = 1$.):

$$H(E) \sim -\frac{1}{N_s} \sum_{i} n_i(n_i - 1) + \log N_s \tag{15}$$

$$= -\frac{2}{N_s} \sum_{i} n_i(n_i - 1)/2 + \log N_s \tag{16}$$

$$= \frac{2}{N_s} \left\{ N_s(N_s - 1)/2 - \sum_{i} n_i(n_i - 1)/2 \right\} - (N_s - 1) + \log N_s \tag{17}$$

$$= (N_s - 1) s DI - (N_s - 1) + \log N_s \tag{18}$$

Since n_i is the number of samples in one expression level, $n_i(n_i - 1)/2$ is the number of indistinguishable pairs in that expression level when there is no replicate. The term within the curly bracket is then the number of distinguishable pairs, leading to eq.(18).

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

More formally, since both $h(p) = \sum n_i \log n_i$ and $d(p) = \sum n_i(n_i - 1) = \sum n_i^2 - N_s$ are Schur-convex functions[1] on partitions of N_s, $p = (n_1, n_2, ..., n_k)$, when partition p_1 majorizes p_2 then, $h(p_1) \geq h(p_2)$ and $d(p_1) \geq d(p_2)$. When partition length is 2, that is when expression levels are discretized into only 2 levels, corresponding to ON/OFF, then, all of the partitions can be ordered by majorization relationship, therefore, $h(p)$ and $d(p)$ are order-preserved transformation of each other (Figure 3 Supplement 1C left). When partition length is greater than 2, this relationship is not true. However, they are still highly correlated to each other (Figure 3 Supplement 1C right).

When DI is calculated from global discretization (as in the above case), the maximum number of pairs distinguishable happens when all the samples are equally distributed to each bin and the number of distinguishable pairs is $\left(\frac{N_s}{N_e}\right)^2 N_e(N_e - 1)/2$. Therefore,

$$max(DI) = \left(\frac{N_s}{N_e}\right)^2 \frac{N_e(N_e - 1)/2}{N_s(N_s - 1)/2} \tag{19}$$

$$= \left(1 - \frac{1}{N_e}\right) / \left(1 - \frac{1}{N_s}\right) \tag{20}$$

$$\sim 1 - \frac{1}{N_e} \quad (when \quad N_s \gg 1) \tag{21}$$

As stated above, this is also when the entropy $H(E)$ takes the maximum value of $\log_2 N_e$ in the unit of bits. (Figure 3 Supplement 1C)

[1] A Schur-convex function is a function $f : \mathbb{R}^k \to \mathbb{R}$ which satisfies $f(x) \geq f(y)$ for all x, y where x majorizes y. For $x = (x_1, x_2, ..., x_k) \in \mathbb{R}^k where (x_1 \geq x_2 \geq ... \geq x_k)$ and $y = (y_1, y_2, ..., y_k) \in \mathbb{R}^k where (y_1 \geq y_2 \geq ... \geq y_k)$. x majorizes y when $\sum_{i=1}^k x_i = \sum_{i=1}^k y_i and \sum_{i=1}^j x_i \geq \sum_{i=1}^j y_i for all j = 1, ..., k$. When x majorizes y, it follows $x_i \geq y_i$ for all i, so it is easy to see $h(x) \geq h(y)$ and $d(x) \geq d(y)$.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

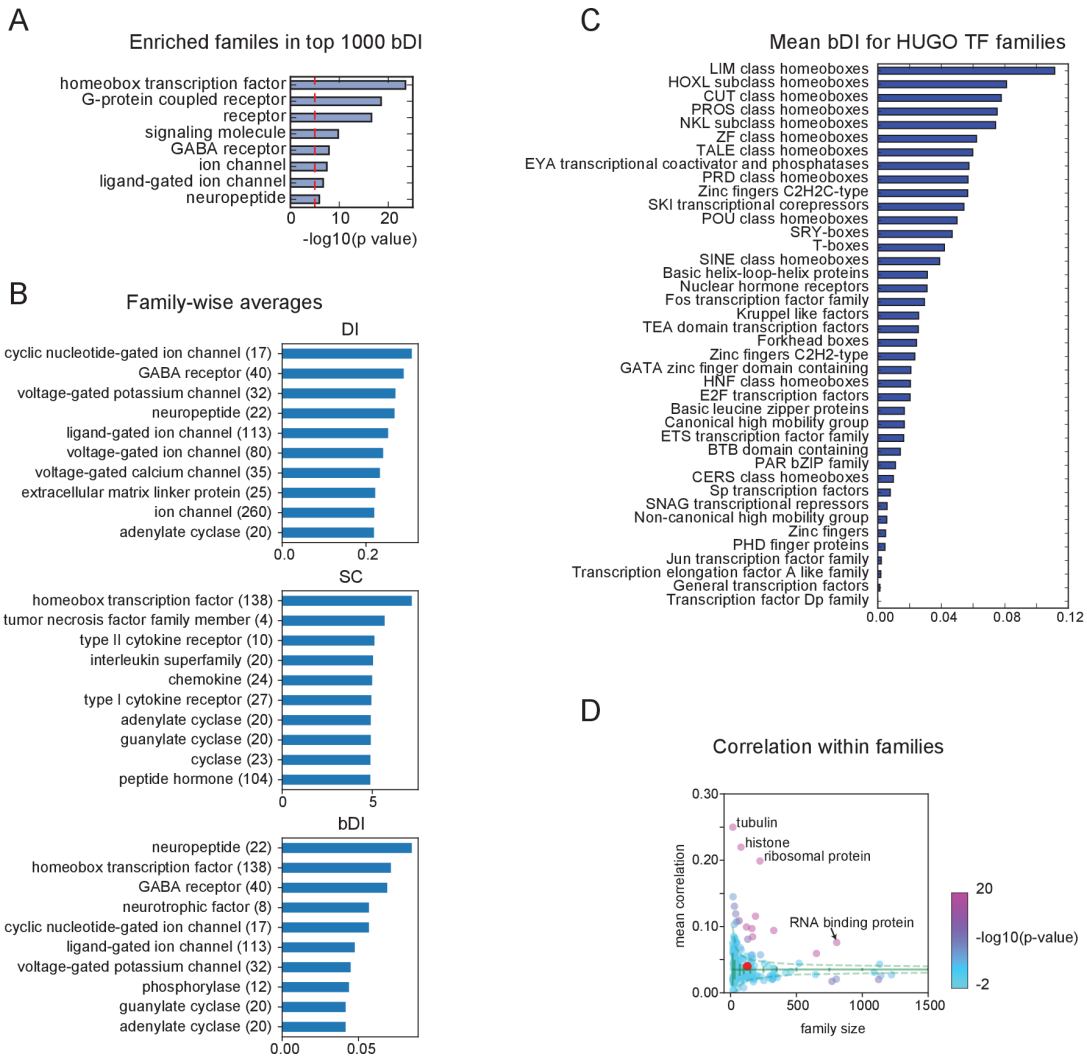


Figure 4–Supplement 1.

(A) PANTHER families enriched in the top 1000 bDI genes. **(B)** Averages of metrics (DI,SC,bDI) for PANTHER families. Only top 10 are shown. Numbers in parenthesis indicate family size. **(C)** Average bDI calculated for each TF family in HUGO protein families (***Gray et al., 2014***). **(D)** Mean Pearson's corr. coef. between genes within PANTHER families. Homeobox TF family is indicated by the red dot. Most of the PANTHER family genes are decorrelated within families but genes in some families, such as ribosomal protein, histone, tubulin, and RNA binding protein have highly significant correlation within families.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

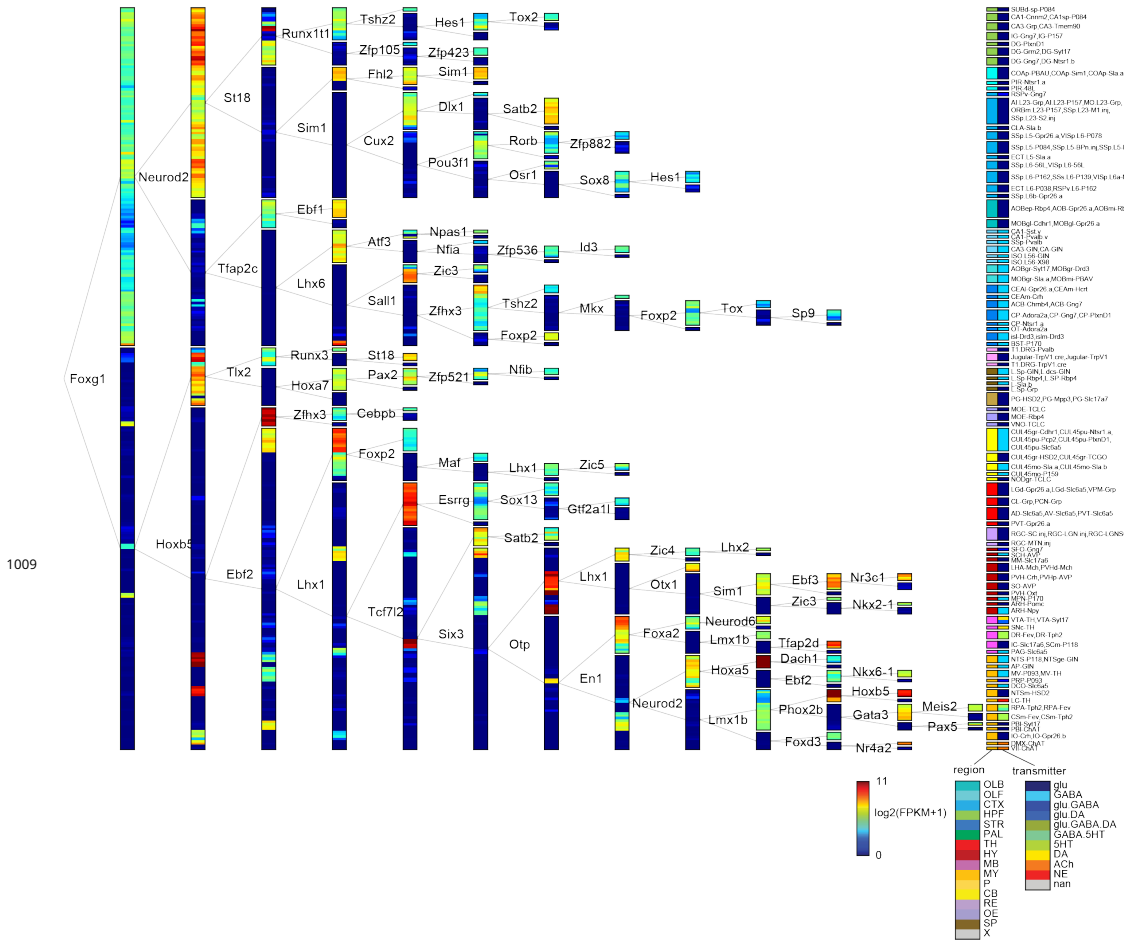


Figure 5–Supplement 1.

TF tree constructed using stronger anatomical constraints. Similar to Figure 5, but the constraints on anatomical boundaries are enforced during each bisection. However, TF expression was not constrained to be uniform within a group, leading to some subgroups that do not match the expression of the dividing gene.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

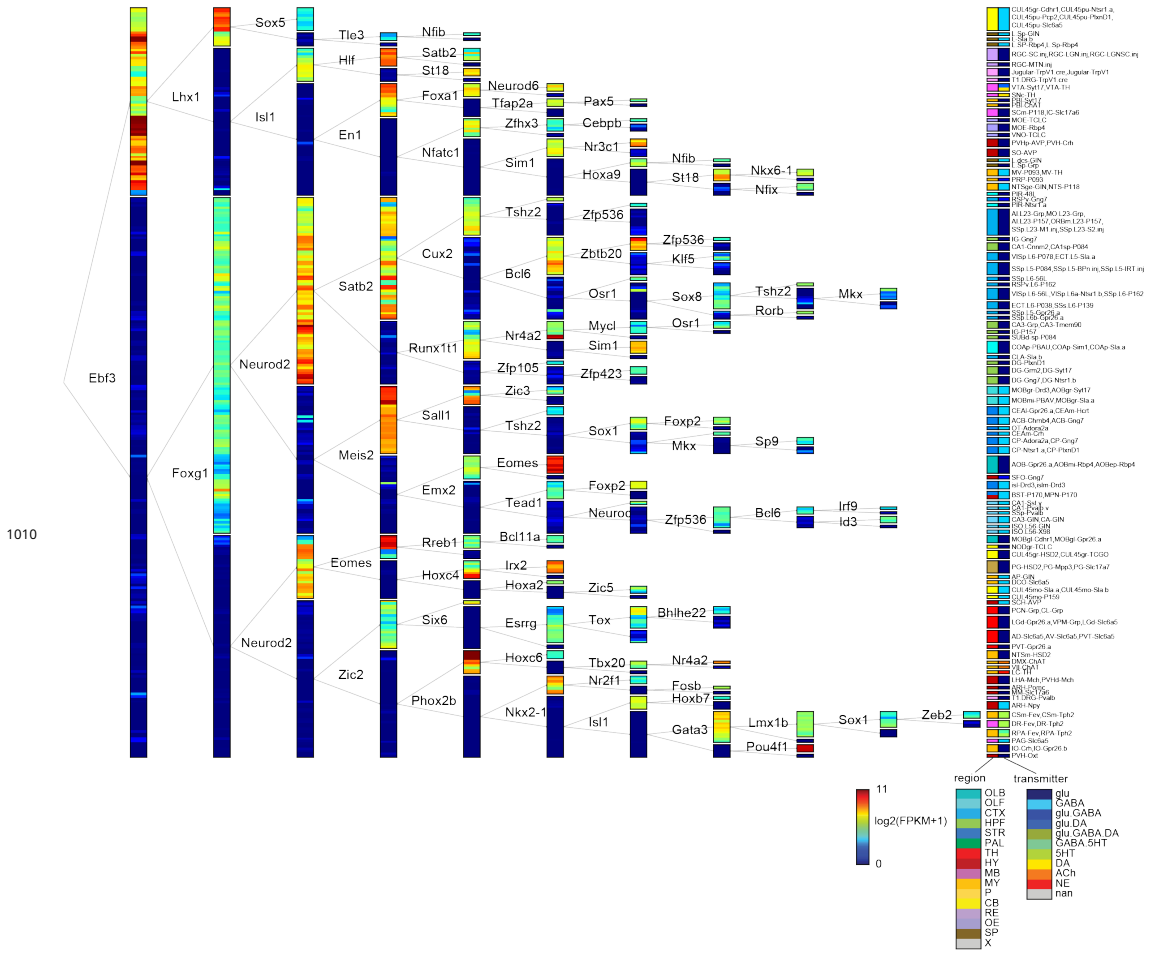


Figure 5–Supplement 2.

TF tree constructed without anatomical constraints. Similar to Figure 5 but anatomical sub-regions were not constrained to be grouped together.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

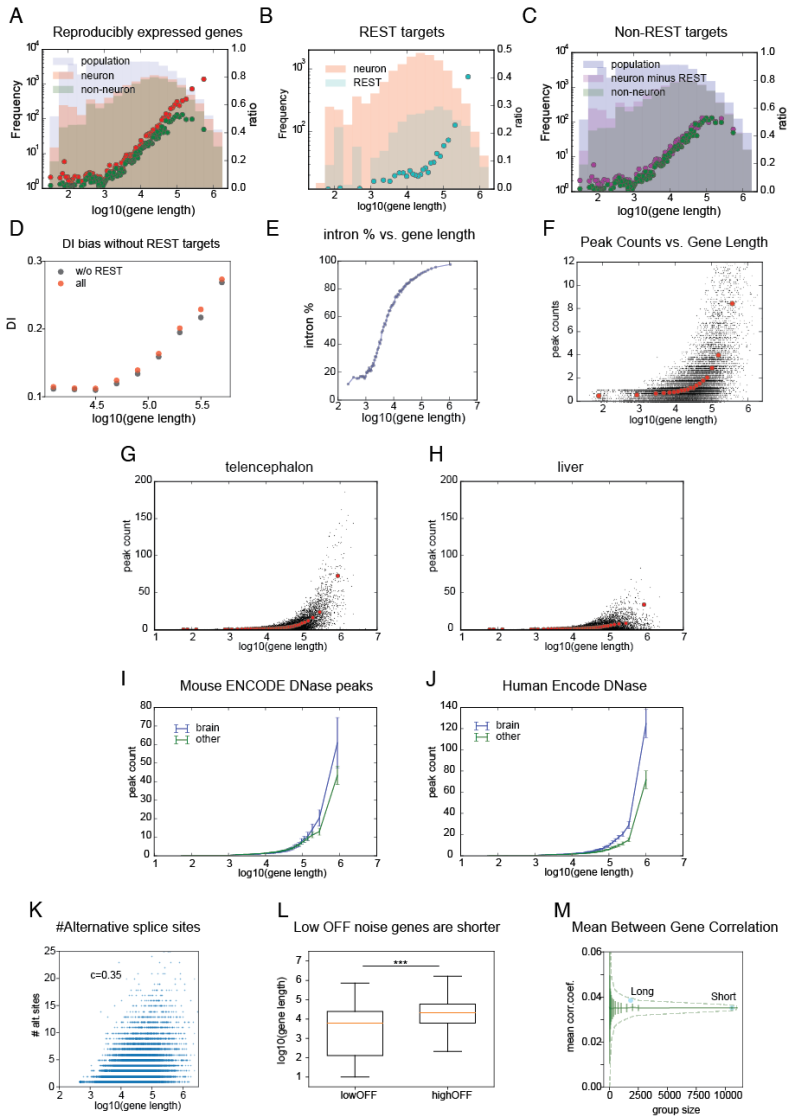


Figure 6–Supplement 1.

Properties of long genes in current and prior datasets.(A) Number (histogram) and ratios (dots) of genes expressed in neurons (pink histogram, red dots) and non-neurons (brown histogram, green dots) relative to the number of genes in the entire population (grey histogram) as a function of gene length (ratios computed per bin of 500 genes). **(B)** Number (cyan histogram; left axis) and ratios (cyan dots; right axis) of genes with nearby NRSE relative to the numbers of neuronally expressed genes (pink histogram). **(C)** (Magenta dots) ratio of neuronally expressed non-REST target genes to the population. Other components are same as in A. **(D)** DI dependence of length without REST target genes compared to all genes. DI is still strongly length dependent because REST targets are a small fraction of expressed long genes. **(E)** Fraction of gene length attributable to intron length. **(F)** Length dependence of peak counts in the ATAC-seq data from the current study. **(G)- (J)** Length dependence of peak counts in ENCODE DNase hypersensitivity data. Examples from mouse ENCODE data in forebrain (telencephalon) **(G)** and liver **(H)** samples showing individual peaks (black dots) and binned averages (red dots) as a function of gene length. Average mouse **(I)** and human **(J)** peak counts from brain(blue) and non-brain(green) samples. **(K)** Number of alternative splice sites for each gene (in Gencode mouse v14) plotted against gene length. **(L)** Distribution of gene lengths for low OFF noise genes (Figure 4B red dashed region) and high OFF noise genes (Figure 4B blue dashed region). Red lines are medians and whiskers indicate 1.5 IQR. (***:p<1e-100, Student's t-test.) **(M)** Similar to Figure 4 Supplement 1D, mean Pearson's correlation coefficients between genes within long and short gene groups relative to mean and S.D. (green solid lines) and 99% confidence interval (green dashed lines) calculated from randomly selected groups of genes.

bioRxiv preprint doi: https://doi.org/10.1101/208355; this version posted November 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

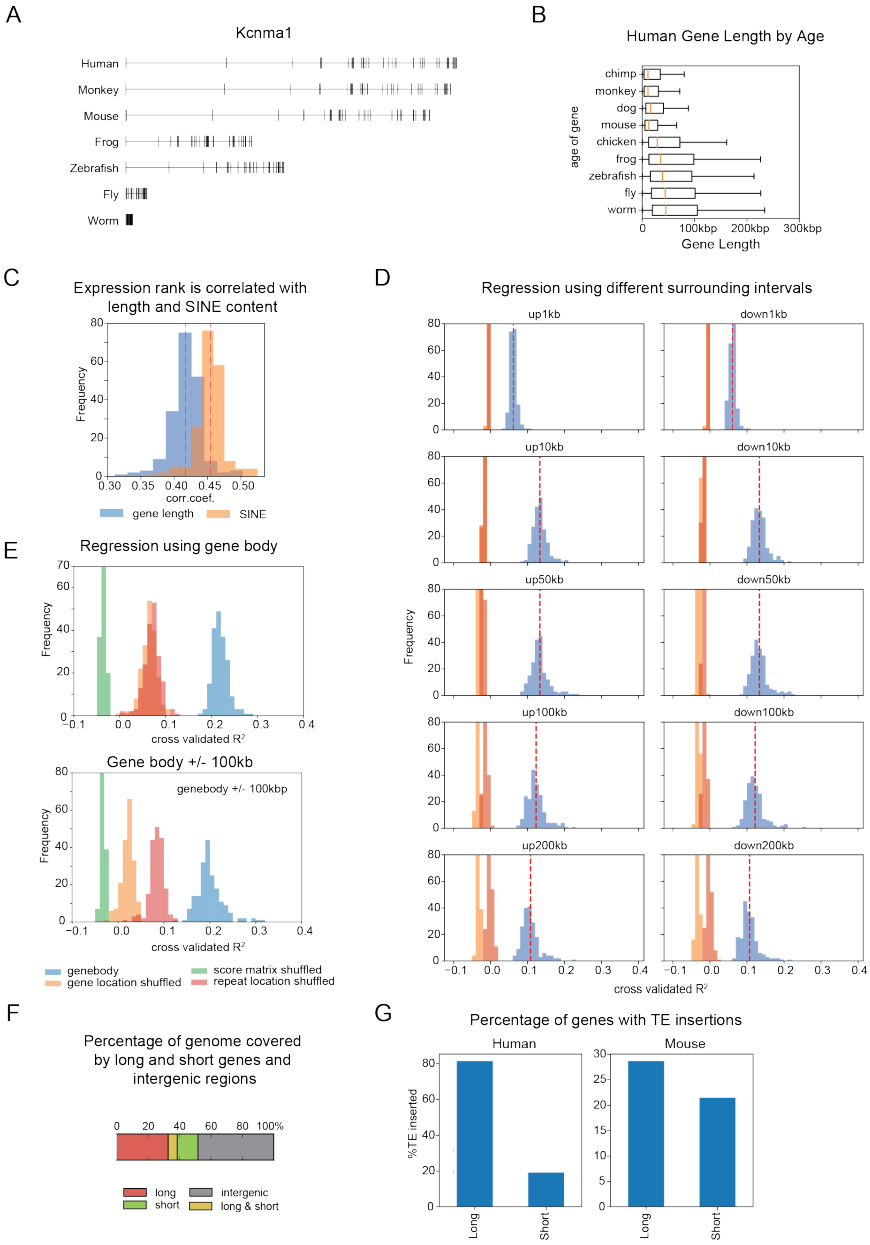


Figure 7–Supplement 1.

Supplementary to Figure 7. TE insertions elongate genes and contain information about gene expression (A) Example of gene length differences between species for *Kcnma1* (a calcium-activated potassium channel, also called *slopoke*, in *Drosophila*). **(B)** Estimated evolutionary age of human genes correlates with their length. The length distribution of human genes is plotted as a function of age, estimated from their most distant homologs. Genes common to all vertebrates (or to all listed genomes) are longer than genes common only to mammals (mouse) or common only to primates (monkey). **(C)** Correlation between gene expression rank and gene length (blue) and SINE repeat score (orange) calculated for all cell types. Because of their abundance, SINE repeat scores are correlated with gene length. **(D)** Similar to Figure 7E but using repeat scores calculated from different sized intervals surrounding each gene (not including the gene body). Average R^2 is maximal near 10kb for both upstream and downstream intervals. Shuffling conditions are colored as in Figure 7E. **(E)** Similar to Figure 7E but for repeat scores calculated from gene body only (upper panel) or gene body+/−100kb (lower panel). **(F)** Fraction of genome spanned by long genes (orange) is greater than that spanned by short genes (green), despite being fewer in number. Some genomic regions contain overlapping long and short genes (yellow). **(G)** Percentage of inserted sequences calculated in Figure 7A (Human vs. Chimp and Mouse vs. Rat), that overlap TEs within long (≥ 100kbp) or short (<100kbp) genes.