A topographic axis of transcriptional identity in thalamus

Authors: *James W. Phillips[1,3]*, Anton Schulmann[1,2]*, Erina Hara[1,†], Chenghao Liu[2], Brenda C. Shields[1,⊘], Wyatt Korff[1], Andrew Lemire[1], Joshua Dudman[1], Sacha B. Nelson[1,2,3], Adam Hantman[1,3]*
**co-first authors, Corresponding authors[3]*

[1]HHMI Janelia Research Campus, Ashburn, Virginia, USA
[2]Brandeis University, Waltham Massachusetts, USA
[3]Corresponding Authors
[*]Co-first authorship
[†]Present Address: City College Of New York, 160 Convent Ave, New York City, USA
[⊘]Present Address: Duke University - Pratt School of Engineering 311 Research Drive, Durham, NC 27710
Correspondence to: James Phillips (phillipsj10@janelia.hhmi.org), Sacha Nelson (nelson@brandeis.edu), Adam Hantman (hantmana@janelia.hhmi.org)

Abstract:

A fundamental goal in neuroscience is to uncover common principles by which different modalities of information are processed. In the mammalian brain, thalamus acts as the essential hub for forebrain circuits handling inputs from sensory, motor, limbic, and cognitive pathways[1]. Whether thalamus imposes common transformations on each of these modalities is unknown. Molecular characterization offers a principled approach to revealing the organization of thalamus. Using near-comprehensive and projection-specific transcriptomic sequencing, we found that almost all thalamic nuclei fit into one of three profiles. These profiles lie on a single axis of genetic variance which is aligned with the mediolateral spatial axis of thalamus. Genes defining this axis of variance include receptors and ion channels, providing a systematic diversification of input/output transformations across the topography of thalamus. Single cell transcriptional profiling revealed graded heterogeneity within individual thalamic nuclei, demonstrating that a spectrum of cell types and potentially diverse input/output transforms exist within a given thalamic nucleus. Together, our data argue for an archetypal organization of pathways serving diverse input modalities, and provides a comprehensive organizational scheme for thalamus.

Main Text

We produced a near-comprehensive transcriptomic atlas of murine thalamus via microdissection of retrogradely labeled cells and subsequent RNAseq analysis (22 nuclei, 120 samples). Anterograde tracing of inputs to thalamus was used when identification of nuclear boundaries was ambiguous (Figure 1A,B for an example dissection approach). Labeled neurons from a given thalamic nucleus were pooled and profiled, an approach chosen because nuclear

identity and projection target are major definers of cell type. See supplementary tables 1 and 2 for details on labelling strategies.

To identify the relationship between the transcriptomes of thalamic nuclei in an unsupervised manner, we selected the top 500 most differentially expressed genes between thalamic nuclei (via an ANOVA-like test, see methods). Using this differentially expressed gene set, we grouped the thalamic nuclei via hierarchical clustering (1 - Spearman's correlation as distance metric), identifying five major subdivisions of nuclei (Figure 1C). We termed these profiles, not clusters, as we do not mean to imply discreteness between the classifications (see below). These profiles are not explained by the modality of information processed, since subnuclei of motor, somatosensory and visual complexes all split across different profiles. For example, the blue profile (Figure 1C) included both the cerebellar-recipient motor thalamus and the retina-recipient visual nucleus (lateral geniculate, LGd), whilst the red profile contained the basal ganglia recipient motor thalamus and the superior colliculus-recipient visual nucleus (lateral posterior, LP). Interestingly, the anterior dorsal nucleus (AD) and nucleus reuniens (RE) had distinct profiles and each formed profiles of their own. When schematic sections of thalamus were colored by these profiles, a mediolateral separation was observed, with the 'blue' nuclei being most lateral and 'yellow' nuclei most medial (Figure 1D). Thus, the largest fraction of linear gene expression variance in thalamus is not explained by modality, but rather a topographic axis spanning the structure's mediolateral extent.

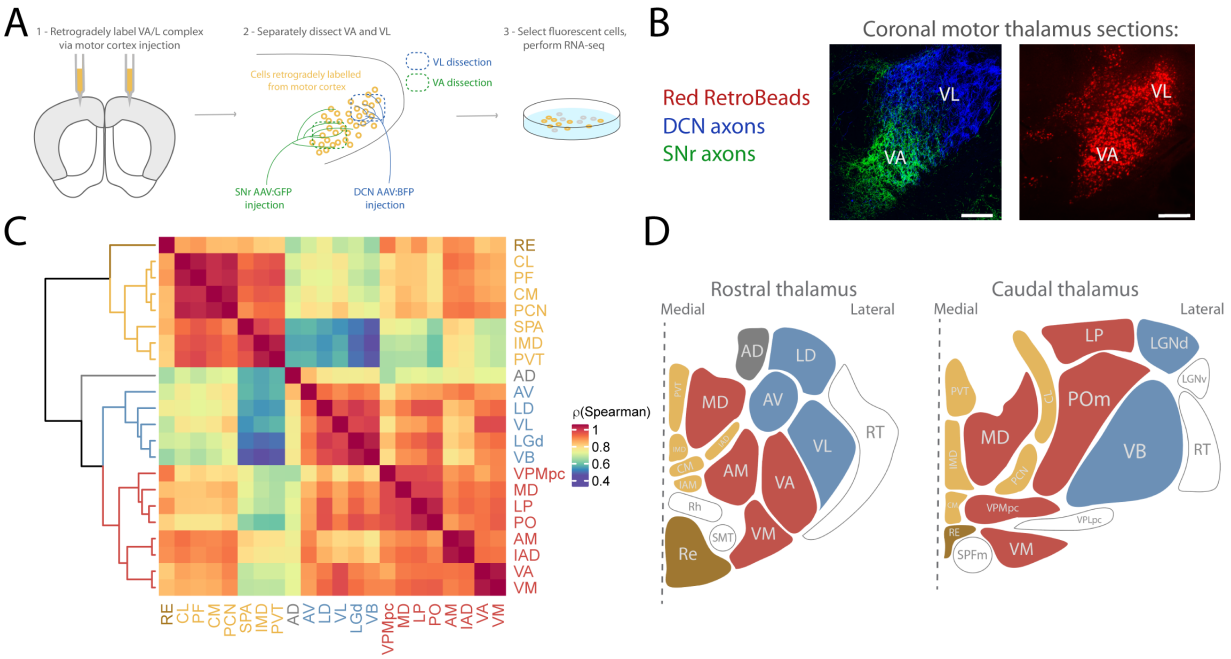


Figure 1 – *A comprehensive transcriptomic atlas allows unsupervised grouping of thalamic gene expression profiles*

A. Schematic of experimental pipeline to obtain the thalamic RNAseq atlas. In this example, motor thalamic neurons were retrogradely labelled from their primary projection target (motor cortex),

manually dissected and sorted. Viruses expressing green and blue fluorescent proteins (GFP, BFP) were injected to the deep cerebellar nuclei and substantia nigra reticulata to mark nuclear subdivisions (Ventral lateral (VL) and ventral anterior (VA), respectively).

B. Example labelling from scheme shown in A. Coronal sections, scale bars = 200um.

C. Hierarchical clustering of thalamic nuclei using assessed Spearman's correlation of top 500 most differentially expressed genes (supplementary figure 1B) across all 22 nuclei. Major clusters defined as the top 5 branches of cluster dendrogram.

D. Topographic localization of profiles in thalamus, colored as in figure 1C. Rostral and caudal coronal thalamic section schematics showing the profile assignment of each nucleus. Thalamic nuclei not sampled are uncolored.

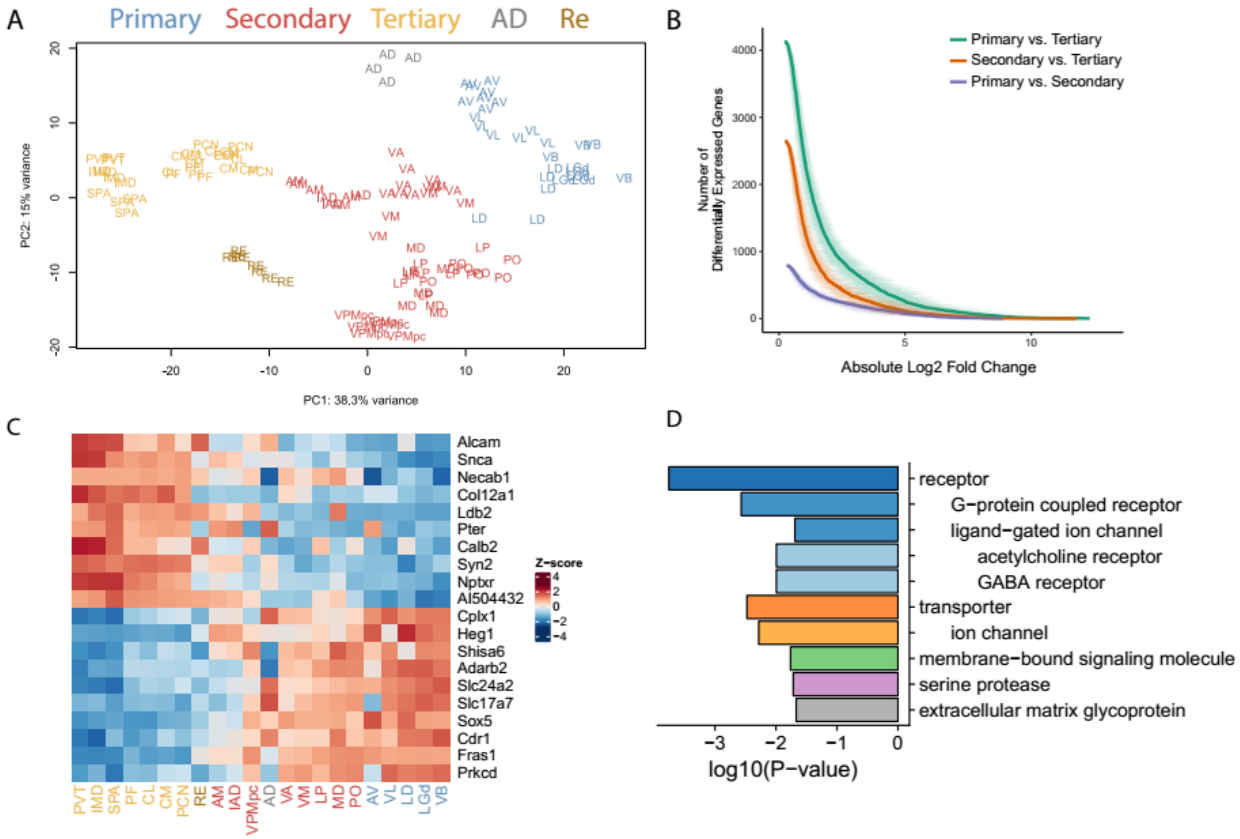


Figure 2 – A single axis of variance separates major thalamic profiles

A. Principal component analysis reveals separation of functional nuclear profiles in the first two principal components (>50% combined explained variance). The underlying gene set and color scheme are the same as in 1C. Gray = AD. Blue = primary thalamus, Red = secondary thalamus, Orange = tertiary thalamus. Brown = Reuniens.

B. Primary nuclei are farthest from tertiary nuclei, with secondary nuclei being intermediate. Plot shows the number differentially expressed genes at each log fold change level for the three comparisons. Differential gene expression was analyzed using edgeR (see methods). A gene was considered to be differentially expressed at a false discovery rate below 0.05. To avoid bias due to differences in sample number (i.e. more degrees of freedom), groups were subsampled

(with replacement) to the size of the smallest group. Shaded bands correspond to standard deviation of bootstrapped coefficients expressed as log2 fold change over 100 iterations.

C. Heatmap of genes with strongest positive and negative loadings on principal component 1. Nuclei are ordered by their mean position on principal component 1 of 2A. Colors represent gene-wise Z-scores.

D. Genes relevant to neurotransmission are overrepresented amongst the top 100 genes with the highest absolute PC1 loadings in our dataset. The ten most highly overrepresented PANTHER protein class terms (a consolidated version of gene ontology for molecular function) are shown. P-values based on hypergeometric test. Indentation indicates gene subfamily.

To better understand the relationship between these profiles, we performed principal component analysis on our data and compared this to the unsupervised profiles. We found that the first principal component (38% explained variance) separated nuclei into the same major profiles identified via hierarchical clustering (Figure 1C). Based on their relative positions in this first component, we named the three identified profiles that contained more than one nucleus primary, secondary, and tertiary. This progressive difference from primary to tertiary nuclei was also evident in the number of genes differentially expressed between the groups, with the primary and tertiary nuclei being most distinct, and the other two comparisons being less so (Figure 2B). Genes contributing to the primary axis separating nuclei included many neurofunctionally relevant genes (Figure 2C), with receptors, ion channels, and signaling genes being most significantly enriched among genes contributing to this axis (Figure 2D). However, other gene classes including kinases, synaptic genes and cell adhesion molecules were also present (Figure 2D). Thus, the expression of a diverse set of genes implicated in neuronal signaling contributes to the progressive differences between nuclei of primary, secondary and tertiary profiles.

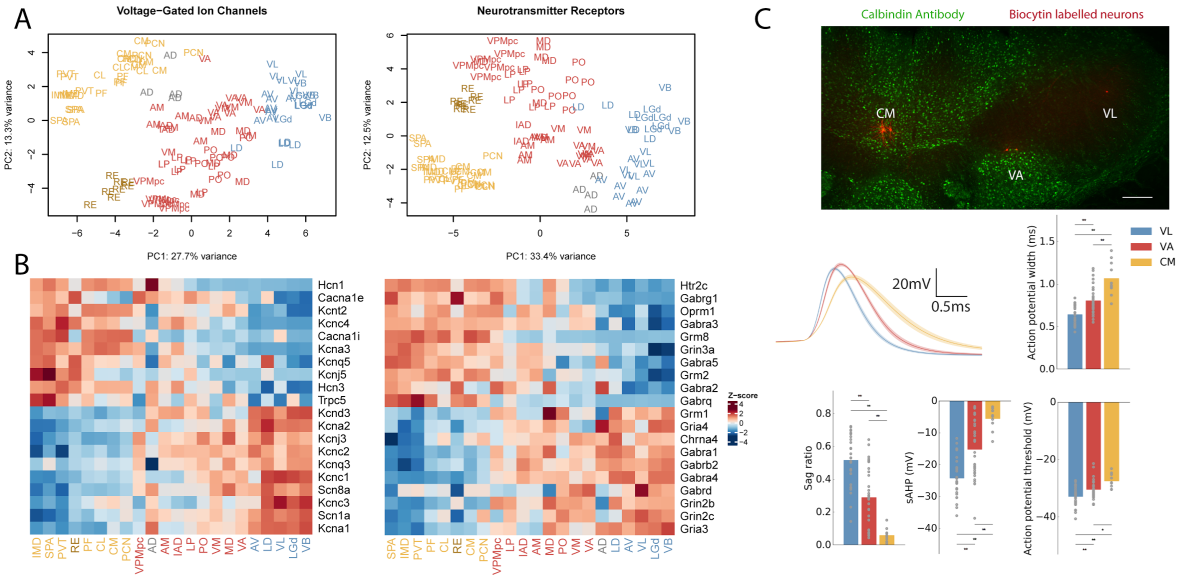


Figure 3 – *Functionally relevant genes vary systematically across nuclear profiles*

A. Principal components analysis including only genes encoding voltage-gated ion channels and neurotransmitter and neuromodulator receptors. Colors as in figure 1C.
B. Heatmap for genes that have the highest loading in principal component 1 from figure 3A. Voltage-gated ion-channels on the left and neurotransmitter receptors on the right. Colors represent gene-wise Z-scores.
C. Confirmation of electrophysiological variation across profiles. Whole cell recordings from ventrolateral (primary thalamus), ventroanterior (secondary thalamus) and central medial thalamus (tertiary thalamus). Top row: Neurons were labelled with biocytin (red) and localized to individual nuclei with the aid of Calbindin-D28K immunolabelling (green). Scale bar = 100um. Middle row, left shows average (mean) action potential shape for VL, VA and CM neurons (error bars are standard error of the mean). Remaining panels show comparisons for four physiological measurements across these nuclei (One way ANOVA with post-hoc Tukey HSD test, all comparisons P<0.05). See supplementary figure 2 for measurements that were not significant for all comparisons. Sample contained 29 VL neurons, 34 VA neurons and 10 CM neurons.

We next tested whether this axis of variance could provide insights into functional differences between thalamic groups. We again performed principal component analysis again, this time providing only voltage-gated ion channel or neurotransmitter/modulator receptor encoding genes (Figure 3A, left and right respectively). Analysis with this limited gene set reproduced the separation of groups in the first principal component (Figure 3B), demonstrating that nuclei of the same profile have a similar ion channel and receptor profile. We noted that expression of genes linked to high firing rates via faster channel kinetics tended to be progressively elevated toward primary group nuclei, and hypothesized that action potential width may progressively narrow from tertiary to primary nuclei. These genes include Kv3 channels (*Kcnc1*,*Kcnc3*), the *Scn8a* channel, and the *Kcnab3* subunit [2-4]. Whole cell recordings from the motor-related nuclei CM, VA and VL (representing the three main nuclear groups) confirmed this prediction (Figure 3D, top row), with VL having the narrowest action potential width and CM the widest. Thalamic

nuclear groups identified through RNAseq thus vary in a functionally relevant axis. This suggests that the timescale of signal processing is a key organizing principle across thalamic pathways[5].

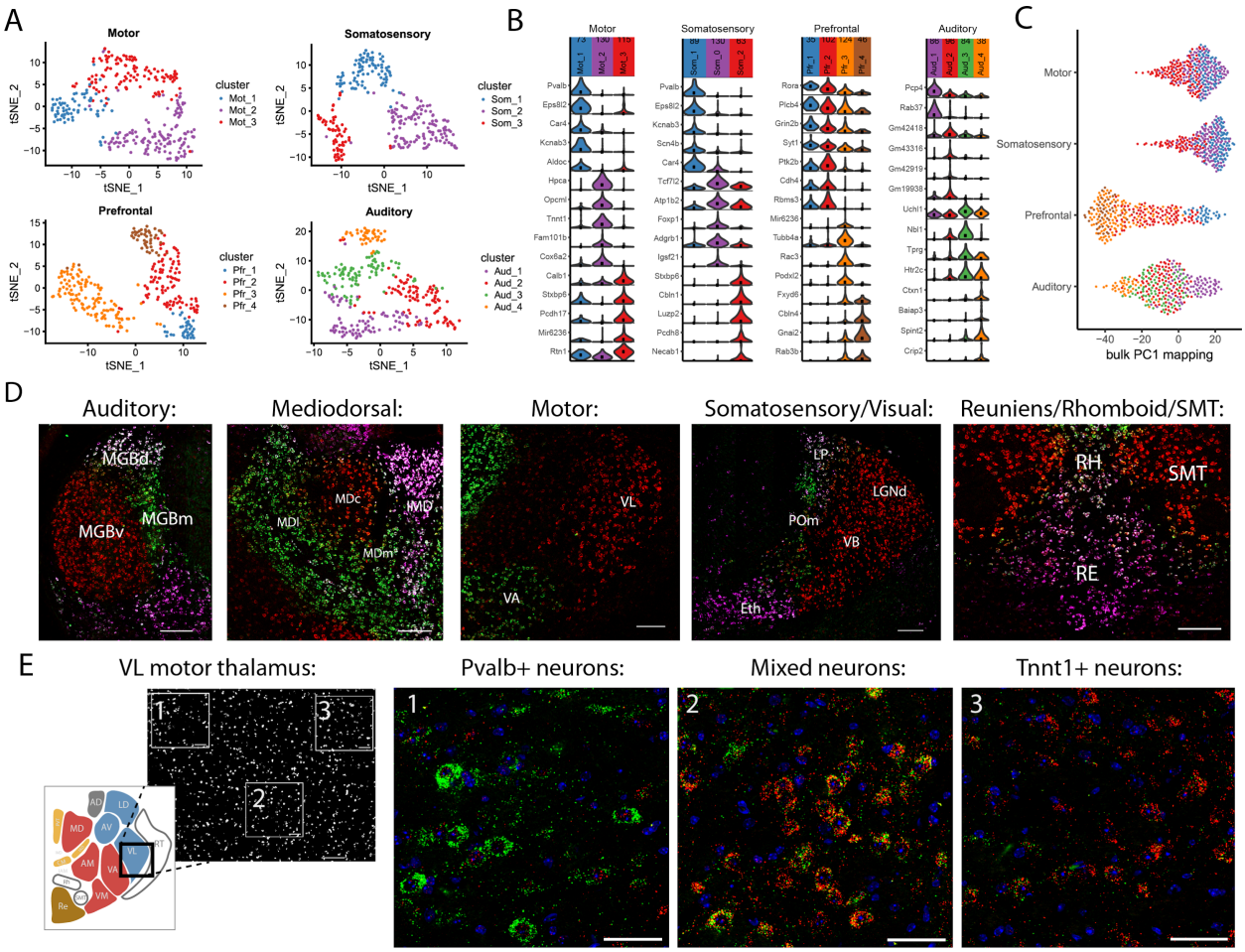


Figure 4 - A topographic spectrum of thalamic cell identities within and between thalamic nuclei

A. Overview of clusters (identified using the Seurat package, see methods) within each modality visualized using tSNE, cells colored by cluster identity.
B. Violin plots for top marker genes for each cluster (Likelihood Ratio Test, p-value<10^-5, logFC>0.5, top 4-5 signal-to-noise ratio).
C. Projection of single-cell data onto bulk principal component 1 (as in figure 2), each dot is a single cell, colored by the clusters from A and B. See supplementary figure 4 for separate plotting of each cluster.
D. Topographic distribution of marker genes within 6 major thalamic modalities. Multi-fluorescent *in situ* hybridization with probes for *Calb2* (pink), *Tnnt1* (red) and *Necab1*

(green). See supplementary figures 5,6,7 for expanded views and quantification. Scale bars = 200um.

E. Multi-fluorescent *in situ* hybridization within VL thalamus comparing genes marking clusters in single-cell data. Left panel shows field of view (scale bar in bottom right corner = 100um). Right three panels show (scale bars = 50um) expansion of the three sub-boxes, moving left to right. Middle box shows intermediate cells expressing both single cell cluster markers. (Red = *Tnnt1* and Green = *Pvalb*, Blue = DAPI). Coronal sections.

Thus far we analyzed thalamic nuclei by pooling projection neurons from their anatomical position. The resulting transcriptional profiles could represent homogenous populations, discrete subtypes, or cells with graded differences. To probe these possibilities, we profiled the transcriptomes of individual neurons from motor, somatosensory, auditory, and prefrontal projection classes (supplementary table 3). Analysis of the single-cell RNAseq dataset resulted in multiple clusters for each projection class which, when projected onto the first principal component of our bulk RNAseq dataset, spanned the same axis of variance identified from bulk data (Figure 4 A,B,C, supplementary figure 4). Markers for this axis were spatially separated at the single-cell level, with clear intermediate cells present (Figure 4D, supplementary figure 5,6,7). However, some individual single-cell clusters had similar PC1 positions (e.g. clusters 1 and 2 of the motor projections neurons, or clusters 1 and 2 of the somatosensory nuclei, supplementary figure 4). This suggests that distinct clusters within a single modality could exist in similar anatomical positions. We examined this possibility by performing RNAscope for genes which distinguished these clusters (*Pvalb* and *Tnnt1*) . Using motor thalamus as an example, *Pvalb* and *Tnnt1* expression were both found within the confines of a single thalamic nucleus, VL. While neurons selectively expressing each gene could be found, intermediate cells were also apparent. *Pvalb*-selective, intermediate, and *Tnnt1*-selective cells were distributed along the mediolateral axis (Figure 4E). Therefore spatially organized genetic gradients can exist even within individual thalamic nuclei.

In this study we produced a near-comprehensive transcriptomic atlas of thalamus via RNAseq in order to investigate its organizational structure. Though input modality has historically defined thalamic nuclei, several alternative, overlapping organizational classifications have been proposed [1,6]. Early work perturbing thalamus showed that nuclei of the same modality can have strikingly different contributions to behavior, leading to a division between arousal-linked intralaminar nuclei and the rest of thalamus[7-9]. More recently, theories developed in the sensory thalamus have subdivided thalamic nuclei into two groups, based upon anatomical or neurochemical features[10-12]. However, precise delineation of groups is not consistent between theories, and no theory is comprehensive. If and how these later theories apply beyond sensory thalamus, for example, is unknown.

Here, we reconcile and extend these theories by placing almost all thalamic nuclei into one of three profiles, lying on a single topographic axis identified via unsupervised analysis. This axis extends from the midline/intralaminar groups through to a profile that includes the classic ‘relay’ sensory nuclei. Though prior theories did not incorporate electrophysiological properties, we

found that genes directly affecting these properties were highly differentially expressed between groups, implying conserved, systematic variation of response properties across thalamus. Understanding how the pattern of intra-thalamic variability identified here is used to support different circuit functions will be an important challenge[13].

Prior theories of thalamic organization have placed thalamic cells into discrete nuclei, and thalamic nuclei into discrete groups. However, it remains unclear if subtypes of cells exist within nuclei and how discrete vs. continuous the cells type differences are. Recent studies in other non-thalamic areas have argued that gradients of projection neuron identities can exist[14,15]. Our single-cell transcriptomic data revealed heterogeneity within thalamic nuclei and suggest that thalamic genetic profiles exist on a spectrum, with mixtures of genetic profiles often existing in single neurons. Intermediate cell types exist both within and between nuclei, and are topographically localised. Thalamic projection neurons therefore exist on a topographic spectrum, rather than in discrete cell classes, suggesting that not all inputs to a given thalamic nucleus are treated the same. Future work should investigate whether anatomical properties of thalamic neurons also rest on the same, topographically organized spectrum.

Together, this open access dataset provides the information critical for more detailed investigation of individual nuclei and thalamus-wide organization. Our analysis shows striking covariation in the organization of thalamic pathways serving all modalities, establishing a comprehensive transthalamic organizational scheme based upon the entire transcriptome.

Conflicts of interest

The authors declare no competing interests.

Author contributions:

JP: contributed to all aspects of this project. AS: Analyzed and collected data, planned project and wrote the paper. EH: Planned project and collected data. CL: Collected and analyzed electrophysiology data. BS: Collected data. WK: Supervised project, AL: Collected data and developed methods JD: Supervised project SN: Supervised project. AH: Initiated and supervised project, wrote the paper.

Acknowledgements

This project was funded as a small project team (ThalamoSeq) collaboration at the Janelia Research Campus, following a pilot project in the Dudman/Hantman labs. We thank Karel Svoboda, Albert Lee and Amy Chuong for critical input throughout the project. We thank Matthew Phillips and Yves Weissenberger for comments on the manuscript. We thank Vilas Menon, Damian Kao and Mark Cembrowski for help with single-cell RNAseq analysis and many helpful discussions. We thank Kim Ritola and the Janelia Virology core for production of viruses. We are grateful to Amy Hu for assistance with confocal microscopy. We thank Daniel Morozoff, Yajie Llang, Justin Little, Amy Chuong and Na Ji for surgical protocols and assistance identifying nuclei for dissection, and the Janelia vivarium services for animal care and surgeries.

References (includes methods references)

1. Jones, E. The Thalamus. *Cambridge University Press* (2007).
2. Rudy, B. & McBain, C. J. Kv3 channels: voltage-gated K+ channels designed for high-frequency repetitive firing. *Trends in Neurosciences* **24,** 517–526 (2001).
3. Freeman, S. A., Desmazières, A., Fricker, D., Lubetzki, C. & Sol-Foulon, N. Mechanisms of sodium channel clustering and its influence on axonal impulse conduction. *Cellular and Molecular Life Sciences* **73,** 723–735 (2015).
4. Leicher, T., Bähring, R., Isbrandt, D. & Pongs, O. Coexpression of the KCNA3B gene product with Kv1.5 leads to a novel A-type potassium channel. *J. Biol. Chem.* **273,** 35095–35101 (1998).
5. Fogerson, P. M. & Huguenard, J. R. Tapping the Brakes: Cellular and Synaptic Mechanisms that Regulate Thalamic Oscillations. *Neuron* **92,** 687–704 (2016).
6. Willis, T. *Cerebri Anatome*. (1664).
7. Hunter, J. & Jasper, H. H. Effects of thalamic stimulation in unanaesthetised animals. *Electroencephalography and Clinical Neurophysiology* **1,** 305–324 (1949).
8. Smith, Y. *et al.* The thalamostriatal system in normal and diseased states. *Front Syst Neurosci* **8,** 5 (2014).
9. Liu, J. *et al.* Frequency-selective control of cortical and subcortical networks by central thalamus. *Elife* **4,** e09215 (2015).
10. Jones, E. Viewpoint: The core and matrix of thalamic organization. *Neuroscience* (1998).
11. Sherman, S. M. & Guillery, R. W. The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences* **357,** 1695–1708 (2002).
12. Frangeul, L. *et al.* A cross-modal genetic framework for the development and plasticity of sensory pathways. *Nature* **538,** 96–98 (2016).
13. Guo, Z. V. *et al.* Maintenance of persistent activity in a frontal thalamocortical loop. *Nature* **545,** 181–186 (2017).
14. Gokce, O. *et al.* Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. *CellReports* **16,** 1126–1137 (2016).
15. Cembrowski, M. S. *et al.* Spatial Gene-Expression Gradients Underlie Prominent Heterogeneity of CA1 Pyramidal Neurons. *Neuron* **89,** 351–368 (2016).
16. Tervo, D. G. R. *et al.* A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. *Neuron* **92,** 372–382 (2016).
17. Hempel, C. M., Sugino, K. & Nelson, S. B. A manual method for the purification of fluorescently labeled neurons from the mammalian brain. *Nat Protoc* **2,** 2924–2929 (2007).
18. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Meth* **10,** 1096–1098 (2013).
19. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* 1–13 (2014). doi:10.1101/003236
20. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2012).
21. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).
22. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 31–21 (2014).

23. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
24. Butler, A. & Satija, R. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv* 1–18 (2017). doi:10.1101/164889

Methods:

EXPERIMENTAL METHODS

Animal care

Experimental procedures were approved by the institutional care and use committee (IACUC) at the Janelia Research Campus. Mice were housed on a 12 hour light/dark cycle, with ad libitum food and water. The majority of mice were 8-12 weeks old.

Acquisition of samples

Cells were fluorescently labelled to enable manual dissection. This was done through retrograde labelling via either viral or tracer injection into the major projection field of the nucleus of interests. For viral injections, rAAV2-retro viruses expressing flexed GFP or TdTomato under the CAG promoter were injected, with volumes of 50-100nL at 3 depths[16]. Typical survival time was 3 weeks post-injection. Viruses were prepared by Janelia Virus Services. Retrograde tracer labelling used the lipophilic tracer DiI (2.5mg/ml in DMSO, injecting volumes of 50-200nL per site, from Molecular probes) or Lumafluor red retrobeads (1:4 dilution in PBS, 50-200nL per site). For supportive anterograde viral labelling, AAV2/1-CAG-GFP and/or AAV2/1-CAG-BFP was used, while CTB-488 (Molecular probes) was used for retinal axon labelling (see supplementary table 1 and 2).

Manual cell sorting and RNAseq

Fluorescent cells were collected and sequenced as previously described[15]. Briefly, animals were deeply anaesthetized with isoflurane and euthanized. Coronal slices (200-300μm) were cut and then digested for 1 hour at room temperature with proteases and neural activity blockers in artificial cerebrospinal fluid (ACSF). Following digestion, relevant regions were microdissected. The resulting cell suspensions were then diluted with filtered ACSF and washed at least 3 times by transferring them to clean dishes. This process has been shown, and is shown here, to produce negligible contamination with non-fluorescent tissue (supplementary figure 1A)[17]. After the final wash, samples were aspirated in a small volume (1~3*μ*l) and lysed in 47μl XB lysis buffer using the Picopure kit (KIT0204, ThermoFisher) in a 200μl PCR tube (Axygen), incubated for 30 min at 40C on a thermal cycler and stored at -80℃.

Library preparation and sequencing was performed by the Janelia Quantitative Genomics core. RNA was isolated from each sample using the PicoPure RNA isolation kit (Life technologies) and on-column RNAase-free DNase I treatment (Qiagen). ERCC control RNA was added (Life technologies). Amplification was then performed using the Ovation RNA-seq v2 kit (NuGEN), yielding 4-8ug of cDNA. The Ovation rapid DR multiplexing kit (NuGEN) was used to make libraries for sequencing, which were sequenced on a HiSeq 2500 (Illumina).

Raw data from all collected samples, including validation samples, are available online (see below). At no stage were experimenters blinded to sample identity.

Sequencing of single cell RNAseq (scRNAseq) samples

Retrogradely labeled cells were dissected as described above, and collected into 8-well strips containing 3 uL Smart-seq2 lysis buffer, flash-frozen on dry ice, and stored at -80º until further use[18].

Upon thawing cells were were re-digested with Proteinase K, and barcoded RT primers added. cDNA synthesis was done using the Maxima H Minus RT kit (Thermo Fisher, Waltham, MA, US) and E5V6NEXT template switch oligo, followed by heat inactivation reverse transcriptase. PCR amplification using the HiFi PCR kit (Kapa Biosystems) and SINGV6 primer was performed with a modified thermocycling protocol (98ºC for 3 min, 20 cycles of 98ºC for 20s, 64ºC for 15s, 72ºC for 4 min, final extension at 72ºC for 5 min). Samples were then pooled across strips, purified with Ampure XP beads (Beckman Coulter), washed twice with 70% ethanol and eluted in water. These pooled strips were then combined to create the plate-level cDNA pool for tagmentation, and concentration was determined using Qubit High-Sensitivity DNA kit (Thermo Fisher Scientific).

Tagmentation and library preparation using 600 pg cDNA from each plate of cells was then performed with a modified Nextera XT (Illumina) protocol, but using the P5NEXTPT5 primer and tagmentation time extended to 15 minutes[19]. The libraries were then purified following the Nextera XT protocol (0.6x ratio) and quantified by qPCR using Kapa Library Quantification (Kapa Biosystems). 6-10 plates were run twice on a NextSeq 550 flow cell reading 25 bases in read 1, 8 bases in the i7 index read, and 125 bases in read 2. Read 1 contains the spacer, barcode, and UMI. Read 2 represents a cDNA fragment from the 3' end of the transcript.

RNAscope multiplex fluorescent *in situs*

C57Bl/6J mice (8 weeks old) were anesthetized with isoflurane then fixed via transcardial perfusion with PBS followed by 4% paraformaldehyde in PBS, pH 7.4. Brains were post-fixed at 4C overnight (~20hrs), washed 3 times with PBS, and cryoprotected in a sucrose series of 10%, 20% then 30% in PBS at 4C. All solutions were prepared RNAse-free. Brains were sectioned (14um) on a Leica CM3050S cryostat, mounted onto Fisher SuperFrost Plus slides, and stored at -80C.

In *situ* hybridization was performed using the RNAscope Multiplex Fluorescent Assay platform from ACDBio (Newark, CA, USA), following the manufacturer's protocol. The probes used were: Calbindin2 (ref 313641-C3), Necab1 (ref 428541 and 428541-C2), Rorb (ref 444271 and 444271-C3), and Tnnt1 (ref 466911-C2). Fluorescent dyes were DAPI, Alexa 488, Atto 550 and Atto 647. Images were acquired using a Zeiss LSM 880 confocal microscope. Images were acquired at with an air 20X (0.8NA) objective unless otherwise specified.

Electrophysiology methods

Acute brain slices were prepared from p20-25 mice. Animals were deeply anesthetized with ketamine/xylazine/acepromizine and transcardially perfused with ice-cold oxygenated cutting solution containing (in mM): 74 NaCl, 3 KCl, 1 NaH_2PO_4, 25 $NaHCO_3$, 6 $MgCl_2$, 0.5 $CaCl_2$, 5 Sodium Ascorbate, 75 Sucrose, 10 Glucose. 300 mm coronal slices containing the thalamus were cut on a vibratome (Leica), and then recovered for 15 min at 33 °C and for 15 min at room temperature in oxygenated cutting solution followed by at least another 1 hr at room temperature in oxygenated ACSF containing (in mM): 126 NaCl, 3 KCl, 1 NaH_2PO_4, 25 $NaHCO_3$, 2 $MgCl_2$, 2 $CaCl_2$, 10 Glucose. During recordings, slices were perfused with oxygenated 34-35 °C ASCF with 35 mM d,l-2-amino-5-phosphonovaleric acid (APV), 20 mM 6,7-dinitroquinoxaline-2,3-dione (DNQX) to block glutamatergic synaptic transmission and 50 mM picrotoxin to block GABAergic synaptic transmission. Target neurons in the central medial (CM), ventral anterior (VA) and ventral lateral (VL) nuclei of the thalamus were identified based on their distance to the mammillothalamic tract and nuclear borders were confirmed with calbindin immunostaining *post hoc*. Whole-cell recording pipettes (6 – 8 MW) were filled with internal solution containing (in mM): 100 K-gluconate, 20 KCl, 10 HEPES, 4 Mg-ATP, 0.3 Na-GTP, 10 Na-phosphocreatine, and 0.1% biocytin. Current-clamp recordings were obtained with Multiclamp 700B amplifiers (Molecular Devices) digitized at 10 kHz using IGOR Pro (WaveMetrics). Resting membrane potentials were adjusted to -65 mV and steady state series resistance was compensated. Neurons with series resistance > 30 MΩ or membrane potentials that changed by > 3 mV were excluded.

Custom IGOR scripts were used to analyze the data. For each neuron, threshold, amplitude, afterhyperpolarization and half width at half height of the 16-19 action potentials in trials with 20 to 40 Hz firing rate were averaged.

Immunohistochemistry

Recorded slices were fixed with 4% paraformaldehyde and 2.5% sucrose in 5x phosphate-buffered saline (PBS) at 4 °C for 2-10 days. After wash with PBS, slices were blocked in PBS with 0.3% Triton and 3% BSA at 4 °C overnight and then incubated in PBS with 0.3% Triton and 3% BSA and rabbit anti-calbindin D-28k (Swant, 1:1000) at 4 °C overnight. After washing, they were incubated in PBS with 0.3% Triton, 3% BSA and 5% goat serum with fluorescent protein conjugated goat anti-rabbit IgG (Invitrogen, 1:1000) and streptavidin (Invitrogen, 1:1000) at 4 °C overnight to label calbindin-expressing neurons and biocytin-filled neurons.

ANALYSIS METHODS

Bulk RNAseq analysis

Data processing and quality control
After removing Illumina adapter sequences using cutadapt, reads were mapped to the mouse reference genome (GRCm38) using STAR with 'ENCODE settings' for RNAseq[20]. Mean mapping rate was 82.29% with a standard deviation of 2.25%. Unique unambiguous exon-mapping reads were summarized at the gene level using Gencode version M13.

Contamination with common astrocytic, oligodendrocytic, erythrocytic and microglial transcripts was low, consistent with a lack of substantial contamination by non-fluorescent cells (Supplementary figure 1a). To ensure the specificity of our dissections and to control for potential batch effects, we collected several nuclei through multiple independent labelling approaches, and showed that these samples cluster in a highly similar manner (Supplementary figure 1c).

Differential gene expression
Differential expression was assessed using the Bioconductor package edgeR[21]. Low counts were removed by requiring a Transcripts per million (TPM) > 5 in at least 3 samples. This yielded a list of approx. 17,000 expressed genes. Counts were then fitted to a negative binomial generalized linear model, where each factor level represents a different thalamic nucleus, and a log ratio test was used to assess differential expression between groups. P-values were adjusted for multiple tests using the Benjamini-Hochberg method. Genes with false-discovery rate < 0.05 were considered differentially expressed. For selecting the most differentially expressed genes between any thalamic nuclei, we used an ANOVA-like test (as in edgeR User manual 3.2.6), testing for differences between any nuclei, and used the 500 genes with the highest P-value.
For visualization, clustering and machine learning of gene expression data, we used variance-stabilized counts produced by the variance-stabilizing transformation in the DESeq2 R package)[22,23] .
For assessing the role of modality vs. hierarchical class on distinguishing thalamic nuclei, we used elastic-net regularized logistic regression classifiers. Models were trained with different numbers of randomly selected genes as features over 100 iterations. To avoid bias due to variable group size, groups were subsampled to the size of the smallest group. Model tuning was performed using the glmnet and caret packages in R, and accuracy of the best model was assessed using 5-fold cross-validation.

Unsupervised clustering and functional enrichments
Hierarchical clustering was performed using 1 - Spearman's correlation as a distance metric, and groups were defined by splitting the tree at the level of 5 branches. Principal component analysis was done using the singular value decomposition based prcomp function in R. For functional enrichment of differentially expressed genes, we used the PATHER Protein Class Ontology (http://data.pantherdb.org/PANTHER13/ontology/Protein_Class_13.0), which is a

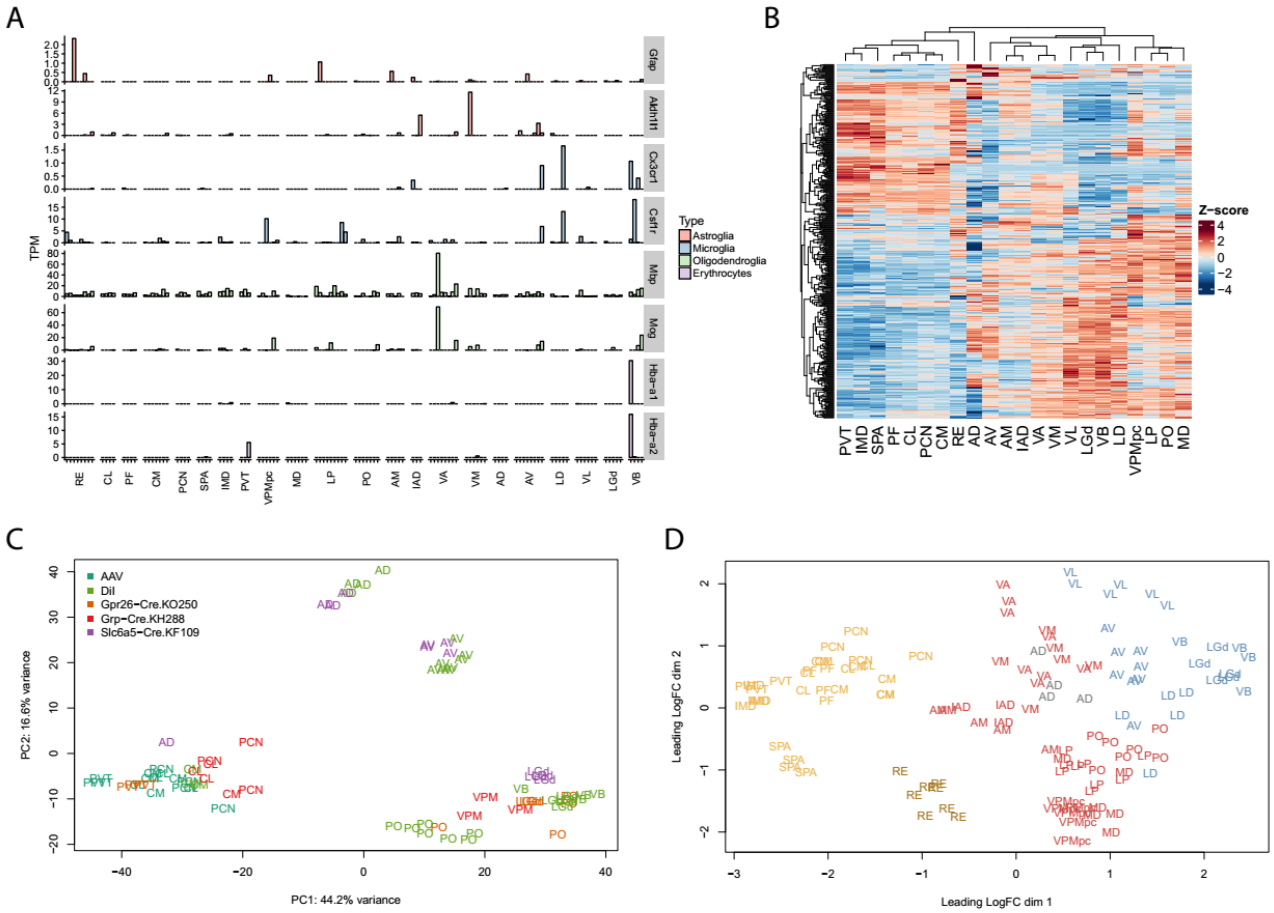
consolidated version of molecular function gene ontology. Over-representation in the top 100 genes with the highest PC1 loadings was assessed via hypergeometric test.
For defining voltage-gated ion channels and neurotransmitter receptors, we downloaded the IUPHAR-BPS database (http://www.guidetopharmacology.org/DATA/targets_and_families.csv). Voltage-gated ion channels were the genes defined in the database, while for neurotransmitter receptors we included ionotropic and metabotropic receptors for glutamate, GABA, glycine, acetylcholine, 5-HT, dopamine, trace amine, histamine, and opioids.

Single-cell RNAseq analysis

Data processing, and quality control
Single cell RNAseq data was trimmed for adapters using cutadapt and aligned to the mouse genome (GRCm38) using STAR. To demultiplex cells, collapse UMIs and produce gene-wise counts for each cell, we used a modified version of Dropseqtools.
Single-cells were required to have more than 20,000 UMIs and more than 2,500 genes detected per cell, which yielded a total of 1,220 cells (supplementary figure 3A). Of these, 9 cells were found to be significantly contaminated with oligodendrocyte transcripts (supplementary figure 3C), leaving 1,211 cells for all downstream analysis. Genes were considered expressed if their expression was detected in more than 10 cells.
Our single cell sequencing was not comprehensive, and with improved sequencing approaches further genetic subdivisions may be identified. Single-cell and bulk dissections were not precisely matched, for example motor-projecting midline nuclei were not dissected for single-cell. However, bulk and single-cell RNAseq is in close agreement, indicating that our results are robust to collection method (supplementary figure 3B).

Single-cell clustering and marker genes
Single-cell clusters were defined using the Seurat R package (version 2.0)[24]. Data were log transformed and scaled. Single-cell clustering was performed separately for each modality using the graph-based clustering approach in Seurat and the top 10 principal components for calculating euclidian distances. Clusters of cells were visualized using t-distributed stochastic neighbor embedding (tSNE) using the top 3 principal components as input. Marker genes for each cluster were required to be expressed in at least 80% of the cells in the cluster, to have a P-value <10^-5 (likelihood ratio test), a log2 fold change > 0.5, and were ordered by signal-to-noise ratio (ratio of log fold changes for pairs between and within clusters) with top 4-5 genes listed for each cluster. Projection of single-cell data onto bulk principal components was obtained by multiplying (inner product) log-transformed and scaled single-cell data by the bulk principal component loadings.

Supplementary Figures



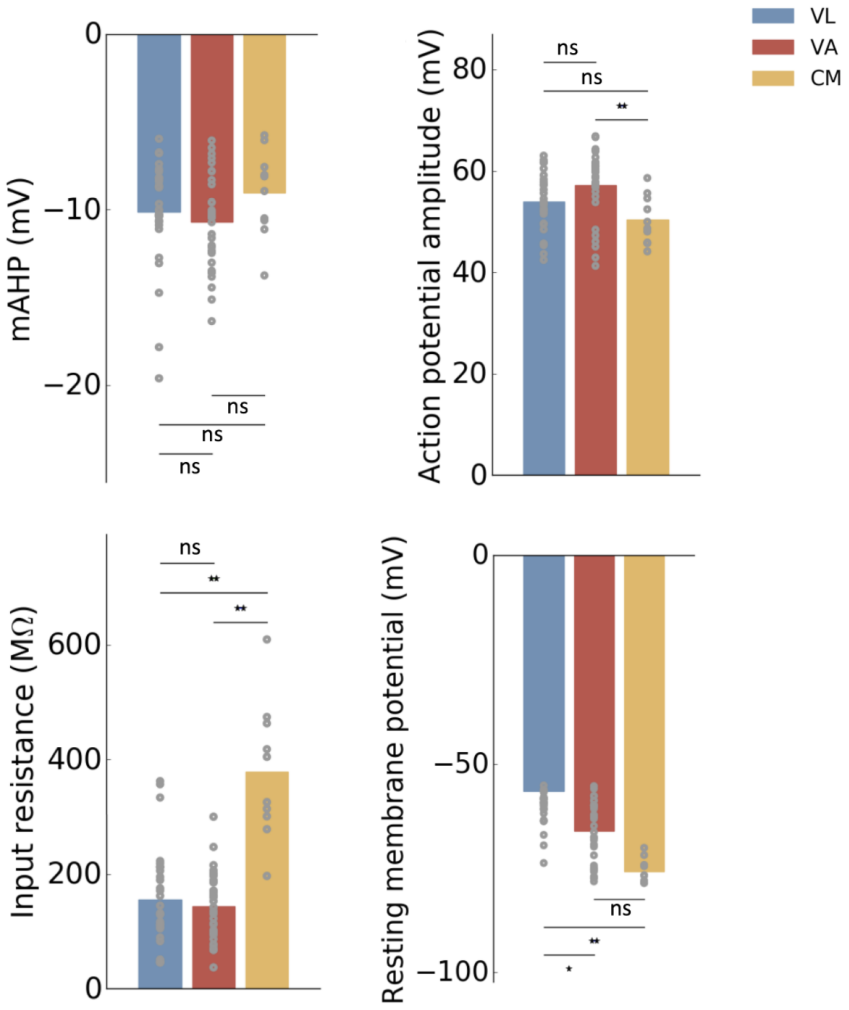
Supplementary figure 1 - Bulk Sequencing quality controls and additional analysis

A) Markers of non-neuronal sample contamination are low across our dataset. Expression (TPM) in bulk samples shown for 8 genes marking astroglia (top two rows), microglia (third and fourth rows), oligodendrocytes (five and sixth rows) and erythrocytes (bottom rows). Only a small number of samples showed concordant expression of contamination markers.

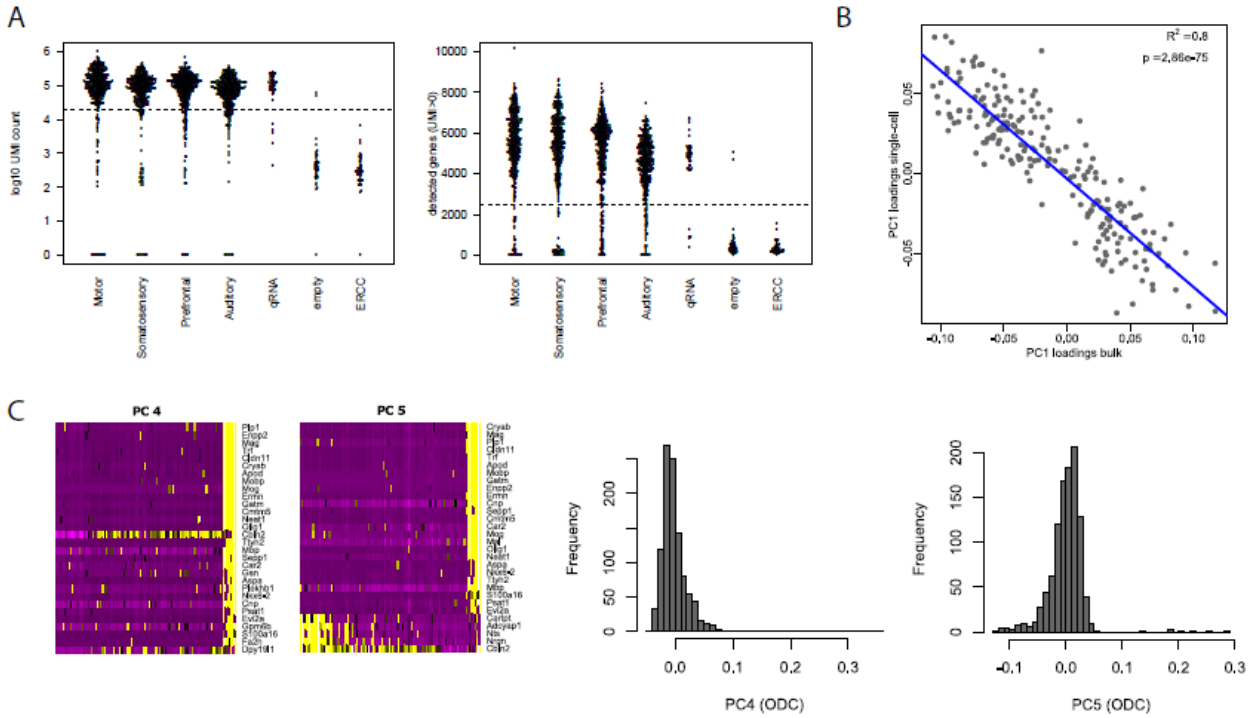
B) Heatmap of the top 500 differentially expressed genes, ranked by hierarchical clustering using Euclidean distance. Colors represent gene-wise Z-scores.

C) Samples of the same nucleus obtained via different labelling methods cluster similarly. Principal components analysis of those samples, for which multiple collection methods were used (i.e. GENSAT lines in addition to retrograde labeling) using the top 500 genes with highest variance. Samples are colored by collection approach or transgenic line used.

D) Multidimensional scaling using an alternative distance metric also identifies the same leading axis of variance. Distance was defined as the quadratic mean of the log fold changes of the top 500 differentially expressed genes between any two samples (meaning that the gene set used for the distance comparison varies between each sample comparison made).



Supplementary Figure 2 - Additional electrophysiology comparisons made
All statistical tests and experimental details are the same as in figure 3D.

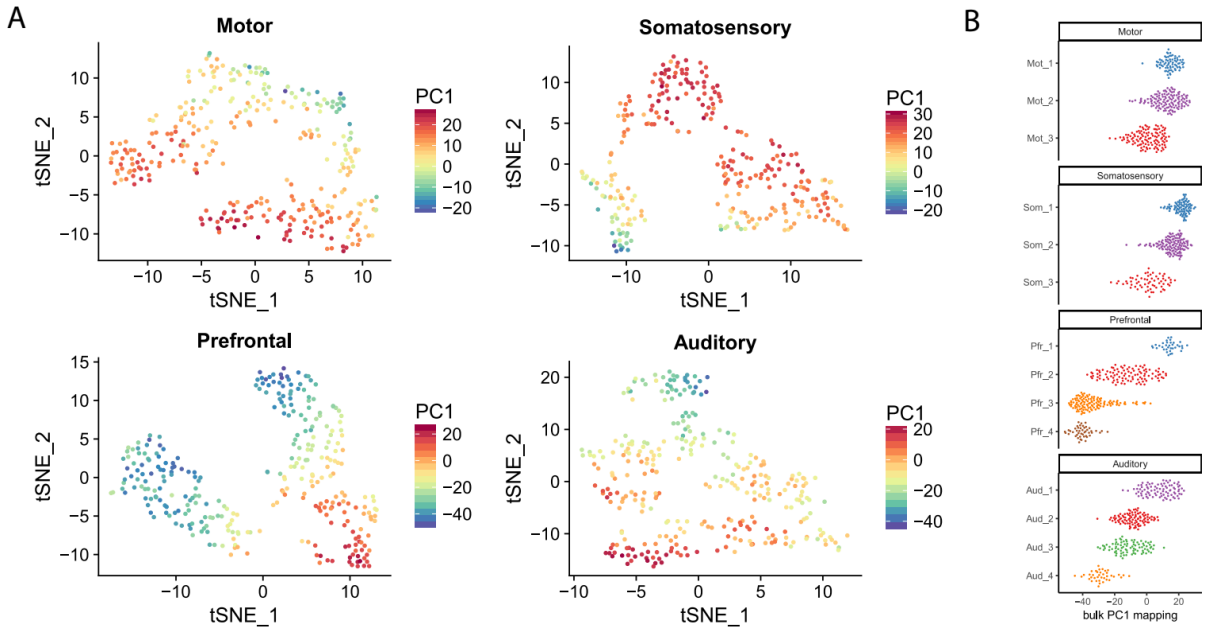


Supplementary Figure 3 - Single Cell Quality Control data

A) UMI count (left) and gene detection rate (right) for all collected single cells. Cutoffs for downstream use were >20,000 detected transcripts and >2,500 detected genes, and are indicated by dashed lines.

B) PC1 loadings for the most differentially expressed genes in bulk (gene set as in figure 1) are highly similar in bulk and single-cell data.

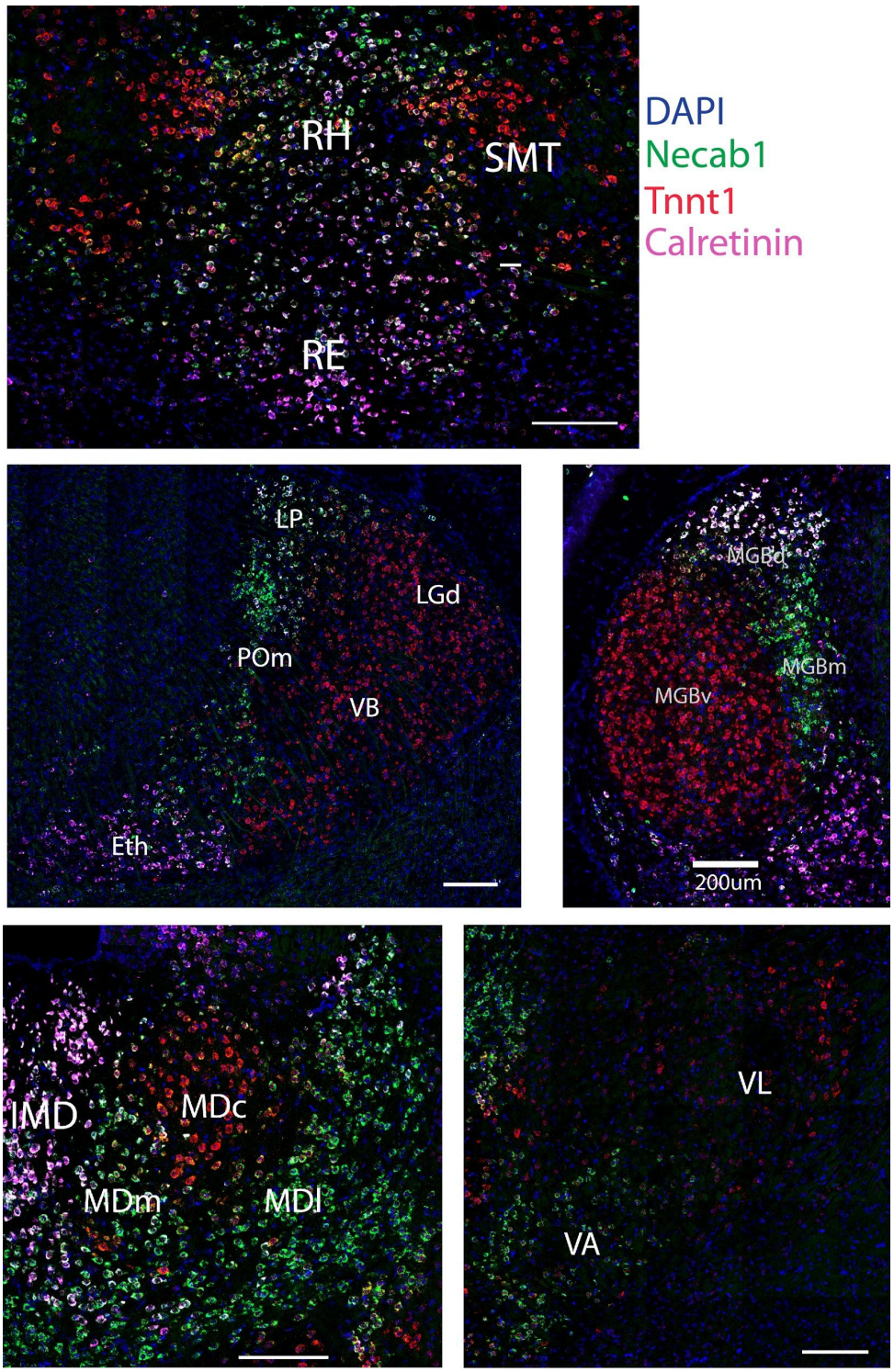
C) Principal component analysis on the single-cell data revealed that PC4 and PC5 represented non-neuronal contamination from oligodendrocytes. For all downstream analysis, the 9 cells whose PC4 and PC5 positions were above 0.1 were removed.



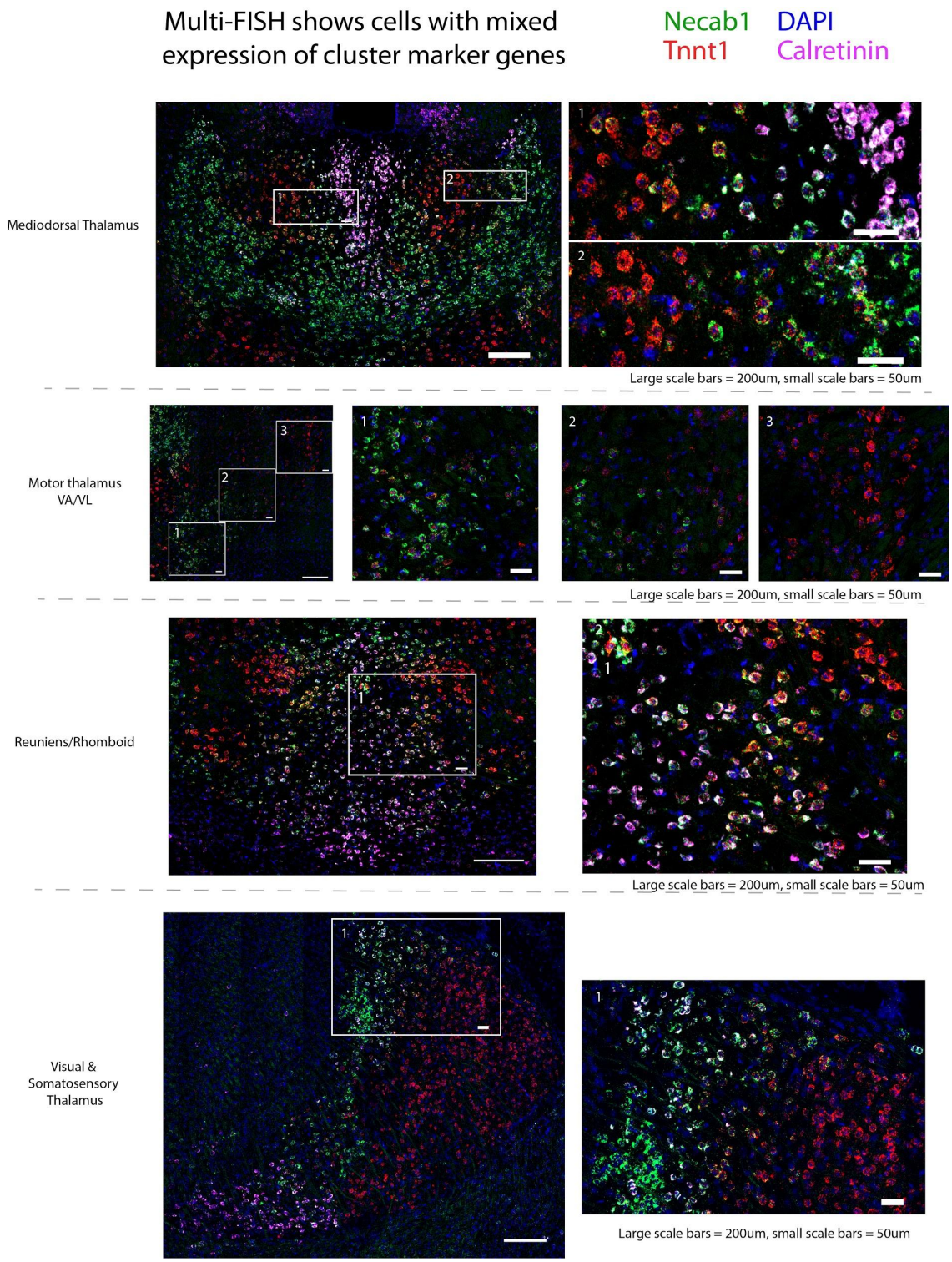
Supplementary Figure 4 - Single Cell Projection onto Bulk Data

A) t-SNE plots for each projection type with cells colored by its projection onto bulk Principal component 1. Cells from the negative extreme of principal component 1 are absent from motor and somatosensory modalities, which may reflect a lack of sampling from the more medial thalamic sections in these dissections.

B) Positions of single cells projected onto bulk PC1, plotted separately for each single-cell cluster.

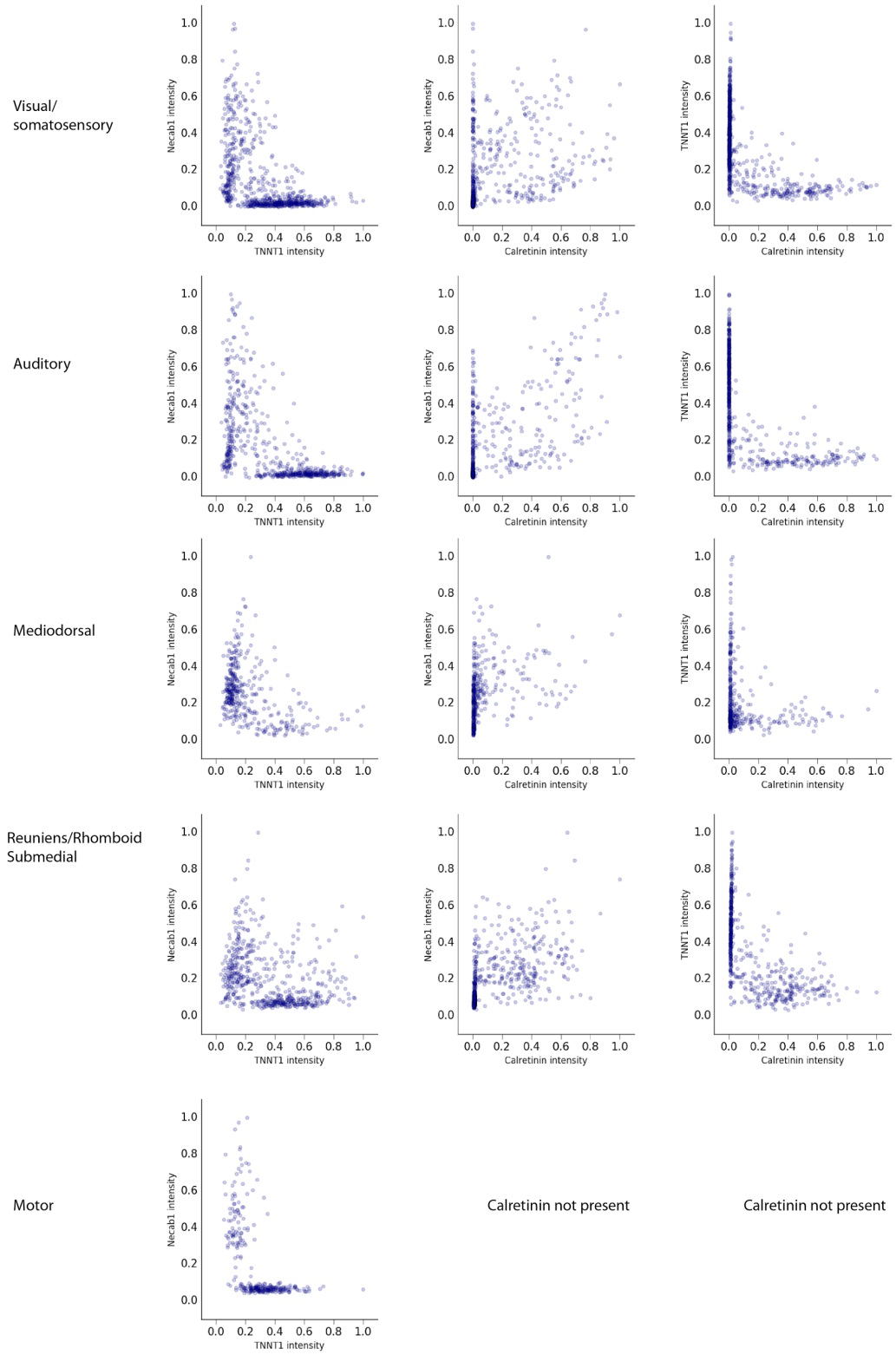


Supplementary Figure 5 - Enlarged view of multplex-FISH
Shows the images from figure 4D in enlarged form. Scale bars all 200um.



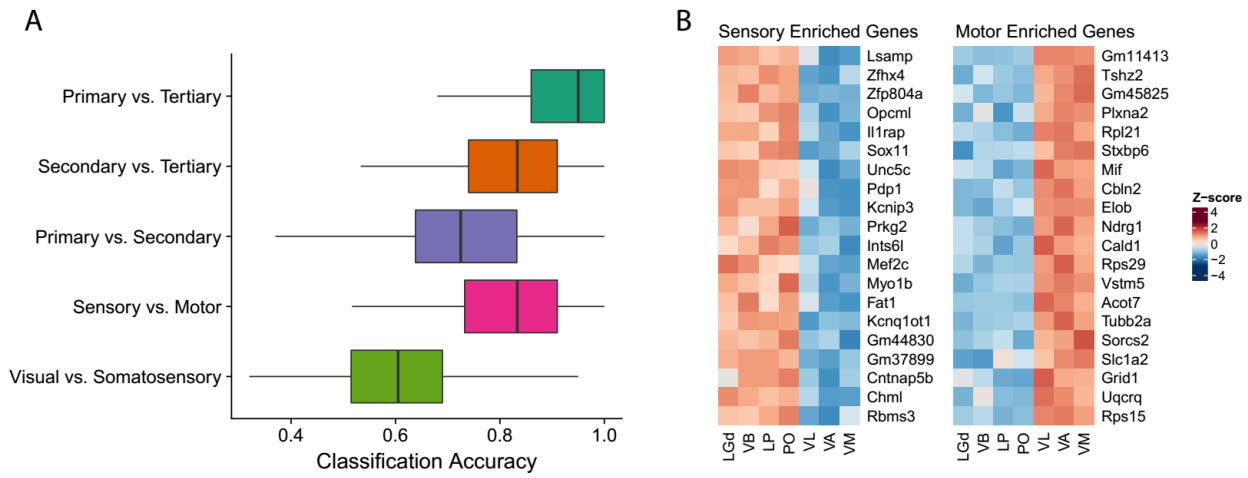
Supplementary Figure 6 - Gradients within RNAscope data

Expanded views of example regions showing intermediate cells expressing combinations of *Tnnt1*, *Necab1* and *Calb2*, which are preferentially expressed in primary, secondary and tertiary nuclear groups respectively. Each row is a different thalamic modality.



Supplementary Figure 7 - Quantification of multiplex-FISH images shows intermediate cell types

Quantification of multiplex-FISH gene expression images. Regions of interest (ROIs) were drawn in imageJ, and then processed in custom python scripts. Intensity was normalised first to the ROI size, then divided by the maximum for that channel. Cells without clear labelling were not drawn, to ensure glia were not included .This will also exclude neurons that lack expression of these genes.



Supplementary Figure 8 - Genetic differences between sensory and motor nuclei

A) Classification accuracy for distinguishing Primary, Secondary, and Tertiary type nuclei, as well as for distinguishing motor (VL,VA,VM) vs. sensory (LGd, LP, VB, PO) and visual (LGd,LP) vs. somatosensory (VB,PO) nuclei. Classifiers were obtained using elastic-net regularized logistic regression using 20 random genes as features (randomly selected from a pool of all genes expressed at >5 TPM across any 3 samples) assessed over 100 iterations. To prevent bias due to sample size difference, groups were subsampled to the size of the smallest group (n=9) at each iteration. Accuracy was assessed using 5-fold cross-validation. A value of 1 corresponds to perfect classification, while a value of 0.5 corresponds to chance level performance.
B) Genes that best distinguish motor from sensory nuclei (LGd,VB,LP vs. VL,VA,VM). Plotted are the top 20 genes with false discovery rate < 10^-3 (likelihood ratio test), fold change > 2, and ordered by highest signal-to-noise ratio (mean log fold change between vs. within group).

Supplementary tables

Supplementary table 1 – Sample collection approaches (Bulk RNASeq)
'Validation' refers to samples in supplementary figure 1C. For transgenic lines used, see supplementary table 4.

Region name:	Labelling methods:
Anterior dorsal (AD)	Retrograde DiI from retrosplenial cortex. Validation: Transgenic line
Anterior medial (AM)	Retrograde DiI from cingulate cortex.
Anterior ventral (AV)	Retrograde DiI to retrosplenial cortex. Validation: transgenic line
Central lateral (CL)	Retrograde virus to striatum. Validation: transgenic line.
Central medial (CM)	Retrograde virus to striatum. Retrograde DiI to striatum.
Interanterodorsal nucleus (IAD)	Retrograde DiI to cingulate cortex
Intermediodorsal nucleus (IMD)	Retrograde virus to striatum.
Laterodorsal nucleus (LD)	Retrograde DiI from retrosplenial cortex.
Lateral Geniculate (dorsal) nucleus (LGd)	Retrograde DiI from visual cortex with anterograde CTB-488 tracer injected to retina. Validation: transgenic line.
Lateral posterior nucleus (LP)	Retrograde DiI from visual cortex with anterograde CTB-488 tracer injected to retina.
Mediodorsal nucleus (MD) - Both subdivisions	Retrograde DiI to frontal cortex.

Subparafascicular nucleus	Retrograde virus from striatum.
Paracentral nucleus (PCN)	Retrograde virus from striatum. Validation: transgenic line.
Parafascicular nucleus (PF)	Retrograde virus from striatum.
Paraventricular thalamus (PVT)	Retrograde virus from striatum Validation: transgenic line
Posterior medial nucleus (PO)	Retrograde DiI from somatosensory cortex. Validation: transgenic line
Reuniens nucleus (RE)	Retrograde DiI from cingulate cortex
Ventral anterior nucleus (VA)	Retrograde DiI from M1, Anterograde labelling of inputs from Basal Ganglia and Cerebellum with viral tracers (see supplementary table 2).
Ventral medial nucleus (VM)	Retrograde DiI from M1. Anterograde labelling of inputs from Basal Ganglia and Cerebellum with viral tracers (see supplementary table 2).
Ventral lateral nucleus (VL)	Retrograde DiI from M1. Anterograde labelling of inputs from Basal Ganglia and Cerebellum with viral tracers (see supplementary table 2).
Ventrobasal nucleus (VB)	Retrograde DiI from primary somatosensory cortex (S1)
Ventroposteromedial parvocellular (VPMpc)	Retrograde DiI from insula

Supplementary table 2 – Coordinates for retrograde labelling/trace injections (Bulk RNASeq)

All depths relative to brain surface. If depth not stated, injections were made at 300um and 600um deep.

Sample area:	**Injection coordinates (in millimeters)**	***Additional comments***
AD	1.7 caudal, 0.25 lateral,	

	0.4,1 deep.	
AM	1.35 rostral, 0.2 lateral, 1.25+1.8mm depth	
AV	1.7 caudal, 0.25 lateral, 0.4,1 deep.	
CL	0.7 rostral, 1.9 lateral, 2.75, 3.00, 3.25 deep.	
CM	0.7 rostral, 1.9 lateral, 2.75, 3.00, 3.25 deep.	
IAD	1.35 rostral, 0.2 lateral, 1.25+1.8mm depth	
LD	1.7 caudal, 0.25 lateral, 0.4,1 deep.	
LGd	2.9 caudal, 2.4 lateral. 2.9 caudal, 1.7 lateral. 3.8 caudal, 3.0 lateral 3.8 caudal, 2.0 lateral	Retina was injected with anterograde tracer to distinguish LGd from LP
LP	2.9 caudal, 2.4 lateral. 2.9 caudal, 1.7 lateral. 3.8 caudal, 3.0 lateral 3.8 caudal, 2.0 lateral	Retina was injected with anterograde tracer to distinguish LGd from LP
MD	2 rostral, 1 lateral, 3.0, 2.5,2.0 deep. 2 rostral, 1.6 lateral, 3.0,2.4,1.5 deep. 1.7 rostral, 0.4 lateral, 3.0, 2.4, 1.5 deep	
PCN	0.7 rostral, 1.9 lateral, 2.75, 3.00, 3.25 deep.	
PF	0.7 rostral, 1.9 lateral, 2.75, 3.00, 3.25 deep.	
PO	0.85, 1.5 caudal,2.75 lateral, 0.4 deep. 1.2 caudal, 2.25 lateral, 0.4 deep.	

RE	1.35 rostral, 0.2 lateral, 1.25+1.8mm dept	
SPA	0.7 rostral, 1.9 lateral, 2.75, 3.00, 3.25 deep.	
VA	0.3 rostral, 1.6 lateral 0.8 rostral, 1.6 lateral 0.5 rostral, 1.4 lateral	GFP was virally expressed in the substantia nigra to distinguish VA/M from VL.
VM	0.3 rostral, 1.6 lateral 0.8 rostral, 1.6 lateral 0.5 rostral, 1.4 lateral	GFP was virally expressed in the substantia nigra to distinguish VA/M from VL.
VL	0.3 rostral, 1.6 lateral 0.8 rostral, 1.6 lateral 0.5 rostral, 1.4 lateral	GFP was virally expressed in the substantia nigra to distinguish VA/M from VL.
VB	0.85, 1.5 caudal,2.75 lateral, 0.4 deep. 1.2 caudal, 2.25 lateral, 0.4 deep.	
VPMpc	0.25 caudal, 4 lateral, 2.4,2.7,2.9 deep.	

Supplementary table 3 – Injection coordinates for retrograde labelling (Single-Cell RNASeq)

Sample area:	Injection coordinates (in millimeters)
Auditory Thalamus	3.5 caudal, 4.5 lateral, 2.25 deep. 3.16 caudal, 4.5 lateral, 2.0 deep
Mediodorsal (Frontal) Thalamus	2 rostral, 1 lateral, 3.0, 2.5,2.0 deep. 2 rostral, 1.6 lateral, 3.0,2.4,1.5 deep. 1.7 rostral, 0.4 lateral, 3.0, 2.4, 1.5 deep
Motor Thalamus	0.3 rostral, 1.6 lateral 0.8 rostral, 1.6 lateral 0.5 rostral, 1.4 lateral
Somatosensory thalamus	0.85 and 1.5 caudal, 2.75 lateral, 0.4 deep. 1.2 caudal, 2.25 lateral, 0.4 deep.

Visual Thalamus	2.9 caudal, 2.4 lateral. 2.9 caudal, 1.7 lateral. 3.8 caudal, 3.0 lateral 3.8 caudal, 2.0 lateral

Supplementary table 4 – Transgenic mice used in this study:

Transgenic line:	**Areas collected**
Slc6a5.KF109 (GENSAT)	Anterior dorsal, Anterior ventral, Dorsal lateral geniculate (LGd).
Gpr26.KO250. (GENSAT)	Paraventricular nucleus (PVT), Posterior medial (PO)
Grp.KH288 (GENSAT)	Rostral intralaminar nuclei (Central lateral, paracentral), ventral posteromedial nucleus (VPM).