Title: A topographic axis of transcriptional identity across thalamic projection systems

Authors: ***James W. Phillips[1,3†,*], Anton Schulmann[1,2†], Erina Hara[1,¥], Chenghao Liu[2], Lihua Wang[1], Brenda C. Shields[1,ø], Wyatt Korff[1], Andrew Lemire[1], Joshua Dudman[1], Sacha B. Nelson[1,2,*], Adam Hantman[1,*]***

Affiliations:
[1]HHMI Janelia Research Campus, Ashburn, Virginia, USA

[2]Brandeis University, Waltham Massachusetts, USA

[3]Department of Physiology, Development and Neuroscience, University of Cambridge, UK

*Correspondence to: James Phillips (phillipsj10@janelia.hhmi.org), Sacha Nelson (nelson@brandeis.edu), Adam Hantman (hantmana@janelia.hhmi.org)
[†]Co-first authorship
[¥] Present Address: Department of Molecular, Cellular & Biomedical Sciences, CUNY School of Medicine, 160 Convent Ave, New York, NY 10031
[ø]Present Address: Duke University - Pratt School of Engineering 311 Research Drive, Durham, NC 27710

Abstract:

Thalamus is the central hub for forebrain communication, a function mediated by approximately 30 nuclei. To uncover organizational principles of the thalamic pathways providing input to the forebrain, we produced a near-comprehensive transcriptomic atlas of its major projection classes. We found that almost all nuclei belong to one of three major profiles that lie on a single axis of variance aligned with the mediolateral axis of thalamus. This axis of variance is strongly enriched in genes encoding receptors and ion channels, and we show that each profile exhibits different electrophysiological signatures. Single-cell profiling revealed even further heterogeneity within established nuclear boundaries, suggesting that the same input to a given nucleus might be differentially processed. Together, our analysis shows striking covariation in the organization of thalamic pathways serving all input modalities and output targets, establishing a simple and comprehensive thalamic functional architecture.

One Sentence Summary: The diverse range of information filtered through thalamus passes through three major classes of pathway, organized along a spectrum, and distinguished by functionally relevant genes.

Main Text:

A fundamental goal in neuroscience is to uncover common principles by which different modalities of information are processed. In the mammalian brain, thalamus is the central processing station for diverse modalities of information en route to the forebrain(*1*, *2*). For example, thalamus handles inputs from sensory systems, subcortical motor systems, and cortical areas, with several thalamic nuclei providing input to a given cortical area(*3-5*). Whether there is a conserved architecture across the set of thalamic pathways projecting to each cortical area has remained unresolved.

To understand the organization and diversity of these thalamic pathways, we produced a near-comprehensive transcriptomic atlas of murine thalamus by microdissecting nuclei and pooling cells retrogradely labeled from individual forebrain areas (8 projection targets, 22 nuclei, 120 samples, Tables S1 and S2). Anterograde tracing of inputs to thalamus was used when identification of nuclear boundaries was ambiguous (Fig. 1, A and B, Table S1). We used hierarchical clustering to explore the relationship between the transcriptomes of thalamic nuclei (on the 500 most differentially expressed genes via an ANOVA-like test, see methods, Fig. S1B and Data S2), and identified five major subdivisions of nuclei across thalamus (Fig. 1C). Anterior dorsal nucleus (AD) and nucleus reuniens (RE) each formed profiles of their own, leaving three major multi-nuclei profiles. These major profiles were not explained by cortical projection target or modality, since the multiple nuclei subserving motor, somatosensory or visual cortices split across different profiles. For example, central medial (CM), ventral anterior (VA) and ventral lateral (VL) nuclei all project to motor cortex, but are split across the three profiles. Rather, each of the three major profiles occupied a characteristic position along the mediolateral axis of the thalamus (Fig. 1D). We thus find that the architecture of thalamus is dominated by genetic differences that are organized topographically. By typically receiving input from each of these profiles, each cortical region samples from all three genetically defined pathways.

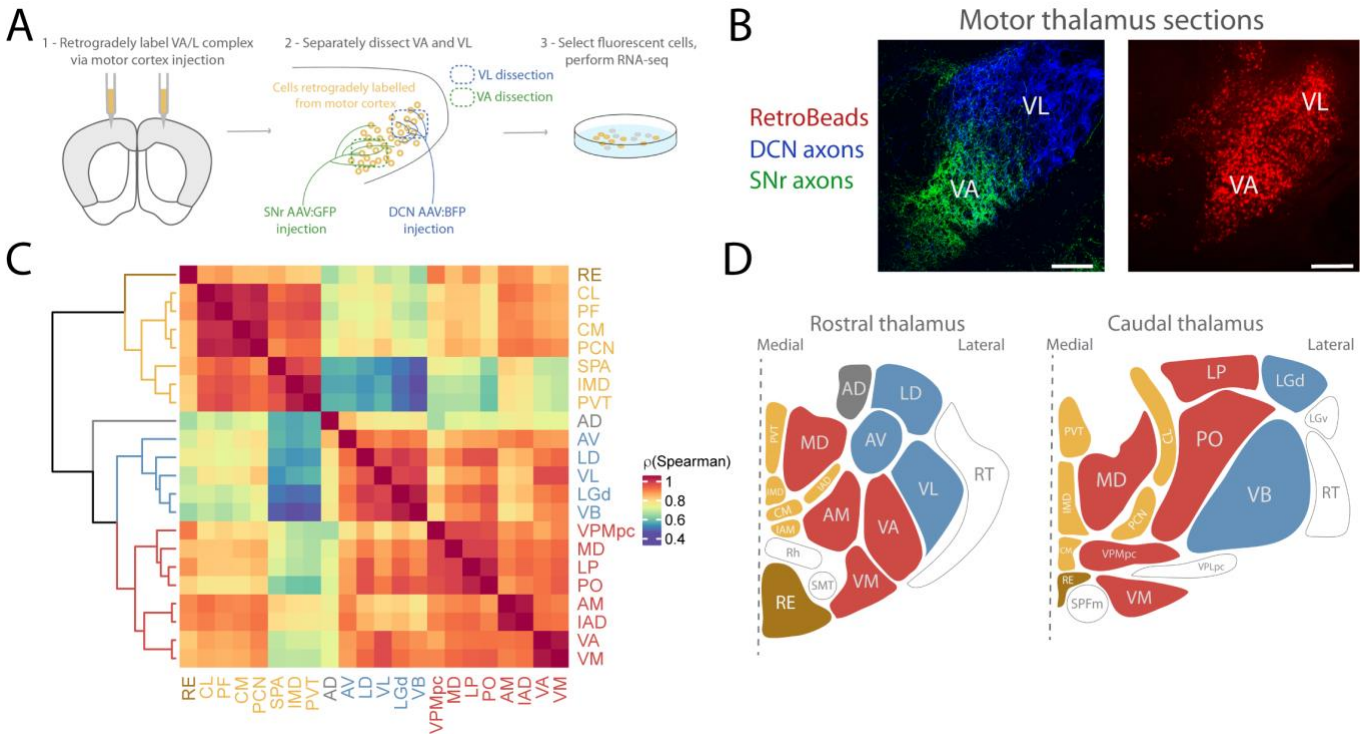

Fig. 1. A near-comprehensive transcriptomic atlas allows unbiased clustering of thalamic gene expression profiles

A. Schematic of experimental pipeline to obtain transcriptomic atlas of the thalamus. 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), respectively, were injected to the deep cerebellar nuclei (DCN) and substantia nigra pars reticulata (SNr) to label motor nuclear subdivisions (ventral lateral (VL) and ventral anterior (VA), respectively) previously identified(*6*, *7*)

B. Example labelling from scheme shown in A. Coronal sections. Scale bars = 200 µm.

C. Hierarchical clustering of thalamic nuclei using Spearman's correlation of top 500 most differentially expressed genes (Fig. S1B) across all 22 nuclei. Major clusters defined as the top 5 branches of cluster dendrogram.

D. Topographic localization of gene expression profiles in thalamus. Coronal thalamic section schematics showing the profile assignment of each nucleus. Thalamic nuclei colored as in Fig. 1C with unsampled nuclei left uncolored. Modified from the Allen Brain Atlas to show VA/L subdivision and to label somatosensory nuclei as VB. (see Fig. 1B).

To understand the pattern of gene expression differences between the thalamic projection profiles, we performed principal component analysis (PCA) on our data. Principal component 1 (PC1, 38% explained variance) separated nuclei into the same major profiles identified via hierarchical clustering (Fig. 2A, Fig. S2A). Again, position on this first dimension strongly correlated with mediolateral position, demonstrating topogenetic architecture in thalamus (Fig. 2B, Fig. S3B). Based on their relative order on this first component, we named the three major profiles primary, secondary, and tertiary. The 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 (Fig. 2C). This primary axis was prominently enriched in genes encoding neurotransmitter receptors, ion channels, and signaling molecules (Fig. 2D and E). Thus, the major differences in gene expression between thalamic nuclei is explained by a single axis closely related to mediolateral position.

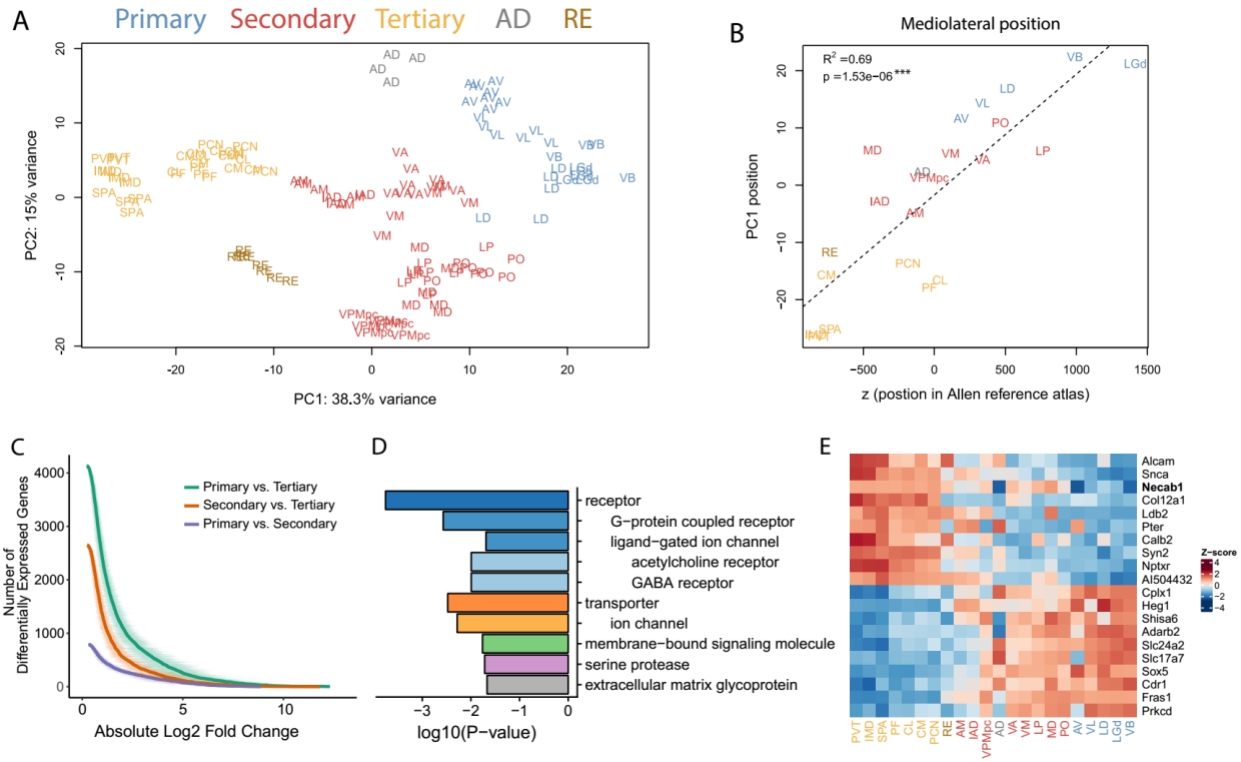

Fig. 2. A common topographic axis of variance separates major thalamic profiles

A. PCA showing 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.

B. PC1 position is highly correlated with mediolateral position of the nuclei. Mediolateral positions are z voxel coordinates of nuclei centers in the Common Coordinate Framework (CCF) of the Allen mouse brain reference atlas, where one voxel corresponds to 10 µm.

C. 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). Shaded bands show standard deviations of bootstrapped log2 fold change values.

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.

E. Heatmap of genes with strongest positive and negative loadings on PC1. Nuclei are ordered by their mean position on PC1 of Fig. 2A. Colors represent gene-wise Z-scores.

Prior work has shown substantial differences in electrophysiological properties between different thalamic nuclei, but to date this has not been incorporated into thalamic organizational schemes (*7*, *11*, *12*). Given the prominent differences in receptor and ion channel expression between thalamic profiles, we asked whether these profiles correspond to functionally distinct classes of neurons. We first performed PCA on the expression profiles of voltage-gated ion channel or neurotransmitter/modulator receptor encoding genes (Fig. 3A, left and right respectively). Analysis with these limited gene sets reproduced the separation of profiles in PC1 (Fig. 3A), confirming that ion channel and receptor profiles are organized along the same axis identified in Fig. 2. Genes linked to high firing rates via faster channel kinetics, such as Kv3 channels (*Kcnc1*, *Kcnc3*), the *Scn8a* channel, and the *Kcnab3* subunit(*8-10*), tended to be progressively elevated toward primary profile nuclei. This raised the possibility that action potential width may progressively narrow from tertiary to primary nuclei (Fig. 3B). Whole-cell recordings from the motor-related nuclei CM, VA and VL (representing the three main nuclear profiles; Fig. 2E) confirmed this prediction (Fig. 3C). Neurons recorded within VL have the narrowest action potential width and those in CM the widest. In addition, many other electrophysiological properties showed a systematic gradient ranging from VL through VA to CM (Fig. 3C, and Fig. S3). Therefore, a topographic organization of genetic profiles results in functional features changing across the mediolateral axis of the thalamus.

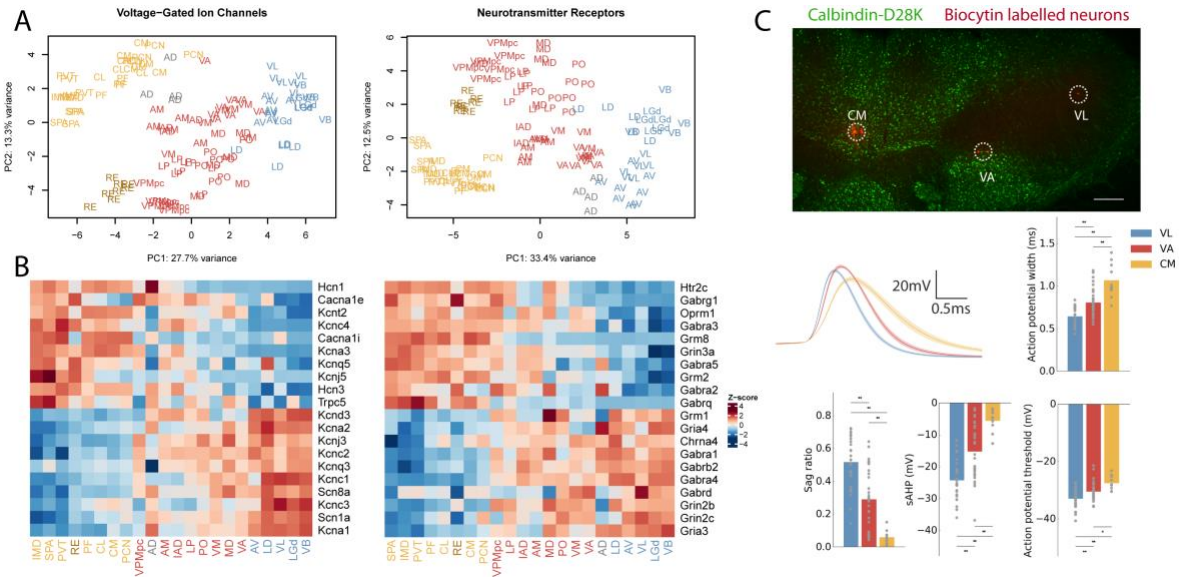

Fig. 3. Functionally relevant genes and electrophysiological properties vary systematically across nuclear profiles

A. PCA including only genes encoding voltage-gated ion channels and neurotransmitter/neuromodulator receptors. Colors as in Fig. 1C.
B. Heatmap for genes with the highest gene loadings in PC1 from Fig. 3A. Voltage-gated ion-channels on the left and neurotransmitter receptors on the right. Colors represent gene-wise Z-scores.
C. Systematic variation of electrophysiological properties across profiles. Whole-cell patch-clamp recordings from VL (primary thalamus), VA (secondary thalamus) and CM (tertiary thalamus). Top row: neurons were labelled with biocytin (red, highlighted with white circle) and localized to individual nuclei with the aid of Calbindin-D28K immunolabelling (green). Scale bar = 100µm. Middle row, left shows average action potential shape for VL, VA and CM neurons (mean±SEM). 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 Fig. S3 for further comparisons. Sample contained 29 VL neurons, 34 VA neurons and 10 CM neurons.

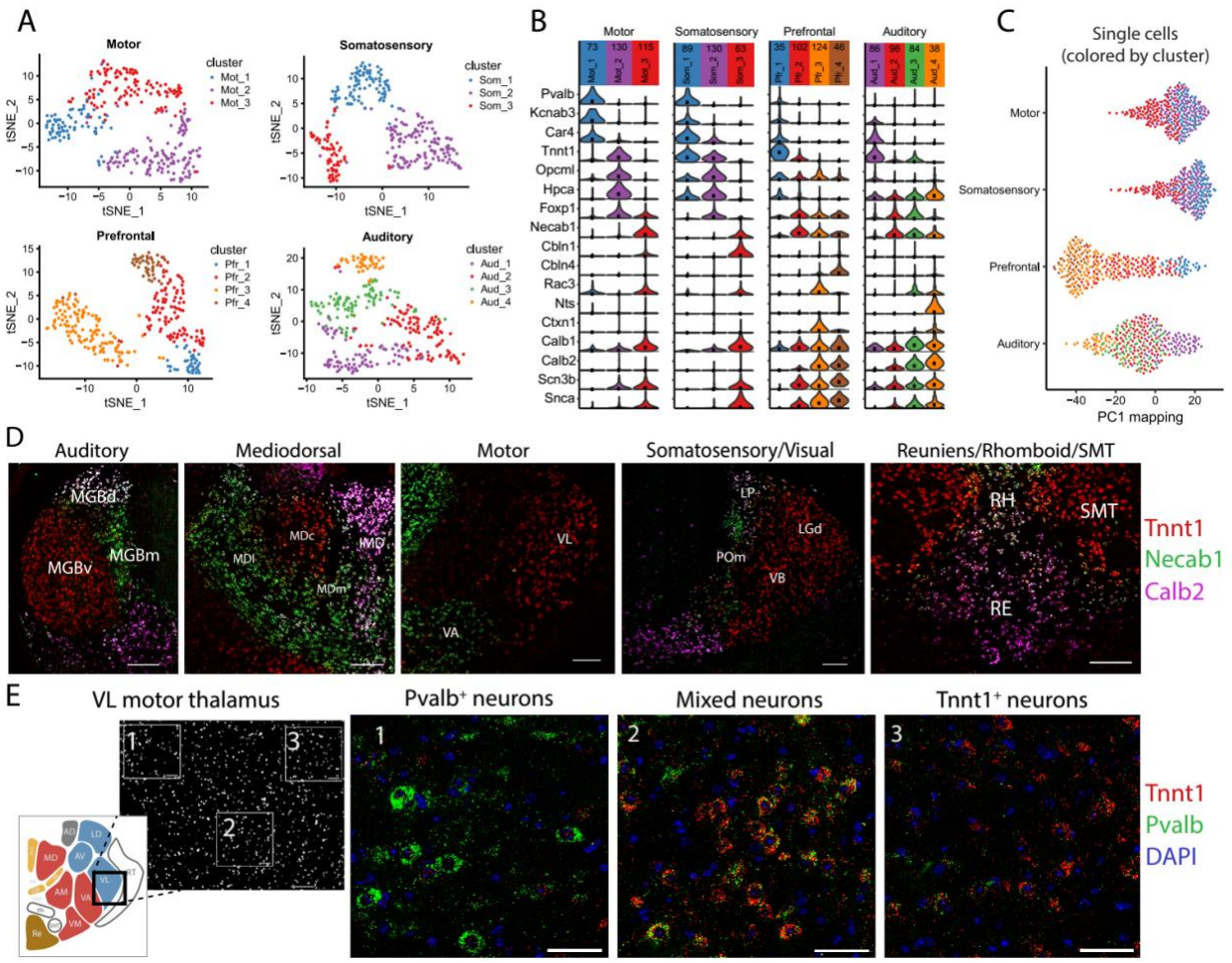

Fig. 4. A topographic spectrum of thalamic cell identities within and between thalamic nuclei

A. Overview of clusters (identified using the *Seurat* package, see methods) within each modality visualized via tSNE, cells colored by cluster identity.
B. Violin plots for marker genes for each cluster (inclusion criteria were Likelihood Ratio Test P-value < 10^{-5}, log2 fold change > 0.5, for clusters in each projection type).
C. Projection of single-cell data onto pooled-cell PC 1 from Fig. 2A, each dot is a single cell colored by the clusters from A and B. See Fig. S5B for separate plotting of each cluster.
D. Topographic distribution of marker genes within 6 major thalamic modalities. Multi-FISH with probes for *Calb2* (pink), *Tnnt1* (red) and *Necab1* (green). See Fig. S6 and S7 for expanded views and quantification. Scale bars = 200 µm.
E. Multi-FISH within VL thalamus comparing genes marking clusters in single-cell data. Left panel shows field of view (coronal section, scale bar = 100 µm). Right three panels

show (scale bars = 50µm) expansion of the three boxed areas, moving left to right. Middle box shows intermediate cells expressing both single cell cluster markers. (red = *Tnnt1*, green = *Pvalb*, blue = DAPI).

Thus far we analyzed thalamic nuclei by pooling projection neurons from specific anatomical positions. 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 (Table S3). Analysis of this single-cell RNAseq dataset resulted in multiple clusters for each projection class, and cluster markers included many genes that also distinguished nuclei (Fig. 4B). Single-cell clusters also separated along the first principal component derived from our pooled-cell RNAseq dataset (Fig. 4C, Fig. S5). Markers for the single-cell clusters were spatially separated at the single-cell level, but with intermediate cells especially prominent near the anatomical boundaries of thalamic nuclei (Fig. 4D, Fig. S6). This is consistent with spatially organized heterogeneity or a gradient-like organization rather than intermingled, discrete populations(*13*, *14*).

Given the strong relationship between PC1 and anatomical nuclei position (Fig. 2C and S2), the presence of single-cell clusters occupying similar PC1 positions (e.g. clusters 1 and 2 of the motor projections neurons, or clusters 1 and 2 of the somatosensory nuclei, Fig. S6) suggested that distinct neuron types could also coexist within anatomical boundaries of nuclei. We examined this possibility by performing multi-color fluorescent *in situ* hybridization (multi-FISH) for genes which distinguished amongst the major profiles (*e.g. Pvalb* and *Tnnt1*). Taking motor thalamus as an example, *Pvalb* and *Tnnt1* expressing cells were found within the anatomical boundaries of a single thalamic nucleus (VL; Fig. 4E). Some individual VL cells expressed both *Pvalb* and *Tnnt1; Pvalb*-selective, intermediate, and *Tnnt1*-selective cells were distributed along the mediolateral axis (Fig. 4E). Therefore, spatially organized transcriptomic differences can exist even within individual thalamic nuclei.

A common organizing principle of thalamus divides nuclei into discrete core or matrix subtypes that span modalities(*1*, *15*). To date the strongest evidence for cross-modal organization is that genetic differences can be larger within a sensory modality than between sensory modalities(*16*)(Fig. S8). However, previous studies have focused either on select sensory systems and/or have had limited scope with respect to the molecules investigated(*17*). Here, using the full transcriptomes of nearly all thalamic pathways, we confirm cross-modal organization but replace the core/matrix dichotomy with a spectrum of profiles that span a single axis of genetic variance. For example, we find that the matrix subtype has substantial diversity, splitting into multiple profiles based upon hierarchical clustering. The axis of variance is dominated by genes that directly shape neuronal properties, implying conserved, systematic variation of function. This pattern of variation is not only imposed on sensory thalamocortical systems but also diversifies motor, limbic, and cognitive thalamocortical systems. Understanding how the pattern of intra-thalamic molecular variability intersects with input modalities and behavioral relevance will be an important challenge for future work(*18-22*).

Our single-cell transcriptomics indicated that a given cortical area samples across a spectrum of thalamic profiles. Suggestive of even finer gradation along the spectrum, some projection

neurons exhibited features of more than one profile. Using multi-FISH, we mapped these intermediate cell types both to the boundaries between nuclei and within nuclei. Intermediate cells might exhibit hybrid input-output transforms, morphologies, and/or functions. Since neurons spanning the full range of across-thalamus and within-nucleus profiles provide input to nearly all cortical areas, our data shows that thalamus provides each recipient forebrain area with a broad complement of differentially filtered inputs.

References and Notes:

1. E. Jones, The Thalamus. *Cambridge University Press* (2007).

2. S. M. Sherman, Thalamus plays a central role in ongoing cortical functioning. *Nature Neuroscience*. **19**, 533–541 (2016).

3. S. M. Sherman, R. W. Guillery, 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).

4. Y. Smith *et al.*, The thalamostriatal system in normal and diseased states. *Front Syst Neurosci*. **8**, 5 (2014).

5. F. Clascá, P. Rubio-Garrido, D. Jabaudon, Unveiling the diversity of thalamocortical neuron subtypes. *European Journal of Neuroscience*. **35**, 1524–1532 (2012).

6. E. Kuramoto *et al.*, Complementary distribution of glutamatergic cerebellar and GABAergic basal ganglia afferents to the rat motor thalamic nuclei. *European Journal of Neuroscience*. **33**, 95–109 (2010).

7. K. C. Nakamura, A. Sharott, P. J. Magill, Temporal Coupling with Cortex Distinguishes Spontaneous Neuronal Activities in Identified Basal Ganglia-Recipient and Cerebellar-Recipient Zones of the Motor Thalamus. *Cerebral Cortex*. **24**, 81–97 (2012).

8. S. A. Freeman, A. Desmazières, D. Fricker, C. Lubetzki, N. Sol-Foulon, Mechanisms of sodium channel clustering and its influence on axonal impulse conduction. *Cellular and Molecular Life Sciences*. **73**, 723–735 (2015).

9. T. Leicher, R. Bähring, D. Isbrandt, O. Pongs, Coexpression of the KCNA3B gene product with Kv1.5 leads to a novel A-type potassium channel. *J. Biol. Chem.* **273**, 35095–35101 (1998).

10. B. Rudy, C. J. McBain, Kv3 channels: voltage-gated K+ channels designed for high-frequency repetitive firing. *Trends in Neurosciences*. **24**, 517–526 (2001).

11. P. M. Fogerson, J. R. Huguenard, Tapping the Brakes: Cellular and Synaptic Mechanisms that Regulate Thalamic Oscillations. *Neuron*. **92**, 687–704 (2016).

12. E. Puil, H. Meiri, Y. Yarom, Resonant behavior and frequency preferences of thalamic neurons. *Journal of Neurophysiology*. **71**, 575–582 (1994).

13. M. S. Cembrowski *et al.*, Spatial Gene-Expression Gradients Underlie Prominent Heterogeneity of CA1 Pyramidal Neurons. *Neuron*. **89**, 351–368 (2016).

14. O. Gokce *et al.*, Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. *CellReports*. **16**, 1126–1137 (2016).

15. E. G. Jones, The thalamic matrix and thalamocortical synchrony. *Trends in Neurosciences*. **24**, 595–601 (2001).

16. L. Frangeul *et al.*, A cross-modal genetic framework for the development and plasticity of sensory pathways. *Nature*. **538**, 96–98 (2016).

17. P. Rubio-Garrido, F. Pérez-de-Manzo, F. Clascá, Calcium-binding proteins as markers of layer-I projecting vs. deep layer-projecting thalamocortical neurons: A double-labeling analysis in the rat. *Neuroscience*. **149**, 242–250 (2007).

18. Z. V. Guo *et al.*, Maintenance of persistent activity in a frontal thalamocortical loop. *Nature*. **545**, 181–186 (2017).

19. L. I. Schmitt *et al.*, Thalamic amplification of cortical connectivity sustains attentional control. *Nature*. **545**, 219–223 (2017).

20. S. S. Bolkan *et al.*, Thalamic projections sustain prefrontal activity during working memory maintenance. *Nature Neuroscience*. **20**, 987–996 (2017).

21. S. Lee *et al.*, Bidirectional modulation of fear extinction by mediodorsal thalamic firing in mice. *Nature Publishing Group*. **15**, 308–314 (2011).

22. R. R. Llinas, Bursting of Thalamic Neurons and States of Vigilance. *Journal of Neurophysiology*. **95**, 3297–3308 (2006).

23. D. G. R. Tervo *et al.*, A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. *Neuron*. **92**, 372–382 (2016).

24. G. Paxinos, K. Franklin, The Mouse Brain in Stereotaxic Coordinates, 4th Edition. *Academic Press* (2012).

25. T. Kita, N. Shigematsu, H. Kita, Intralaminar and tectal projections to the subthalamus in the rat. *European Journal of Neuroscience*. **44**, 2899–2908 (2016).

26. C. M. Hempel, K. Sugino, S. B. Nelson, A manual method for the purification of fluorescently labeled neurons from the mammalian brain. *Nat Protoc*. **2**, 2924–2929 (2007).

27. S. Picelli *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in single cells.

Nat Meth. **10**, 1096–1098 (2013).

28. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2012).

29. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**, 139–140 (2009).

30. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 31–21 (2014).

31. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

32. A. Butler, R. Satija, Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv*, 1–18 (2017).

33. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. **33**, 495–502 (2015).

Acknowledgements:

Funding: This project was funded as a small project team (ThalamoSeq) by HHMI at the Janelia Research Campus, following a pilot project in the Dudman/Hantman labs. SN and CL were also supported by grants from NINDS (NS079419) and NIMH (MH105949). AS is funded via the Janelia Visiting Scientists Program. **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. LW: Collected data. BS: Collected data. WK: Supervised project, AL: Collected data and developed methods JD: Supervised project and wrote the paper SN: Supervised project. AH: Initiated and supervised project, wrote the paper. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** All transcriptomic data used will be made publicly available via the Gene Expression Omnibus. **Further acknowledgements:** We thank Karel Svoboda, Albert Lee and Amy Chuong for critical input throughout the project. We thank Matthew Phillips, Kirandeep Ghataorhe, Brett Mensh and Yves Weissenberger for comments on the manuscript. We thank Vilas Menon, Damian Kao and Mark Cembrowski for help with single-cell RNAseq analysis. We thank Kim Ritola and the Janelia Virology and histology cores for production of viruses and histology. 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.

Supplementary Materials:

Materials and Methods

EXPERIMENTAL METHODS

Animal care

Experimental procedures were approved by the Institutional Animal 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 (Data S1,S3).

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 interest. For viral injections, rAAV2-retro expressing cre-dependent GFP or TdTomato under the CAG promoter were injected, with volumes of 50-100nL at 3 depths (see Table S2) (*23*). Minimum survival time was 3 weeks post-injection. Viruses were prepared by Janelia Virus Services. Non-viral 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 (diluted 3x in PBS, 50-200nL per site). Anterograde labelling of inputs to thalamus was also used in a small number of cases (see Table S1,S2).

We referred to the Paxinos and Franklin mouse brain atlas to guide our dissections(*24*). For the majority of regions of thalamus, retrograde tracers labeled populations corresponding to identified thalamic nuclei (Table S1 and S2 for targeting details). However, the caudal intralaminar nuclei (parafascicular complex) were less clearly delineated. This likely reflects additional heterogeneity within this complex beyond that shown in atlases(*25*).

At no stage were experimenters blinded to sample identity.

Manual cell sorting and RNAseq

Sorted pooled-cell RNAseq

Fluorescent cells were collected and sequenced as previously described(13, 26). Briefly, animals were deeply anaesthetized with isoflurane and euthanized. Coronal slices (200-300 μm) were cut and placed for 1 hour at room temperature with pronases and neural activity blockers in artificial cerebrospinal fluid (ACSF). Relevant regions were then microdissected, and the tissue dissociated. The resulting cell suspensions were diluted with filtered ACSF and washed at least 3 times by transferring them to clean dishes. This process produces negligible contamination with

non-fluorescent tissue (Fig. S1A)(*26*). After the final wash, samples were aspirated in a small volume (3 µl) and lysed in 47µl XB lysis buffer using the Picopure kit (KIT0204, ThermoFisher) in a 200µl PCR tube (Axygen), incubated for 30 min at 40 ºC on a thermal cycler and stored at -80 ºC.

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). 1µL ERCC RNA spike-in mix (Life technologies) was added to each sample. Amplification was then performed using the Ovation RNA-seq v2 kit (NuGEN), yielding 4-8 µg of cDNA. The Ovation rapid DR multiplexing kit (NuGEN) was used to make libraries for sequencing, which were sequenced on a HiSeq 2500 (Illumina).

Sorted single-cell RNAseq

Retrogradely labeled cells were isolated as described above, and collected into 8-well strips containing 3 µL Smart-seq2 lysis buffer, flash-frozen on dry ice, and stored at -80ºC until further use(*27*).

Upon thawing, cells were re-digested with Proteinase K and barcoded RT primers were added. cDNA synthesis was done using the Maxima H Minus RT kit (Thermo Fisher) 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).

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(*30*). The libraries were then purified following the Nextera XT protocol (at 0.6X dilution) and quantified by qPCR using Kapa Library Quantification (Kapa Biosystems). 6-10 plates were run twice on a NextSeq 550 flow cell. Read 1 contained the spacer, cell barcode, and unique molecular identifier (UMI). Read 2 was a cDNA fragment from the 3' end of the transcript.

Multi-FISH

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 4 ºC overnight, washed 3 times with PBS, and cryoprotected in a sucrose series of 10%, 20%

then 30% in PBS at 4 ºC. All solutions were prepared RNase-free. Brains were sectioned (14µm) on a Leica CM3050S cryostat, mounted onto Fisher SuperFrost Plus slides, and stored at -80 ºC .

Multi-FISH was performed using the RNAscope Multiplex Fluorescent Assay platform from ACDBio, following the manufacturer's protocol. The probes used were: *Calb2* (ref 313641-C3), *Necab1* (ref 428541 and 428541-C2), and *Tnnt1* (ref 466911-C2). Fluorescent dyes were DAPI, Alexa Fluor 488, Atto 550 and Atto 647. Images were acquired using a Zeiss LSM 880 confocal microscope. Images were acquired with an air 20x (0.8NA) objective unless otherwise specified.

Electrophysiology

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 µm 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 hour 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 CM, VA and VL 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 MΩ) 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 16th-19th action potentials in trials with 20 to 40 Hz firing rate were averaged.

Immunohistochemistry

After recordings, slices were fixed with 4% paraformaldehyde and 2.5% sucrose in 5x phosphate-buffered saline (PBS) at 4 °C for 2-10 days. After washing 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

Pooled-cell 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(*28*). 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 (Fig. S1A). 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 (Fig. S1C).

Differential gene expression

Differential expression was assessed using the Bioconductor package *edgeR*(*29*). Low counts were removed by requiring a Transcripts per million (TPM) > 5 in at least 3 samples. This yielded a list of approximately 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 Likelihood 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 (ANODEV test for generalized linear models, as described in edgeR User manual 3.2.6), testing for differences between any of the 22 nuclei, and used the 500 genes with the highest P-value. To avoid bias due to differences in sample number when comparing numbers of differentially expressed genes between different profiles in Fig. 2C, the groups were subsampled (with replacement) to the size of the smallest group. Bootstrapped log2 fold changes were obtained over 100 iterations.

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)(*30*, *31*) .

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. We termed these profiles, not clusters, as we do not mean to imply discreteness between the classifications. PCA was done using the singular value decomposition based *prcomp* function in R. For functional enrichment of differentially expressed genes, we used the PANTHER 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 *Drop-seq_tools-1.12*.

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 (Fig. S4A). Of these, 9 cells were found to be significantly contaminated with oligodendrocyte transcripts (Fig. S4C), 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 pooled-cell dissections were not precisely matched, for example motor-projecting midline nuclei were not dissected for single-cell. However, pooled-cell and single-cell RNAseq are in close agreement, indicating that our results are robust to collection method (Fig. S4B).

Single-cell clustering and marker genes

Single-cell clusters were defined using the *Seurat* R package (version 2.0) (*32*, *33*). Data were log transformed and scaled. For identifying variable genes, the FindVariableGenes function was used with default parameters (x.low.cutoff=0.1, x.high.cutoff=8, y.cutoff=1, n.bin=20). Single-cell clustering was performed separately for each modality using shared nearest neighbor (SNN) clustering approach in Seurat, limiting the analysis to the top 10 principal components for calculating Euclidian distances. Clustering resolution was set to 1. 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. Projection of single-cell data onto pooled-cell principal components was obtained by multiplying (inner product) log-transformed and scaled single-cell data by the pooled-cell principal component loadings.

Supplementary Figures

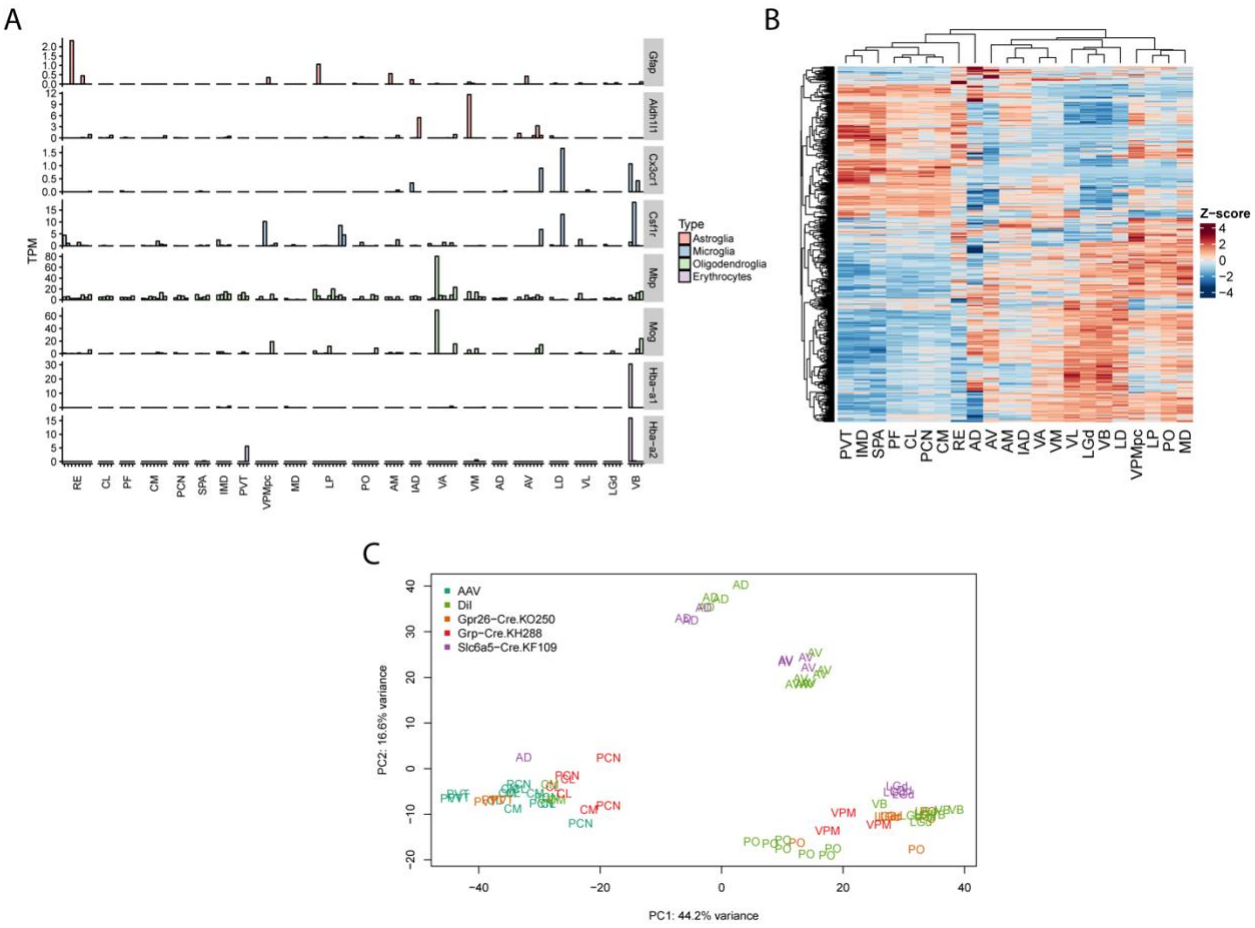

Fig. S1 – Pooled-cell RNAseq quality controls and additional analyses

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

B) Heatmap of the top 500 differentially expressed genes. Rows and columns are ordered by hierarchical clustering with Euclidean distance metric. 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.

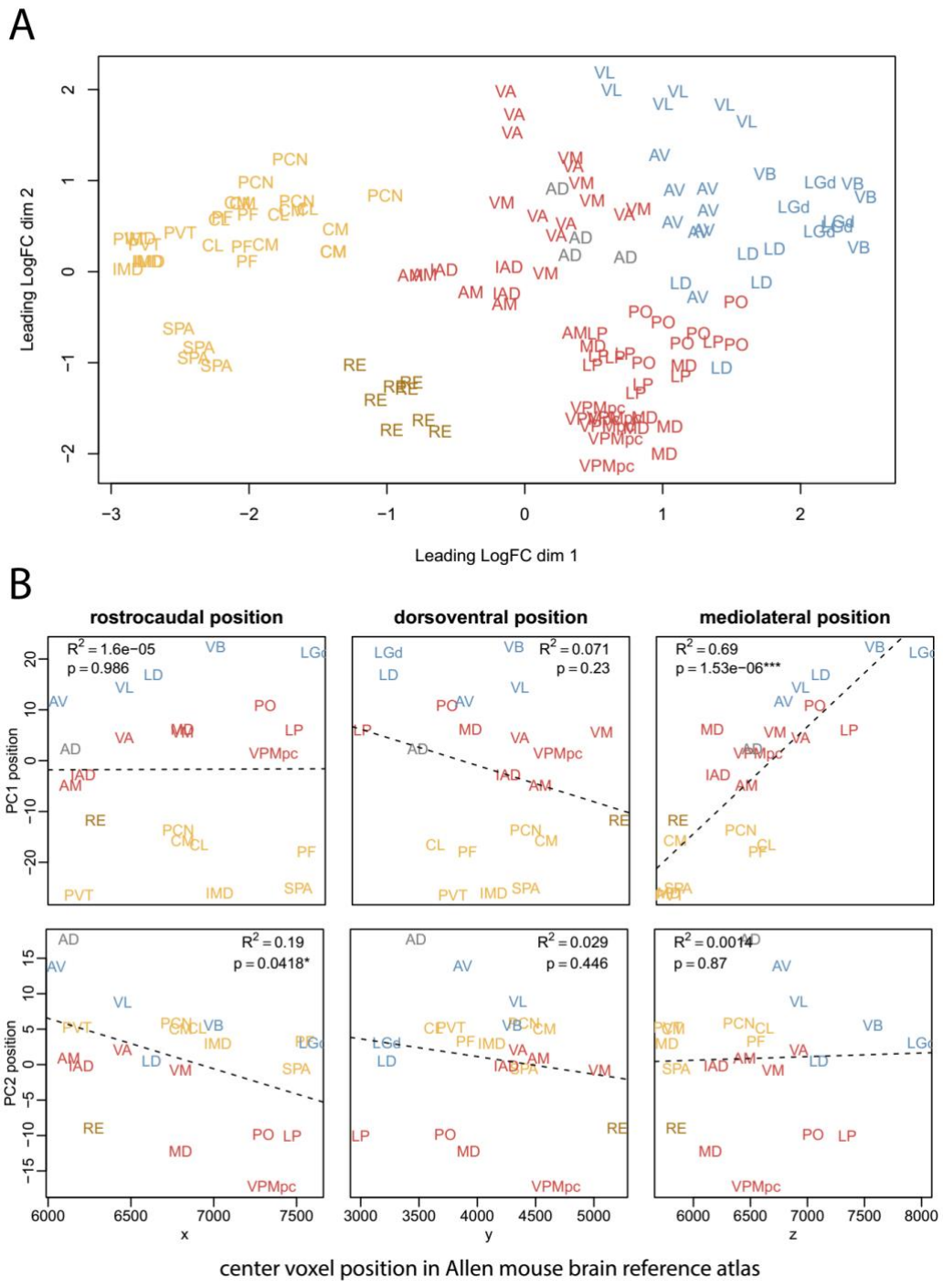

Fig. S2 – Additional analysis of topogenetic axis

A) 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).

B) Relationship of PC1 and 2 with topographical position of nuclei. Rostrocaudal, dorsoventral, and mediolateral positions are the x, y, and z voxel coordinates, respectively, in the Allen Mouse Brain reference atlas. One voxel corresponds to 10µm.

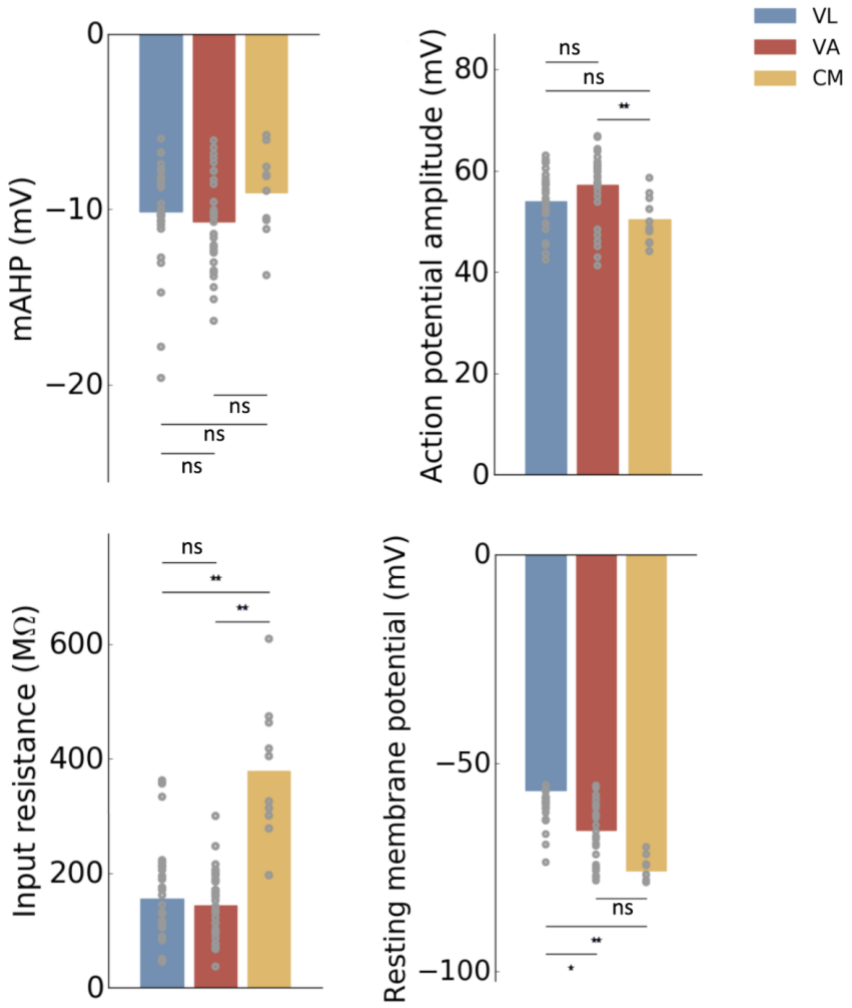

Fig. S3 - Additional electrophysiological properties between thalamic nuclear profiles. All statistical tests and experimental details are the same as in figure 3C.

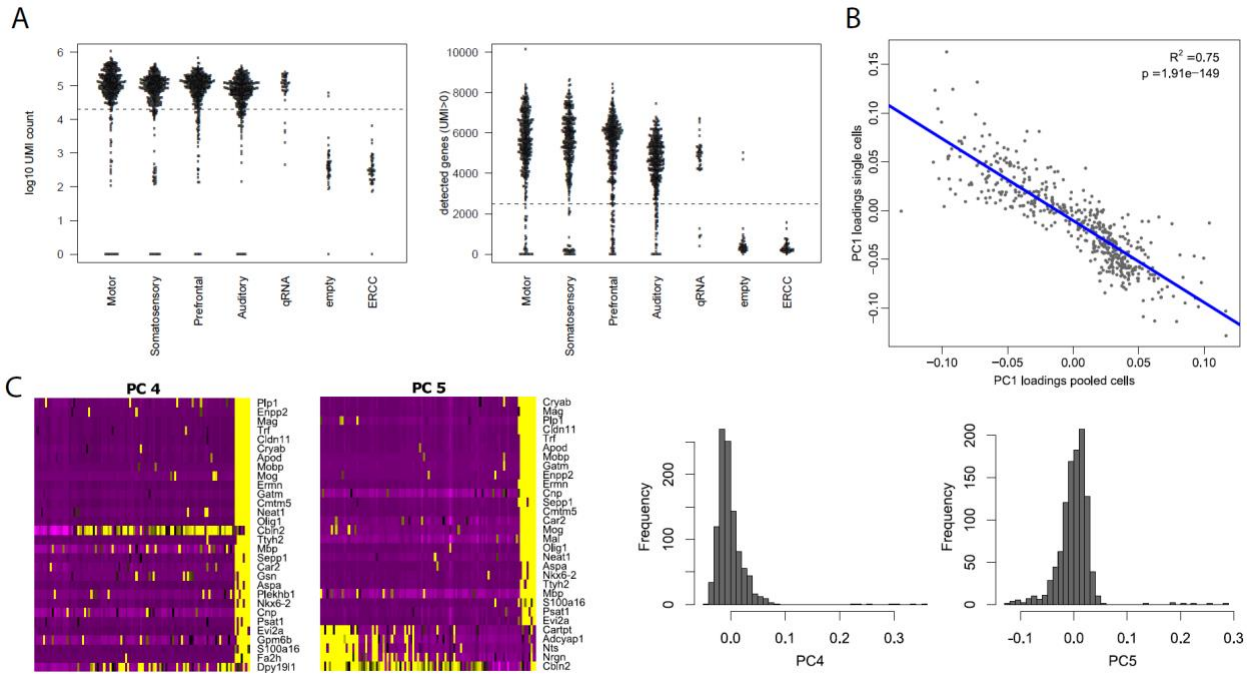

Fig. S4 – Quality control for single-cell RNAseq data

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 between nuclei (gene set as in Fig.1) are highly similar in pooled-cell and single-cell RNAseq data.

C) PCA on the single-cell RNAseq data revealed that principal components 4 and 5 represented non-neuronal contamination from oligodendrocytes. For all downstream analysis, the 9 cells whose principal component 4 and 5 positions were above 0.1 were removed.

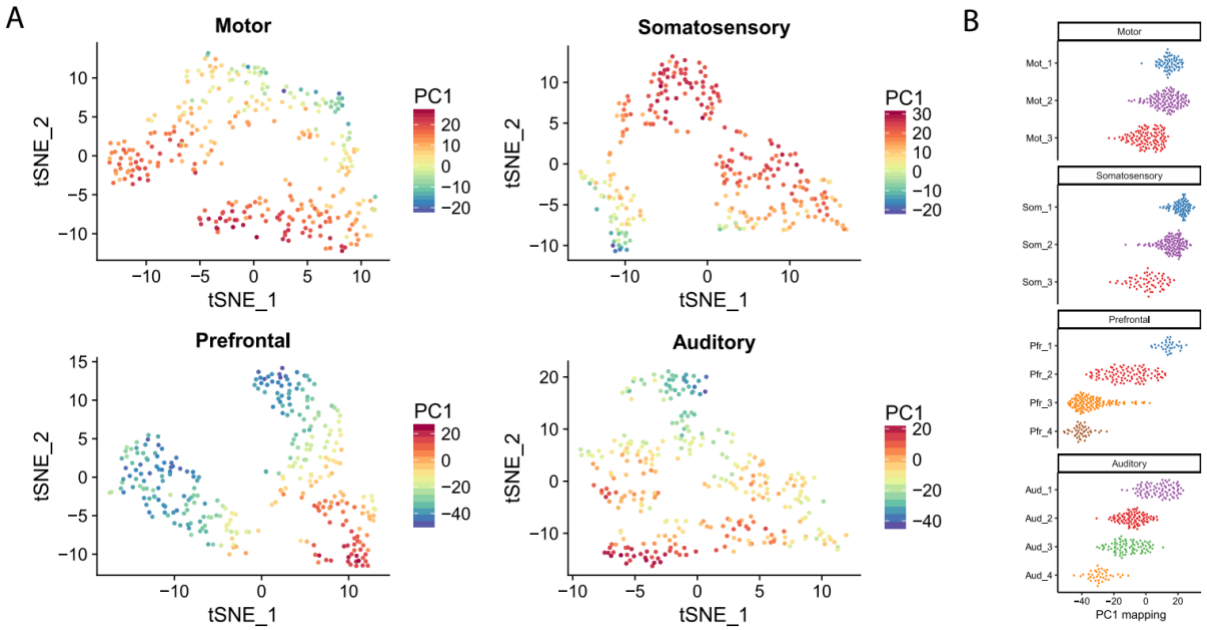

Fig. S5 – Projection of single cells onto PC1 of pooled-cell RNAseq data

A) tSNE plots for each projection type with cells from single-cell RNAseq colored by its projection onto pooled-cell PC1. Cells from the negative extreme of PC1 are absent from motor and somatosensory modalities, which may reflect a lack of sampling from the more medial thalamic sections.

B) Positions of single cells projected onto pooled-cell PC1 from Fig. 2, plotted separately for each single-cell cluster.

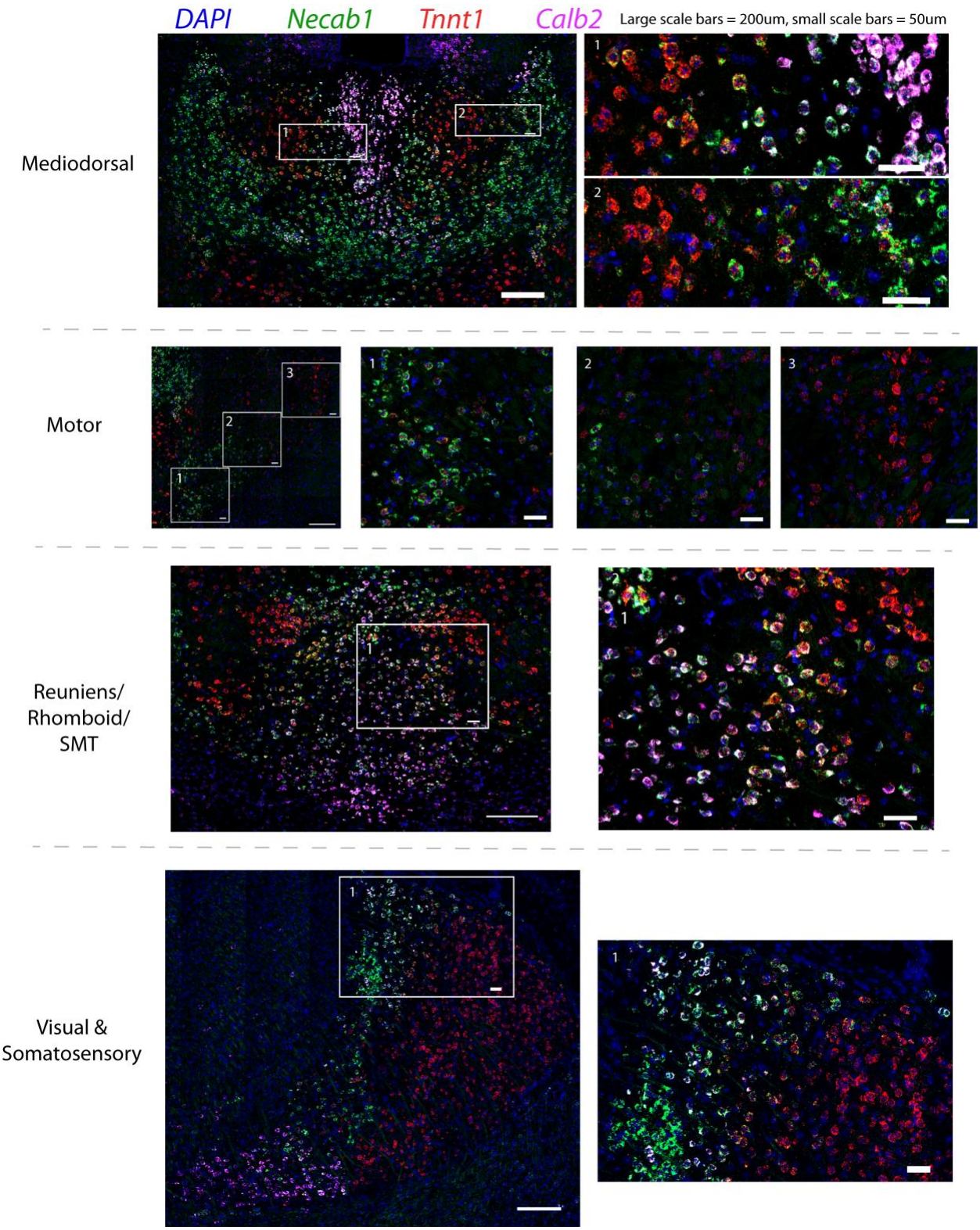

Fig. S6 – Multi-FISH show cells with mixed expression of profile marker genes
Expanded views of example regions showing intermediate cells expressing combinations of *Tnnt1*, *Necab1* and *Calb2*, which are preferentially expressed in primary, secondary and tertiary nuclear profiles respectively.

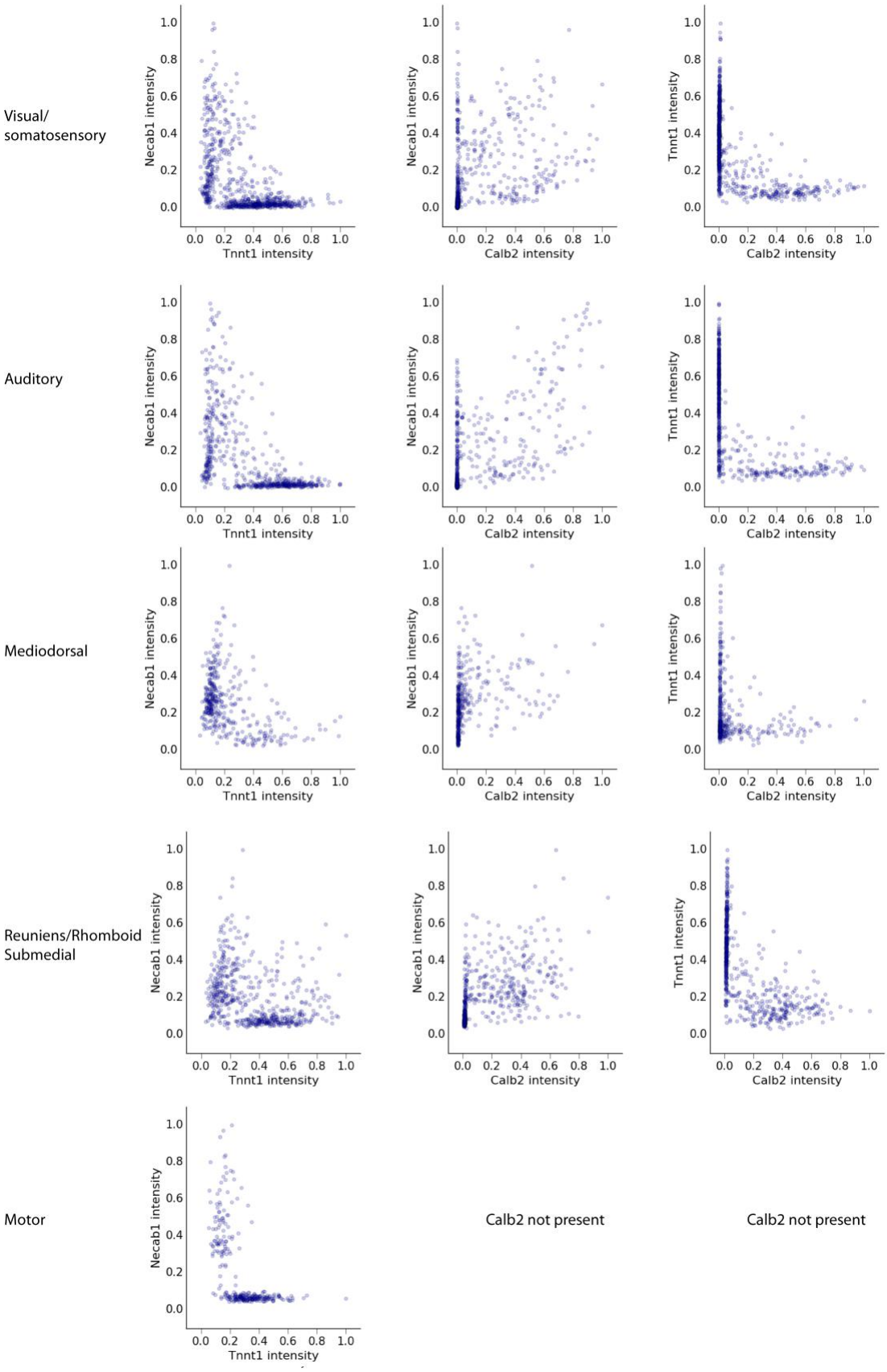

Fig. S7 - Quantification of multi-FISH images shows intermediate cells

Quantification of multi-FISH gene expression images. Regions of interest (ROIs) were drawn in ImageJ, and then processed in custom Python scripts. Intensity was normalized first to the ROI size, then divided by the maximum for that channel. Only cells that express at least one of the marker genes were included.

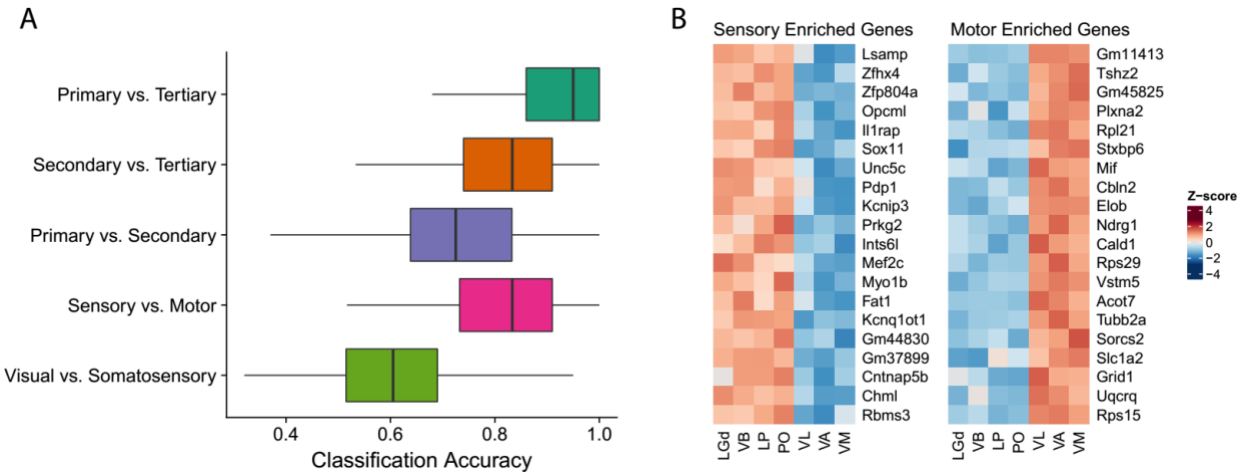

Fig. S2 – Differential gene expression between sensory and motor thalamic nuclei

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, PO 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

Table S1 – Sample collection approaches (pooled-cell RNASeq)

[Validation] refers to samples in Fig. S1C. For transgenic lines used, see supplementary Table S4.

Region name:	Labelling methods:	Replicates:
Anterior dorsal (AD)	Retrograde DiI from retrosplenial cortex. [Validation: Slc6a5-Cre.KF109 transgenic line]	4,[3]
Anterior medial (AM)	Retrograde DiI from cingulate cortex. Retrograde AAV from cingulate cortex.	4,1
Anterior ventral (AV)	Retrograde DiI from retrosplenial cortex. [Validation: Slc6a5-Cre.KF109 transgenic line]	8,[4]
Central lateral (CL)	rAAV2-retro from striatum. [Validation: Grp-Cre.KH288 transgenic line]	4
Central medial (CM)	rAAV2-retro from striatum. Retrograde DiI to striatum. [Validation: Grp-Cre.KH288 transgenic line]	4,3,[1]
Interanterodorsal nucleus (IAD)	Retrograde DiI from cingulate cortex	3
Intermediodorsal nucleus (IMD)	rAAV2-retro from striatum.	4
Laterodorsal nucleus (LD)	Retrograde DiI from retrosplenial cortex.	5
Lateral Geniculate (dorsal) nucleus (LGd)	Retrograde DiI from visual cortex with anterograde CTB-488 tracer injected to retina (Cholera Toxin Subunit B, Alexa Fluor 488™, Molecular Probes). [Validation: Slc6a5-Cre.KF109 and Gpr26-Cre.KO250 transgenic lines]	5,[3],[4]

Lateral posterior nucleus (LP)	Retrograde DiI from visual cortex with anterograde CTB-488 tracer injected to retina (Cholera Toxin Subunit B, Alexa Fluor 488™, Molecular Probes).	9
Medial geniculate body (MGB)	Not collected for pooled-cell RNAseq in this study. Consists of ventral (MGBv), dorsal (MGBd) and medial (MGBm) subdivisions.	na
Mediodorsal nucleus (MD)	Retrograde DiI from frontal cortex. Collected without subdividing the three subdivisions (central: MDc, medial: MDm, lateral: MDl).	6
Paracentral nucleus (PCN)	rAAV2-retro from striatum. [Validation: Grp-Cre.KH288 transgenic line]	4,[3]
Parafascicular nucleus (PF)	rAAV2-retro from striatum. This did not label the entire PF complex, and likely reflects a genetic subset.	4
Posterior medial nucleus (PO)	Retrograde DiI from somatosensory cortex. [Validation: Gpr26-Cre.KO250 transgenic line]	7,[4]
Paraventricular thalamus (PVT)	rAAV2-retro from striatum [Validation: Gpr26-Cre.KO250 transgenic line]	3,[3]
Reuniens nucleus (RE)	Retrograde DiI from cingulate cortex	8
Rhomboid nucleus (RH)	Not collected for pooled-cell sequencing in this study.	na
Subparafascicular nucleus (SPA)	rAAV2-retro from striatum. This did not label the entire parafascicular complex, and likely reflects a genetic subset (medially biased).	4
Submedial thalamus (SMT)	Not collected for pooled-cell RNAseq in this study.	na
Ventral anterior nucleus (VA)	Retrograde DiI from motor cortex. Anterograde labelling of inputs from SNr and DCN with viral tracers (AAV2/1-CAG-GFP and/or AAV2/1-CAG-BFP).	8
Ventral medial nucleus (VM)	Retrograde DiI from motor cortex. Anterograde labelling of inputs from SNr and DCN with viral tracers (AAV2/1-CAG-GFP and/or AAV2/1-	6

	CAG-BFP).	
Ventral lateral nucleus (VL)	Retrograde DiI from motor cortex. Anterograde labelling of inputs from from SNr and DCN with viral tracers (AAV2/1-CAG-GFP and/or AAV2/1-CAG-BFP).	6
Ventrobasal nucleus (VB)	Retrograde DiI from somatosensory cortex [Validation: Grp-Cre.KH288 transgenic line for VPM]	4,[3]
Ventroposteromedial parvocellular (VPMpc)	Retrograde DiI from insular cortex.	6

Table S2 – Coordinates for retrograde labelling/trace injections (pooled-cell RNAseq)

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

Sample area:	**Injection coordinates (in millimeters, from bregma, depth from brain surface)**	***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 (CTB) 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 (CTB) 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	
VM	0.3 rostral, 1.6 lateral	

	0.8 rostral, 1.6 lateral 0.5 rostral, 1.4 lateral	
VL	0.3 rostral, 1.6 lateral 0.8 rostral, 1.6 lateral 0.5 rostral, 1.4 lateral	
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.	

Table S3 – Injection coordinates for retrograde labelling (Single-cell RNAseq)

Sample area:	**Injection coordinates (in millimeters, from bregma, depth from brain surface)**
Auditory Thalamus	3.5 caudal, 4.5 lateral, 2.25 deep. 3.16 caudal, 4.5 lateral, 2.0 deep
Mediodorsal (Prefrontal) 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.

Table S4 – Transgenic mice used in this study:

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