Addressing the looming identity crisis in single cell RNA-seq

Megan Crow, Anirban Paul, Sara Ballouz, Z. Josh Huang, Jesse Gillis*

Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, NY 11724, USA

mcrow@cshl.edu, paula@cshl.edu, sballouz@cshl.edu, huangj@cshl.edu, jgillis@cshl.edu

*corresponding author

Abstract

Single cell RNA-sequencing technology (scRNA-seq) provides a new avenue to discover and characterize cell types, but the experiment-specific technical biases and analytic variability inherent to current pipelines may undermine the replicability of these studies. Meta-analysis of rapidly accumulating data is further hampered by the use of *ad hoc* naming conventions. Here we demonstrate our replication framework, MetaNeighbor, that allows researchers to quantify the degree to which cell types replicate across datasets, and to rapidly identify clusters with high similarity for further testing. We first measure the replicability of neuronal identity by comparing more than 13 thousand individual scRNA-seq transcriptomes, then assess cross-dataset evidence for novel pyramidal neuron and cortical interneuron subtypes identified by scRNA-seq. We find that 24/45 cortical interneuron subtypes and 10/48 pyramidal neuron subtypes have evidence of replication in at least one other study. Identifying these putative replicates allows us to re-analyze the data for differential expression and provide lists of robust candidate marker genes. Across tasks we find that large sets of variably expressed genes can identify replicable cell types and subtypes with high accuracy, indicating many of the transcriptional changes characterizing cell identity are pervasive and easily detected.

Introduction

Single cell RNA-sequencing (scRNA-seq) has emerged as an important new technology enabling the dissection of heterogeneous biological systems into ever more refined cellular components. One popular application of the technology has been to try to define novel cell subtypes within a given tissue or within an already refined cell class, as in the lung (Treutlein et al., 2014), pancreas (Baron et al., 2016; Muraro et al., 2016; Segerstolpe et al., 2016; Wang et al., 2016), retina (Macosko et al., 2015; Shekhar et al., 2016), or others (Grun et al., 2015; Klein et al., 2015; Min et al., 2015). Because they aim to discover completely new cell subtypes, the majority of this work relies on unsupervised clustering, with most studies using customized pipelines with many unconstrained parameters, particularly in their inclusion criteria and statistical models (Grun et al., 2015; Habib et al., 2016; Macosko et al., 2015; Zeisel et al., 2015). While there has been steady refinement of these techniques as the field has come to appreciate the biases inherent to current scRNA-seq methods, including prominent batch effects (Hicks et al., 2015), expression drop-outs (Lun et al., 2016; Pierson and Yau, 2015), and the complexities of normalization given differences in cell size or cell state (Buettner et al., 2015; Vallejos et al., 2015), the question remains: how well do novel transcriptomic cell subtypes replicate across studies?

In order to answer this, we turned to the issue of cell diversity in the brain, a prime target of scRNA-seq as neuron diversity is critical for construction of the intricate, exquisite circuits underlying brain function. The heterogeneity of brain tissue makes it particularly important that results be assessed for replicability, while its popularity as a target of study makes this goal particularly feasible. Because a primary aim of neuroscience has been to derive a taxonomy of cell types (Ascoli et al., 2008), already more than twenty single cell RNA-seq experiments have been performed using mouse nervous tissue (Poulin et al., 2016). Remarkable strides have been made to address fundamental questions about the diversity of cells in the nervous system,

48 including efforts to describe the cellular composition of the cortex and hippocampus (Tasic et
49 al., 2016; Zeisel et al., 2015), to exhaustively discover the subtypes of bipolar neurons in the
50 retina (Shekhar et al., 2016), and to characterize similarities between human and mouse
51 midbrain development (La Manno et al., 2016). In spite of this wealth of data, there have been
52 few attempts to compare, validate and substantiate cell type transcriptional profiles across
53 scRNA-seq datasets, and no systematic or formal method has been developed for
54 accomplishing this task.

55 To address this gap in the field, we propose a simple, supervised framework, MetaNeighbor
56 (**meta**-analysis via **neighbor** voting), to assess how well cell type-specific transcriptional profiles
57 replicate across datasets. Our basic rationale is that if a cell type has a biological identity rooted
58 in the transcriptome then knowing its expression features in one dataset will allow us to find
59 cells of the same type in another dataset. We make use of the cell type labels supplied by data
60 providers, and assess the correspondence of cell types across datasets by taking the following
61 approach (see schematic, Figure 1):

62 1) First we construct a kernel: we calculate correlations between all pairs of cells that we
63 aim to compare across datasets based on the expression pattern of a set of genes. This
64 generates a network where each cell is a node and the edges are the strength of the
65 correlations between them.
66 2) Next, we do cross-dataset validation: we hide all cell type labels ('identity') for one
67 dataset at a time. This dataset will be used as our test set. Cells from all other datasets
68 remain labeled, and are used as the training set.
69 3) Finally, we predict the cell type labels of the test set: we use a neighbor voting algorithm
70 to predict the identity of the held-out cells based on their similarity to the training data.

Conceptually, this resembles approaches for the validation of sample clustering (Dudoit et al., 2002; Kapp and Tibshirani, 2007) but it has been adapted to operate from within a supervised learning framework. This permits both systematic scoring and carefully defined control experiments to investigate the data features that drive high performance. Our implementation is extremely fast and robust to technical differences between experiments; because prediction is performed only within an individual dataset at a time, we are able to keep many aspects of technical variation constant. This essentially controls for any dataset specific effects that would otherwise swamp the subtler cell identity signal. The method provides a score that indicates the degree to which a cell type replicates for each gene set that is tested. This means that MetaNeighbor doubles as a low-tech 'feature selection tool' that we can use to identify the transcriptional features that are most discriminative between cell types. By comparing the scores returned from using Gene Ontology (GO) functions ("functional gene sets") or sets of randomly chosen genes ("random gene sets"), we can determine whether co-expression of specific gene sets is characteristic of particular cell types, and thus important for cell function or identity.

We evaluate cell identity by taking sequential steps according to the basic taxonomy of brain cells: first classifying neurons vs. non-neuronal cells across eight single cell RNA-seq studies, then classifying cortical inhibitory neurons vs. excitatory neurons, and for our final step, we align interneuron and pyramidal cell subtypes across three studies. Critically, we discover that that almost any sufficiently large and highly variable set of genes can be used to distinguish between cell types, suggesting that cell identity is widely represented within the transcriptome. Furthermore, we find that cross-dataset analysis of pyramidal neurons results in broad definition of cortical vs. hippocampal types, and find evidence for the replication of five layer-restricted subtypes. In contrast, we find that cortical interneuron subtypes show clear lineage-specific structure, and we readily identify 11 subtypes that replicate across datasets, including

96 Chandelier cells and five novel subtypes defined by transcriptional clustering in previous work.
97 Meta-analysis of differential expression across these highly replicable cortical interneuron
98 subtypes revealed evidence for canonical marker genes such as parvalbumin and somatostatin,
99 as well as new candidates which may be used for improved molecular genetic targeting, and to
100 understand the diverse phenotypes and functions of these cells.

101 **Assessing neuronal identity with MetaNeighbor**

102 We aimed to measure the replicability of cell identity across tasks of varying specificity.
103 Broadly, these are divided into tasks where we are recapitulating known cell identities, and ones
104 where are measuring the replicability of novel cell identities discovered in recent research. The
105 former class of task is the focus of this subsection: first, by assessing how well we could
106 distinguish neurons from non-neuronal cells ("task one"), and next assessing the discriminability
107 of excitatory and inhibitory neurons ("task two"). As detailed in the methods, MetaNeighbor
108 outputs a performance score for each gene set and task. This score is the mean area under the
109 receiver operator characteristic curve (AUROC) across all folds of cross-dataset validation, and
110 it can be interpreted as the probability that we will rank a positive higher than a negative (e.g.
111 neuron vs. non-neuronal cell, when using neurons as the positive label set) based on the
112 expression of a set of genes. This varies between 0 and 1, with 1 being perfect classification,
113 0.5 meaning that we have performed as well as if we had randomly guessed the cell's identity,
114 and 0.9 or above being extremely high. Comparison of scores across gene sets allows us to
115 discover their relative importance for defining cell identity.

116 As described above, in task one we assessed how well we could identify neurons and non-
117 neuronal cells across eight datasets with a total of 13928 cells (Table S1). Although this was
118 designed to be fairly simple, we were surprised to find that AUROC scores were significantly
119 higher than chance for all gene sets tested, including all randomly chosen sets ($AUROC_{all}$
120 $_{sets}$=0.78 ± 0.1, Figure 2A). Reassuringly, a bootstrapped sampling of the datasets showed a

121 trend toward increased performance with the inclusion of additional training data, indicating that
122 we are recognizing an aggregate signal across datasets (Figure S1). However, the significant
123 improvement of random sets over the null means that prior knowledge about gene function is
124 not required to differentiate between these cell classes. Randomly chosen sets of genes have
125 decidedly non-random expression patterns that enable discrimination between cell types.

126 Task two aimed to assess how well we could discriminate between cortical excitatory and
127 inhibitory neurons across four studies with a total of 2809 excitatory and 1162 inhibitory neurons
128 (Dueck et al., 2015; Habib et al., 2016; Tasic et al., 2016; Zeisel et al., 2015). Similar to our
129 previous results, we saw that AUROC scores were significantly higher than chance
130 (AUROC=0.69 ± 0.1, Figure 2B), suggesting that transcriptional differences are likely to be
131 encoded in a large number of genes.

132 Consistent with the view that a large fraction of transcripts are useful for determining cell
133 identity, we found a positive dependency of AUROC scores on gene set size, regardless of
134 whether genes within the sets were randomly selected or shared some biological function
135 (Figure 2B). This was further supported by a comparison of scores for task one using 100 sets
136 of either 100 or 800 randomly chosen genes. AUROC score distributions and means were
137 significantly different, with sets of 100 genes having lower scores but higher variability in
138 performance, whereas sets of 800 genes were more restricted in variance and gave higher
139 performance on average (Figure 2C, $AUROC_{100}$=0.80 ± 0.05, $AUROC_{800}$=0.90 ± 0.03, $p<2.2E-$
140 16, Wilcoxon rank sum test). The variability in performance observed while keeping set size
141 constant suggests that even in random sets, there are transcriptional features that contribute to
142 cell identity. We delved into this further by comparing AUROC scores across gene sets chosen
143 based on their mean expression as we have previously shown that this is a critical factor to
144 control for in evaluating single cell gene co-expression (Crow et al., 2016). We performed task

145 one again using expression-level based gene sets and found a strong positive relationship
146 between expression level and our ability to classify cells (Figure 2D, r_s=0.9).

147 These results provide evidence that MetaNeighbor can readily identify cells of the same type
148 across datasets, without relying on specific knowledge of marker genes. In these two examples,
149 all cells could be classified as one of two types, making this a binary classification task. We find
150 that a gene set's size and mean expression level are the key features that allow for cell type
151 discrimination in this setting.

152 **Investigating cortical interneuron subtypes using MetaNeighbor**

153 Cortical inhibitory interneurons have diverse characteristics based on their morphology,
154 connectivity, electrophysiology and developmental origins, and it has been an ongoing goal to
155 define cell subtypes based on these properties (Ascoli et al., 2008). In a related paper (Paul et
156 al., submitted), we describe the transcriptional profiles of GABAergic interneuron types which
157 were targeted using a combinatorial strategy including intersectional marker gene expression,
158 cell lineage, laminar distribution and birth timing, and have been extensively phenotyped both
159 electrophysiologically and morphologically (He et al., 2016). Previously, two studies were
160 published in which new interneuron subtypes were defined based on scRNA-seq transcriptional
161 profiles (Tasic et al., 2016; Zeisel et al., 2015). These found different numbers of subtypes (16
162 in one and 23 in the other), and the authors of the later paper compared their outcomes by
163 looking at the expression of a handful of marker genes, which yielded mixed results: a small
164 number of cell types seemed to have a direct match but for others the results were more
165 conflicting, with multiple types matching to one another, and others having no match at all. Here
166 we aimed to more quantitatively assess the similarity of their results, and compare them with our
167 own data which derives from phenotypically characterized sub-populations; i.e., not from
168 unsupervised expression clustering (see Table S2 for sample information).

169 MetaNeighbor relies on coordinated variation in expression level to detect cell identity, which
170 means that genes with high variability are particularly useful. Our preceding binary
171 classifications showed that genes with high mean expression were more likely to have variation
172 that allowed MetaNeighbor to learn cell identities. In the following analyses, we are examining
173 both rare and common cell types across datasets. In this case, the mean expression level of
174 marker genes should be a proxy for cell incidence: we can expect that the marker expression for
175 a more abundant type would have a higher mean expression. Since variance scales with
176 expression, the most highly variable genes in the dataset would likely only be discriminative for
177 the abundant type. Because we would like to be able to identify both abundant and rare cell
178 types, we select the genes with the highest variance at each mean expression level.

179 We identified 638 genes with high variability given their expression levels (detailed in Methods)
180 and these were used as a 'high variability gene set' to measure AUROC scores between each
181 pair of cells across datasets. When AUROCs were measured using all genes, we saw that
182 clustering was subject to strong lab-specific effects (Figure S2). In contrast, the use of variable
183 genes reproduced the known subtype structure, with major branches for the three main
184 subtypes, Pv, Sst and Htr3a.

185 To examine how the previously identified interneuron subtypes are represented across the three
186 studies, we tested the similarity of each pair of subtypes both within and across datasets using
187 the high variability gene set. For each genetically-targeted interneuron type profiled by Paul et
188 al., we found at least one corresponding subtype from the other two studies, which were defined
189 by having a mean AUROC score across training/testing folds >0.95 (Figure 3). This includes
190 Chandelier cells, a subtype that could not be definitively identified by either Tasic or Zeisel.
191 Using our reciprocal testing and training protocol we find that the Tasic_Pvalb Cpne5 subtype
192 are likely to be Chandelier cells (AUROC=0.99). In addition, expanding our criteria to include all
193 reciprocal best matches in addition to those with ID scores >0.95, we found correspondence

among five subtypes that were assessed only in the Tasic and Zeisel data, Tasic_Smad3/Zeisel_Int14 (AUROC=0.97), Tasic_Sncg/Zeisel_Int6 (AUROC=0.95), Tasic_Ndnf-Car4/Zeisel_Int15 (AUROC=0.95), Tasic_Igtp/Zeisel_Int13 (AUROC=0.94) and Tasic_Ndnf-Cxcl14/Zeisel_Int12 (AUROC=0.91). Overall, based on this high-variance gene set, we could identify 11 subtypes representing 24/45 (53%) types (Figure 3A), with total n for each subtype ranging from 25-189 out of 1583 interneurons across all datasets (1.5-11%). These results were robust to differences in data processing. Tasic *et al.* provided data as both RPKM and TPM values, and while thousands of genes had extremely divergent expression between the two, including some key markers like Vip, reciprocal average AUROCs among corresponding subtypes were nearly identical (Figure S3). Our corresponding subtypes also confirm the marker gene analysis performed by Tasic *et al.* (Table S3), without requiring manual gene curation. Because we quantify the similarity among types we can prioritize matches, and use these as input to MetaNeighbor for further evaluation.

In the above, we identified overlaps using a single gene set. To assess cell identification more broadly, we ran MetaNeighbor with these new across-dataset subtype labels, measuring predictive validity across all gene sets in GO (Figure 3A, far right). The distribution of AUROC scores varied across subtypes but we found that the score from the high variability gene set was representative of overall trends, with high performing groups showing higher mean AUROC scores over many gene sets. As detailed in the previous section, we note that AUROC scores are sensitive both to the number of training samples (n) and to underlying data features (e.g., transcriptome complexity), which complicates direct comparison of ID score distributions. Both the high mean AUROCs across all putative replicate subtypes (>0.6), and the similarity of maximum performance suggest that distinctive gene co-expression can be observed in each subtype (max AUROC=0.92 ± 0.04). As with previous tasks, we found little difference in average

218 AUROCs using functional gene sets compared to random sets (mean $AUROC_{Random}$=0.67 ±
219 0.06, mean $AUROC_{GO}$=0.68 ± 0.1).

220 These results indicate that highly variable gene sets can be used alongside pairwise testing and
221 training as a heuristic to identify replicable subtypes.

222 Investigating pyramidal neuron subtypes using MetaNeighbor

223 The heterogeneity of pyramidal neurons is undisputed, but the organizing principles are still
224 debated, with some suggesting that identity is discrete and modular (Habib et al., 2016; Zeisel
225 et al., 2015) and others purporting that identities are more likely to be described by expression
226 gradients or spectra (Cembrowski et al., 2016). With MetaNeighbor we are able to quantitatively
227 assess the degree to which pyramidal subtypes defined by scRNA-seq replicate across diverse
228 datasets. If cell types are discrete and modular, we would expect to see sharp differences, with
229 some types showing very strong similarity to one another, and strong dissimilarities to other
230 types.

231 To compare pyramidal neuron scRNA-seq datasets we permuted through all combinations of
232 subtypes as testing and training data based on a set of 743 genes with high variability given
233 their expression level (subtypes listed in Table S2). This was the same procedure that was used
234 for cortical interneurons and while there were similar numbers of subtypes in total, a smaller
235 fraction corresponded across datasets (10/48, ~21%) yielding five putative subtypes (Figure
236 3B). The AUROC score heatmap was generally less modular than the heatmap of interneuron
237 scores. The most prominent feature was that types from the hippocampus and cortex tended to
238 cluster separately from one another. Within each region-specific cluster some layer- or area-
239 specific clustering was observed but it was not completely consistent. Particular discrepancy
240 was observed between the cortical layer 5 subtypes which showed more similar AUROC score
241 profiles to the hippocampal subtypes than to other deep layer types (Tasic L5b_Cdh13,

242 L5_Chrna6, L5b_Tph). Note that these were also the same subtypes that Tasic *et al.* found no
243 match for in their marker gene analysis. We suggest that the inclusion of additional datasets
244 may help to resolve this inconsistency.

245 We assessed the five putative subtypes using MetaNeighbor. All subtypes were significantly
246 discernable compared to the null (Figure 3B) and as with the interneuron subtypes, AUROC
247 scores from the high variability gene set were well correlated with mean performance across all
248 of GO (3888 gene sets). In line with previous tasks, we found that functional gene sets
249 performed equally to random gene sets (mean $AUROC_{Random}$=0.71 ± 0.08, $AUROC_{GO}$=0.70 ±
250 0.09).

251 **Comparing gene set performance across tasks**

252 Finally, we compared gene set results from the 11 replicate interneuron subtypes and the 5
253 pyramidal neuron subtypes. In agreement with our previous results, we found that the top
254 groups were all related to neuronal function, which is unsurprising given the large size of these
255 gene sets and their likelihood of expression and variation in these cells (Figure 3C). AUROCs
256 were highly correlated across tasks (r~0.76), with slightly higher performance for identifying
257 interneuron types compared to pyramidal types (Figure 3D). The linearity of the trend across all
258 scores suggests that fundamental data features, like mean expression level and set size,
259 underlie the differential discriminative value of gene sets. The high performance across many
260 sets (mean AUROC ~0.7) also supports the notion that cell identity is encoded promiscuously
261 across the transcriptome, and is not restricted to a small set of functionally important genes.

262 **Identifying subtype specific genes**

263 ScRNA-seq experiments often seek to define marker genes for novel subtypes. Though ideally
264 marker genes are perfectly discriminative with respect to all cells, in practice marker genes are
265 often contextual and defined relative to a particular out-group. Here we aimed to identify

266 possible marker genes that would allow discrimination among interneuron subtypes or
267 pyramidal neuron subtypes. For each of our identified replicate subtypes we generated a ranked
268 list of possible marker genes by performing one-tailed, non-parametric differential expression
269 analysis within each study for all subtypes (e.g., Int1 vs. all other interneurons in the Zeisel
270 study, Int2 vs. all interneurons, etc.) and combining p-values for replicated types using Fisher's
271 method (Table S4). Figure 4A shows the FDR adjusted p-values for the top candidates based
272 on fold change for the ten replicated interneuron subtypes with overlapping differential
273 expression patterns. Figure 4B shows the same for the two pyramidal neuron subtypes with
274 overlapping differential expression patterns. The majority of these genes have previously been
275 characterized as having some degree of subtype- or layer-specific expression, for example we
276 readily identify genes that were used for the Cre-driver lines in the Tasic and Paul studies (*Sst,*
277 *Pvalb, Vip, Cck, Htr3a, Ctgf*). Even though we filtered for genes with high fold changes, we see
278 that many genes are differentially expressed in more than one subtype. Notably, considerable
279 overlap can be observed among the *Htr3a*-expressing types. For example, the Vip Sncg
280 subtype (Tasic_Vip Sncg/Paul_Vip Cck) is only subtly different from the Sncg subtype
281 (Tasic_Sncg/Zeisel_Int6) across this subset of genes, with the Sncg cells lacking differential
282 expression of *Cxcl14* and *Nr2f2*.

283 We also identify some novel candidates, including *Ptn,* or pleiotrophin, which is significantly
284 more expressed in the three Nos1-expressing subtypes than in the others (Figure 4B). It is thus
285 expected to be discriminative of Nos1-positive neurons compared to other interneuron types.
286 We validated *Ptn* expression with *in situ* hybridization and we show clear expression in neurons
287 that are positive for both *Sst* and *Nos1* (Figure 4C). *Ptn* is a growth factor, and we suggest that
288 its expression may be required for maintaining the long-range axonal connections that
289 characterize these cells. These cells are well described by current markers, however this
290 approach is likely to be of particular value for novel subtypes that lack markers, allowing

291 researchers to prioritize genes for follow-up by assessing robustness across multiple data
292 sources.

293 Discussion

294 Single-cell transcriptomics promises to have a revolutionary impact by enabling comprehensive
295 sampling of cellular heterogeneity; nowhere is this variability more profound than within the
296 brain, making it a particular focus of both single-cell transcriptomics and our own analysis into
297 its replicability. The substantial history of transcriptomic analysis and meta-analysis gives us
298 guidance about bottlenecks that will be critical to consider in order to characterize cellular
299 heterogeneity. The most prominent of these is laboratory-specific bias, likely deriving from the
300 adherence to a strict set of internal standards, which may filter for some classes of biological
301 signal (e.g., poly-A selection) or induce purely technical grouping (e.g., by sequencing depth).
302 Because of this, it is imperative to be able to align data across studies and determine what is
303 replicable. In this work, we have provided a formal means of determining replicable cell identity
304 by treating it as a quantitative prediction task. The essential premise of our method is that if a
305 cell type has a distinct transcriptional profile within a dataset, then an algorithm trained from that
306 data set will correctly identify the same type within an independent data set.

307 The currently available data allowed us to draw a number of conclusions. We validated the
308 discrete identity of eleven interneuron subtypes, and described replicate transcriptional profiles
309 to prioritize possible marker genes, including *Ptn*, a growth factor that is preferentially expressed
310 in Sst Chodl cells. We performed a similar assessment for pyramidal neurons but found less
311 correspondence among datasets, suggesting that additional data will be required to determine
312 whether there is evidence for discrete pyramidal neuron types. One major surprise of our
313 analysis is the degree of replicability in the current data. Our AUROC scores are exceptionally
314 high, particularly when considered in the context of the well-described technical confounds of

315 single-cell data. We suspect this reflects the fundamental nature of the biological problem we
316 are facing: discrete cell types can be identified by their transcriptional profiles, and the biological
317 clarity of the problem overcomes technical variation.

318 This is further suggested by our result that cell identity has promiscuous effects within
319 transcriptional data. While in-depth investigation of the most salient gene functions is required to
320 characterize cell types, to simply identify cell types is relatively straightforward. This is
321 necessarily a major factor in the apparent successes of unsupervised methods in determining
322 novel cell types and suggests that cell type identity is clearly defined by transcriptional profiles,
323 regardless of cell selection protocols, library preparation techniques or fine-tuning of clustering
324 algorithms. To us this result recalls the startling finding by Venet *et al.* that "Most random gene
325 expression signatures are related to breast cancer outcome" (Venet et al., 2011). Where, until
326 that point, research had often focused on demonstrating that highly specific genes or gene
327 clusters could predict breast cancer outcome, Venet et al. clearly demonstrated that this was a
328 more straightforward task than targeted analyses would reveal, and was due to the strength of
329 the underlying biological signal: more aggressive cancers divide more, and so anything
330 correlated with fast cycling times will be associated with poor clinical outcomes. Comparison of
331 transcriptional signatures between different cell types provides an equally clear lens. Many gene
332 sets show more correlated expression within than across types, and variation across types is
333 likely to be accounted for by simple important factors, like cell size. This is not to say that more
334 detailed characterization of cell types is not necessary: understanding the differences between
335 cells and how they work will require focused investigation into the precise molecular players that
336 are differentially utilized. However, we hope that this helps to demonstrate that the variations on
337 dimension reduction and clustering methods in single cell RNA-seq are 'working', inevitably by
338 taking advantage of this very clear signal.

339 In this work we opted to use the subtype or cluster labels provided by the original authors, in
340 essence to characterize both the underlying data as well as current analytic practices. However,
341 this has limitations where studies cluster to different levels of specificity. For example, the Tasic
342 paper defines multiple Parvalbumin subtypes but the Zeisel and Paul work do not. Our method
343 makes it extremely easy to identify highly overlapping types at the levels defined by each
344 author, facilitating downstream work to validate the sub-clusters through meta-analysis and at
345 the bench. Given the known noisiness of single-cell expression and the complex and
346 idiosyncratic character of approaches taken to assessing it, the degree of replicability that we
347 see is much higher than could have been expected were there not simple explanations for the
348 derived clusters from individual laboratories. Our work shows that with additional data,
349 comprehensive evaluation and replication is likely to be quantitatively straightforward, making it
350 possible to have high confidence in derived cell sub-types quite rapidly. As this additional data is
351 generated, our approach can provide consistent updates of the field-wide consensus.

352 The simplicity of our method makes it unlikely to be biased toward the exact cell identity tasks
353 assessed here. For example, because of the method's reliance on relative ranks, it is almost
354 entirely immune to normalization as a potential confound. On the one hand, this limits our
355 sensitivity to detect real signals of some type, but this cost is more than offset by the robustness
356 of signals identified. Its simplicity also means that it is scalable, and readily admits to the
357 incorporation of data from individual labs in their ongoing work. Ultimately we hope that by
358 defining what is replicable clearly, MetaNeighbor will allow future studies involving cell-cell
359 comparisons to build on a strong foundation toward a comprehensive delineation of cell types.

360 Experimental Procedures

361 Animals, manual cell sorting and scRNA-seq

362 Mice were bred and cared for in accordance with animal husbandry protocols at Cold Spring
363 Harbor Laboratory, with access to food and water ad libitum and a 12 hour light-dark cycle.

364 Transgenic animals bred to target the six phenotypically characterized subpopulations were
365 generated using the following breeding strategies (detailed in He et al): Nkx2.1-CreER, Pv-ires-
366 Cre animals were bred separately to Ai14 reporter to label ChC and Pv cells in the cortex. ChCs
367 were enriched in frontal cortex with tamoxifen induction at embryonic day 17.5. Intersectional
368 labeling was achieved by breeding (a) Sst-Flp, Nos1-CreER, (b) Sst-Flp, CR-Cre, (c) Vip-Flp,
369 CR-Cre and (d) Vip-Flp, Cck-Cre separately to the Ai65 intersectional reporter that labels cells
370 with tdTomato only when both the lox-STOP-lox and frt-STOP-frt cassettes are excised. Adult
371 animals (P28-35) were sacrificed by cervical dislocation to harvest brains for single cell sorting.
372 Cell sorting and scRNA-seq were performed as described previously (Crow et al., 2016). Single
373 cells were collected by manual sorting then placed into single tubes with 1µl total volume of
374 RNAseOUT (Invitrogen), 1:400K diluted ERCC spike-in RNA, and sample-specific RT primers.
375 Cells were flash frozen in liquid nitrogen then stored at -80ºC until processed. RNA was linearly
376 amplified using the MessageAmp-II kit (Life Technologies) according to the manufacturer's
377 recommended protocol. Reverse transcription of amplified aRNA was done with SuperScript-III
378 (Invitrogen) and cDNA libraries were prepared with the Illumina TruSeq small RNA library
379 preparation kit (7-11 cycles of PCR). Libraries were size-selected with SPRISelect magnetic
380 beads (Agencourt) and sequenced with paired-end 101bp reads using an Illumina HiSeq.
381 PolyA-primed reads were mapped to the mouse reference genome (mm9) with Bowtie (v
382 0.12.7), while paired sequences were used for varietal tag counting. A custom python script was
383 used map and tally sequences with unique tags for each mRNA in each cell (Crow et al., 2016).
384 All data is available to download from GEO (accession GSE92522).

385 **Public expression data**

386 Data analysis was performed in R using custom scripts (github.com/maggiecrow/MetaNeighbor,
387 2016). Processed expression data tables were downloaded from GEO directly, then subset to
388 genes appearing on both Affymetrix GeneChip Mouse Gene 2.0 ST array (902119) and the

389 UCSC known gene list to generate a merged matrix containing all samples from each
390 experiment. The mean value was taken for all genes with more than one expression value
391 assigned. Where no gene name match could be found, a value of 0 was input. We considered
392 only samples that were explicitly labeled as single cells, and removed cells that expressed fewer
393 than 1000 genes with expression >0. Cell type labels were manually curated using sample
394 labels and metadata from GEO (see Tables S1 and S2). Merged data and metadata are linked
395 through our Github page.

396 **Gene sets**

397 Gene annotations were obtained from the GO Consortium 'goslim_generic' (August 2015).
398 These were filtered for terms appearing in the GO Consortium mouse annotations
399 'gene_association.mgi.gz' (December 2014) and for gene sets with between 20-1000 genes,
400 leaving 106 GO groups with 9221 associated genes. Random gene sets were generated by
401 randomly choosing genes with the same set size distribution as GO slim. Sets of high variance
402 genes were generated by binning data from each dataset into deciles based on expression
403 level, then making lists of the top 25% of the most variable genes for each decile, excluding the
404 most highly expressed bin. The high variance set was then defined as the intersect of the high
405 variance gene lists across the relevant datasets.

406 **MetaNeighbor**

407 All scripts, sample data and detailed directions to run MetaNeighbor in R can be found on our
408 Github page (github.com/maggiecrow/MetaNeighbor, 2016).

409 The input to MetaNeighbor is a set of genes, a data matrix and two sets of labels: one set for
410 labeling each experiment, and one set for labeling the cell types of interest. For each gene set,
411 the method generates a cell-cell similarity network by measuring the Spearman correlation
412 between all cells across the genes within the set, then ranking and standardizing the network so

413 that all values lie between 0 and 1. The use of rank correlations means that the method is
414 robust to any rank-preserving normalization (i.e., log2, TPM, RPKM). Ranking and standardizing
415 the networks ensures that distributions remain uniform across gene sets, and diminishes the
416 role outlier similarities can play since values are constrained.

417 The node degree of each cell is defined as the sum of the weights of all edges connected to it
418 (i.e., the sum of the standardized correlation coefficients between each cell and all others), and
419 this is used as the null predictor in the neighbor voting algorithm to standardize for a cell's 'hub-
420 ness': cells that are generically linked to many cells are preferentially down-weighted, whereas
421 those with fewer connections are less penalized. For each cell type assessment, the neighbor
422 voting predictor produces a weighted matrix of predicted labels by performing matrix
423 multiplication between the network and the binary vector (0,1) indicating cell type membership,
424 then dividing each element by the null predictor (i.e., node degree). In other words, each cell is
425 given a score equal to the fraction of its neighbors, including itself, which are part of a given cell
426 type (Ballouz et al., 2016). For cross-validation, we permute through all possible combinations
427 of leave-one-dataset-out cross-validation, sequentially hiding each experiment's cell labels in
428 turn, and then reporting how well we can recover cells of the same type as the mean area under
429 the receiver operator characteristic curve (AUROC) across all folds. A key difference from
430 conventional cross-validation is that there is no labeled data within the dataset for which
431 predictions are being made. Labeled data comes only from external datasets, ensuring
432 predictions are driven by signals that are replicable across data sources. To improve speed,
433 AUROCs are calculated analytically, where the AUROC for each cell type *j*, is calculated based
434 on the sum of the ranks of the scores for each cell *i*, belonging to that cell type. This can be
435 expressed as follows:

$$AUROC_j = \sum_{i}^{\mathrm{N}} \frac{Ranks_i}{N * N_{Neg}} - \frac{N+1}{2 * N_{Neg}}$$

436 where N is the number of true positives, and N_{Neg} is the number of true negatives. Note that for
437 experiments with only one cell type this cannot be computed as there are no true negatives.
438 AUROCs are reported as averages across all folds of cross-validation for each gene set
439 (excluding NAs from experiments with no negatives), and the distribution across gene sets is
440 plotted.

441 To test the dependency of results on the amount of training and testing data we repeated the
442 neuron vs. non-neuronal cell discrimination task after randomly selecting between two and
443 seven datasets ten times each. This was done for 21 representative gene sets. Means for each
444 gene set and each number of included datasets were plotted.

445 **Identifying putative replicates**

446 In cases where cell identity was undefined across datasets (i.e., cortical interneuron and
447 pyramidal subtypes) we treated each subtype label as a positive for each other subtype, and
448 assessed similarity over the high variance gene set described above. For example, Int1 from the
449 Zeisel dataset was used as the positive (training) set, and all other subtypes were considered
450 the test set in turn. Mean AUROCs from both testing and training folds are plotted in the
451 heatmap in Figure 3. A stringent cut-off of mean AUROC >0.95 and/or mutual best matches
452 across datasets identified putative replicated types for further assessment with our supervised
453 framework (detailed above). While lowering this threshold could increase the number of
454 subtypes with some match, we found that reciprocal top hits alone provided an upper bound on
455 the number of replicated types (i.e., lowering the thresholds did not allow for a higher number of
456 subtypes). New cell type labels encompassing these replicate types (e.g. a combined Sst-Chodl
457 label containing Int1 (Zeisel), Sst Chodl (Tasic) and Sst Nos1 (Paul)) were generated for
458 MetaNeighbor across random and GO sets, and for meta-analysis of differential expression.
459 While only reciprocal top-hits across laboratories were used to define novel cell-types,

460 conventional cross-validation within laboratories was performed to fill in AUROC scores across
461 labels contained within each lab.

462 Differential expression

463 For each cell type within a dataset (defined by the authors' original labeling), differential gene
464 expression was calculated using a one-sided Wilcoxon rank-sum test, comparing gene
465 expression within a given cell type to all other cells within the dataset (e.g., Zeisel_Int1 vs all
466 other Zeisel interneurons). Meta-analytic p-values were calculated for each putative replicated
467 type using Fisher's method (Fisher, 1925) then a multiple hypothesis test correction was
468 performed with the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Top
469 differentially expressed genes were those with an adjusted meta-analytic p-value <0.001 and
470 with log2 fold change >2 in each dataset. All differential expression data for putative replicated
471 subtypes can be found in Table S4.

472 Supplementary Material

473 Supplementary tables and figures may be accessed at the following link: http://bit.ly/2s58zPd

474 Author Contributions

475 JG conceived the study. JG, MC, and JH designed experiments. MC and JG wrote the
476 manuscript. MC and SB performed computational experiments. AP performed cell sorting, and
477 generated and parsed the raw sequencing data. JH supervised wet-lab data collection. All
478 authors read and approved the final manuscript.

479 Acknowledgments

480 MC, SB and JG were supported by a gift from T. and V. Stanley. Z.J.H. was supported by NIH
481 5R01MH094705-04, R01MH109665-01 and the CSHL Robertson Neuroscience Fund. A.P. was
482 supported by a NARSAD Postdoctoral Fellowship. The authors would like to thank Paul
483 Pavlidis, Bo Li and Jessica Tollkuhn for their thoughtful feedback on earlier drafts of this

484 manuscript. We would also like to thank the dedicated researchers who have made their data

485 publicly available. Our work would not be possible without their valuable contributions.

486 References

487 Ascoli, G.A., Alonso-Nanclares, L., Anderson, S.A., Barrionuevo, G., Benavides-Piccione, R.,
488 Burkhalter, A., Buzsaki, G., Cauli, B., Defelipe, J., Fairen, A.*, et al.* (2008). Petilla terminology:
489 nomenclature of features of GABAergic interneurons of the cerebral cortex. Nat Rev Neurosci *9*,
490 557-568.

491 Ballouz, S., Weber, M., Pavlidis, P., and Gillis, J. (2016). EGAD: ultra-fast functional analysis of
492 gene networks. Bioinformatics (Oxford, England).

493 Baron, M., Veres, A., Wolock, Samuel L., Faust, Aubrey L., Gaujoux, R., Vetere, A., Ryu,
494 Jennifer H., Wagner, Bridget K., Shen-Orr, Shai S., Klein, Allon M.*, et al.* (2016). A Single-Cell
495 Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population
496 Structure. Cell Systems *3*, 346-360.e344.

497 Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and
498 Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B
499 (Methodological) *57*, 289-300.

500 Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann,
501 S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in
502 single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature biotechnology
503 *33*, 155-160.

504 Cembrowski, M.S., Bachman, J.L., Wang, L., Sugino, K., Shields, B.C., and Spruston, N.
505 (2016). Spatial Gene-Expression Gradients Underlie Prominent Heterogeneity of CA1 Pyramidal
506 Neurons. Neuron *89*, 351-368.

507 Crow, M., Paul, A., Ballouz, S., Huang, Z.J., and Gillis, J. (2016). Exploiting single-cell
508 expression to characterize co-expression replicability. Genome biology *17*, 101.

509 Dudoit, S., Fridlyand, J., and Speed, T.P. (2002). Comparison of Discrimination Methods for the
510 Classification of Tumors Using Gene Expression Data. Journal of the American Statistical
511 Association *97*, 77-87.

512 Dueck, H., Khaladkar, M., Kim, T.K., Spaethling, J.M., Francis, C., Suresh, S., Fisher, S.A.,
513 Seale, P., Beck, S.G., Bartfai, T.*, et al.* (2015). Deep sequencing reveals cell-type-specific
514 patterns of single-cell transcriptome variation. Genome biology *16*, 122.

515 Fisher, R.A. (1925). Statistical methods for research workers (Edinburgh, London,: Oliver and
516 Boyd).

517 github.com/maggiecrow/MetaNeighbor (2016). MetaNeighbor: a method to rapidly assess cell
518 type identity using both functional and random gene sets.

519 Grun, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van
520 Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell
521 types. Nature *525*, 251-255.

522 Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J.J., Hession, C.,
523 Zhang, F., and Regev, A. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare
524 adult newborn neurons. Science (New York, NY) *353*, 925-928.

525 He, M., Tucciarone, J., Lee, S., Nigro, M.J., Kim, Y., Levine, J.M., Kelly, S.M., Krugikov, I., Wu,
526 P., Chen, Y.*, et al.* (2016). Strategies and Tools for Combinatorial Targeting of GABAergic
527 Neurons in Mouse Cerebral Cortex. Neuron *91*, 1228-1243.

528 Hicks, S.C., Teng, M., and Irizarry, R.A. (2015). On the widespread and critical impact of
529 systematic bias and batch effects in single-cell RNA-Seq data.

530 Kapp, A.V., and Tibshirani, R. (2007). Are clusters found in one dataset present in another
531 dataset? Biostatistics (Oxford, England) *8*, 9-31.

532 Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz,
533 D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to
534 embryonic stem cells. Cell *161*, 1187-1201.

535 La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L.E., Stott,
536 S.R., Toledo, E.M., Villaescusa, J.C.*, et al.* (2016). Molecular Diversity of Midbrain Development
537 in Mouse, Human, and Stem Cells. Cell *167*, 566-580.e519.

538 Lun, A.T., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA
539 sequencing data with many zero counts. Genome biology *17*, 75.

540 Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas,
541 A.R., Kamitaki, N., Martersteck, E.M.*, et al.* (2015). Highly Parallel Genome-wide Expression
542 Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.

543 Min, J.W., Kim, W.J., Han, J.A., Jung, Y.J., Kim, K.T., Park, W.Y., Lee, H.O., and Choi, S.S.
544 (2015). Identification of Distinct Tumor Subpopulations in Lung Adenocarcinoma via Single-Cell
545 RNA-seq. PLoS One *10*, e0135817.

546 Muraro, Mauro J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L.,
547 Engelse, Marten A., Carlotti, F., de Koning, Eelco J.P.*, et al.* (2016). A Single-Cell
548 Transcriptome Atlas of the Human Pancreas. Cell Systems *3*, 385-394.e383.

549 Peng, R. (2016). A Simple Explanation for the Replication Crisis in Science. In Simply Statistics.

550 Pierson, E., and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene
551 expression analysis. Genome biology *16*, 241.

552 Poulin, J.-F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J.M., and Awatramani, R. (2016).
553 Disentangling neural cell diversity using single-cell transcriptomics. Nature neuroscience *19*,
554 1131-1141.

555 Segerstolpe, A., Palasantza, A., Eliasson, P., Andersson, E.M., Andreasson, A.C., Sun, X.,
556 Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K.*, et al.* (2016). Single-Cell Transcriptome
557 Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell metabolism *24*, 593-
558 607.

559 Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis,
560 X., Levin, J.Z., Nemesh, J., Goldman, M.*, et al.* (2016). Comprehensive Classification of Retinal
561 Bipolar Neurons by Single-Cell Transcriptomics. Cell *166*, 1308-1323.e1330.

562 Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T.,
563 Sorensen, S.A., Dolbeare, T.*, et al.* (2016). Adult mouse cortical cell taxonomy revealed by
564 single cell transcriptomics. Nature neuroscience *19*, 335-346.

565 Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J.,
566 Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung
567 epithelium using single-cell RNA-seq. Nature *509*, 371-375.

568 Vallejos, C.A., Marioni, J.C., and Richardson, S. (2015). BASiCS: Bayesian Analysis of Single-
569 Cell Sequencing Data. PLoS Comput Biol *11*, e1004333.

570 Venet, D., Dumont, J.E., and Detours, V. (2011). Most Random Gene Expression Signatures
571 Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol *7*, e1002240.

572 Wang, Y.J., Schug, J., Won, K.-J., Liu, C., Naji, A., Avrahami, D., Golson, M.L., and Kaestner,
573 K.H. (2016). Single cell transcriptomics of the human endocrine pancreas. Diabetes.

574 Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A.,
575 Marques, S., Munguba, H., He, L., Betsholtz, C.*, et al.* (2015). Brain structure. Cell types in the
576 mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (New York, NY) *347*,
577 1138-1142.

578

Figures

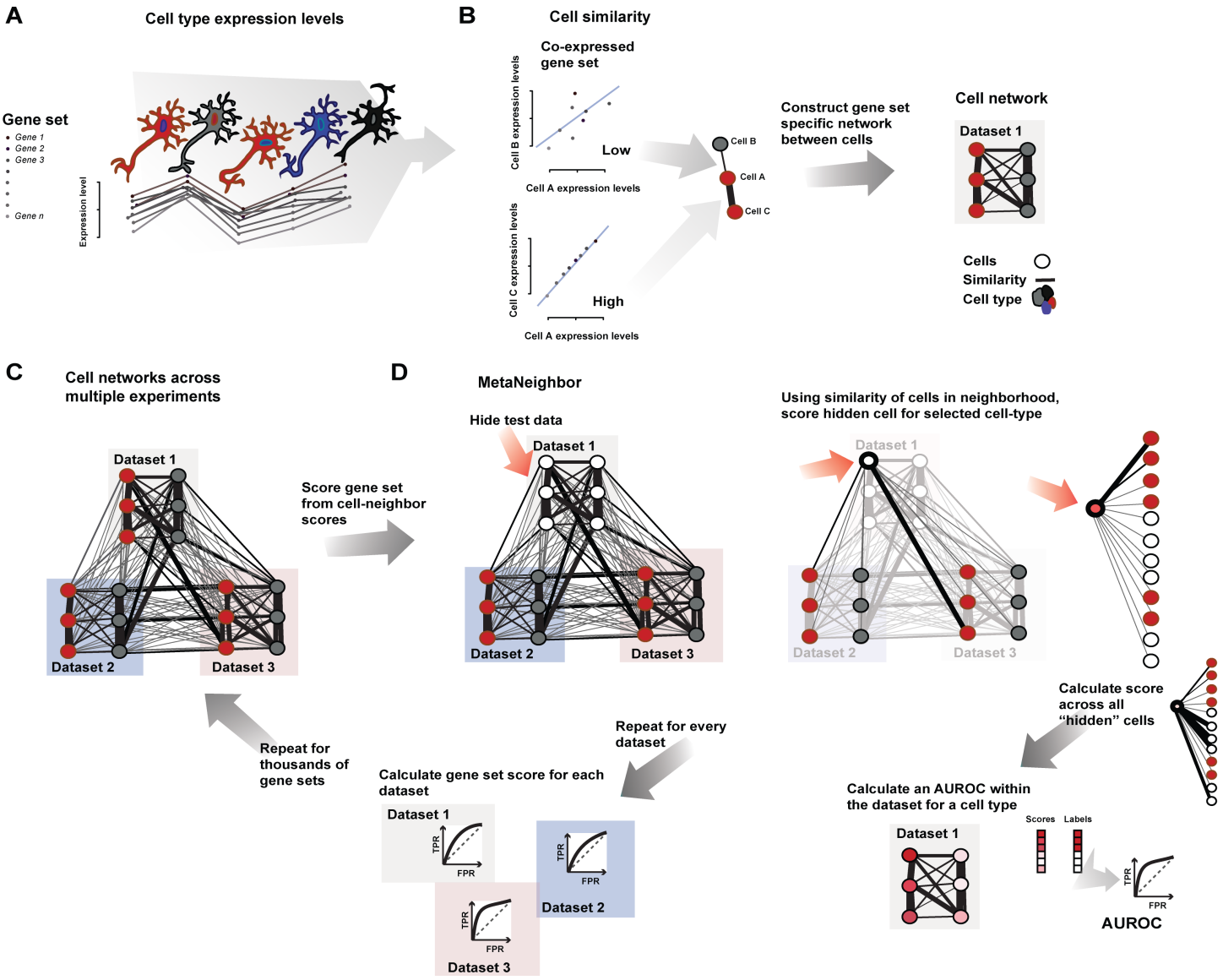


Figure 1 – MetaNeighbor quantifies cell type identity across experiments
A – Schematic representation of gene set co-expression across individual cells. Cell types are indicated by their color. **B** – Similarity between cells is measured by taking the correlation of gene set expression between individual cells. On the top left of the panel, gene set expression between two cells, A and B, is plotted. There is a weak correlation between these cells. On the bottom left of the panel we see the correlation between cells A and C, which are strongly correlated. By taking the correlations between all pairs of cells we can build a cell network (right), where every node is a cell and the edges represent how similar each cell is to each other cell. **C -** The cell network that was generated in B can be extended to include data from multiple experiments (multiple datasets). The generation of this multi-dataset network is the first step of MetaNeighbor. **D –** The cross-validation and scoring scheme of MetaNeighbor is demonstrated in this panel. To assess cell type identity across experiments we use neighbor voting in cross-validation, systematically hiding the labels from one dataset at a time. Cells within the hidden dataset are predicted as similar to the cell types from other datasets, using a neighbor voting formalism. Whether these scores prioritize cells as the correct type within the dataset determines the performance, expressed as the AUROC. In other words, comparative assessment of cells occurs only within a dataset, but this is based only on training information from outside that dataset. This is then repeated for all gene sets of interest.

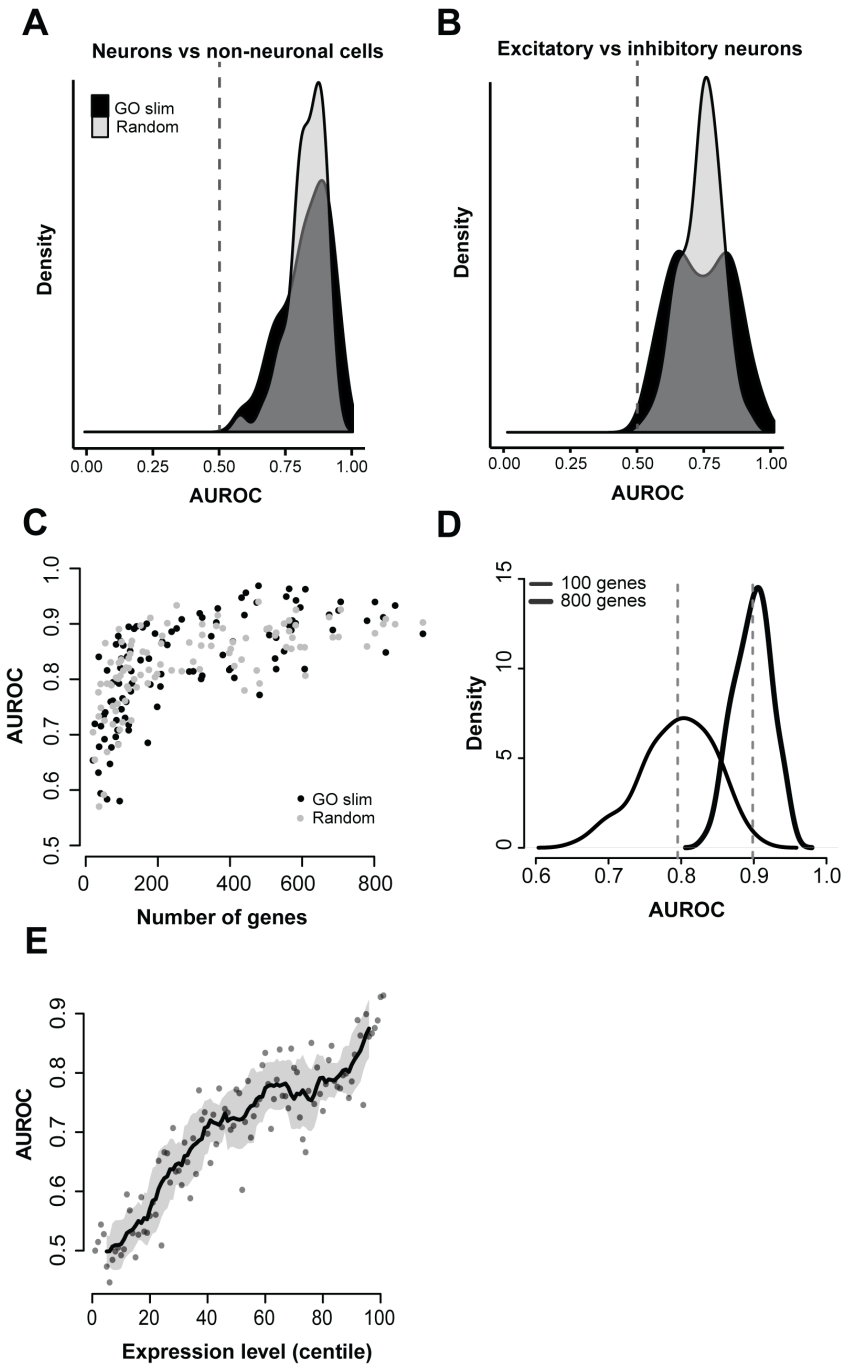
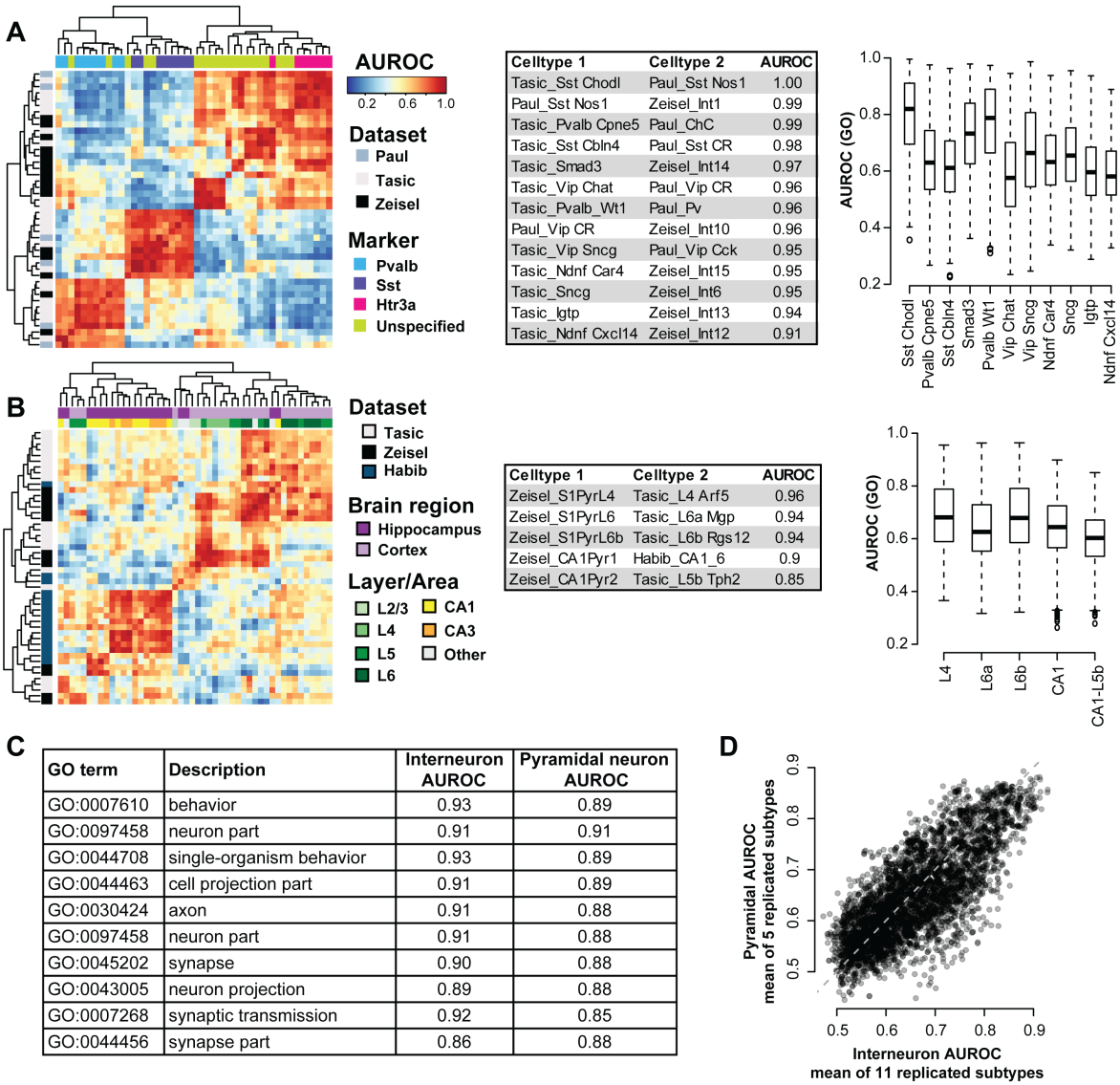


Figure 2 – Cell type identity is widely represented in the transcriptome
A & B – Distribution of AUROC scores from MetaNeighbor for discriminating neurons from non-neuronal cells ("task one", A) and for distinguishing excitatory vs. inhibitory neurons ("task two", B). GO scores are in black and random gene set scores are plotted in gray. Dashed grey lines indicate the null expectation for correctly guessing cell identity (AUROC=0.5). For both tasks, almost any gene set can be used to improve performance above the null, suggesting widespread encoding of cell identity across the transcriptome. **C** – Task one AUROC scores for each gene set are plotted with respect to the number of genes. A strong, positive relationship is observed between gene set size and AUROC score, regardless of whether genes were chosen randomly or based on shared functions. **D** – Distribution of AUROC scores for task one using 100 sets of 100 randomly chosen genes, or 800 randomly chosen genes. The mean AUROC score is significantly improved with the use of larger gene sets (mean 100 = 0.80 +/- 0.05, mean 800 = 0.90 +/- 0.03). **E** – Relationship between AUROC score and expression level. Task one was re-run using sets of genes chosen based on mean expression. A strong positive relationship was observed between expression level and performance (r_s ~0.9).



Celltype 1	Celltype 2	AUROC
Tasic_Sst Chodl	Paul_Sst Nos1	1.00
Paul_Sst Nos1	Zeisel_Int1	0.99
Tasic_Pvalb Cpne5	Paul_ChC	0.99
Tasic_Sst Cbln4	Paul_Sst CR	0.98
Tasic_Smad3	Zeisel_Int14	0.97
Tasic_Vip Chat	Paul_Vip CR	0.96
Tasic_Pvalb_Wt1	Paul_Pv	0.96
Paul_Vip CR	Zeisel_Int10	0.96
Tasic_Vip Sncg	Paul_Vip Cck	0.95
Tasic_Ndnf Car4	Zeisel_Int15	0.95
Tasic_Sncg	Zeisel_Int6	0.95
Tasic_Igtp	Zeisel_Int13	0.94
Tasic_Ndnf Cxcl14	Zeisel_Int12	0.91

Celltype 1	Celltype 2	AUROC
Zeisel_S1PyrL4	Tasic_L4 Arf5	0.96
Zeisel_S1PyrL6	Tasic_L6a Mgp	0.94
Zeisel_S1PyrL6b	Tasic_L6b Rgs12	0.94
Zeisel_CA1Pyr1	Habib_CA1_6	0.9
Zeisel_CA1Pyr2	Tasic_L5b Tph2	0.85

GO term	Description	Interneuron AUROC	Pyramidal neuron AUROC
GO:0007610	behavior	0.93	0.89
GO:0097458	neuron part	0.91	0.91
GO:0044708	single-organism behavior	0.93	0.89
GO:0044463	cell projection part	0.91	0.89
GO:0030424	axon	0.91	0.88
GO:0097458	neuron part	0.91	0.88
GO:0045202	synapse	0.90	0.88
GO:0043005	neuron projection	0.89	0.88
GO:0007268	synaptic transmission	0.92	0.85
GO:0044456	synapse part	0.86	0.88

Figure 3 – Cross-dataset analysis of interneuron and pyramidal neuron diversity
A – (Left) Heatmap of AUROC scores between interneuron subtypes based on the highly variable gene kernel. Dendrograms were generated by hierarchical clustering of Euclidean distances using average linkage. Row colors indicate data origin and column colors show marker expression. Clustering of AUROC score profiles recapitulates known cell type structure, with major branches representing the Pv, Sst and Htr3a lineages. (Middle) Table of reciprocal best matches and subtype pairs with scores >0.95. (Right) Boxplots of GO performance (3888 sets) for each replicated subtype, ordered by their AUROC score from the highly variable gene set. Subtypes are labeled with the names from Tasic *et al*. A positive relationship is observed between AUROC scores from the highly variable set and the average AUROC score for each subtype. Mean AUROCs are all greater than chance (0.5) suggesting robust cross-dataset replication across gene sets. **B** – (Left) Heatmap of AUROC scores between pyramidal subtypes based on the highly variable gene kernel, clustered as in A. Row colors indicate datasets and column colors show brain region, cortical layer or hippocampal area. Clustering of AUROC score profiles shows a separation of cortical and hippocampal subtypes. (Middle) Table of reciprocal best matches. (Right) Boxplots of GO performance (3888 sets) for each replicated subtype, ordered by their AUROC score from the highly variable gene set. Subtypes are labeled by layer. A positive relationship is observed between ID scores from the highly variable set and the average AUROC for each subtype. **C** – The table shows the top GO terms that allow for cross-dataset subtype discrimination, listed by their mean AUROC across tasks. For both tasks, high scores are obtained for terms related to neuronal function. **D** – AUROC scores for each GO function are plotted, with pyramidal scores on the y-axis and interneuron scores on the x-axis. AUROCs are highly correlated across tasks (r_s~0.76), suggesting limited functional specificity.

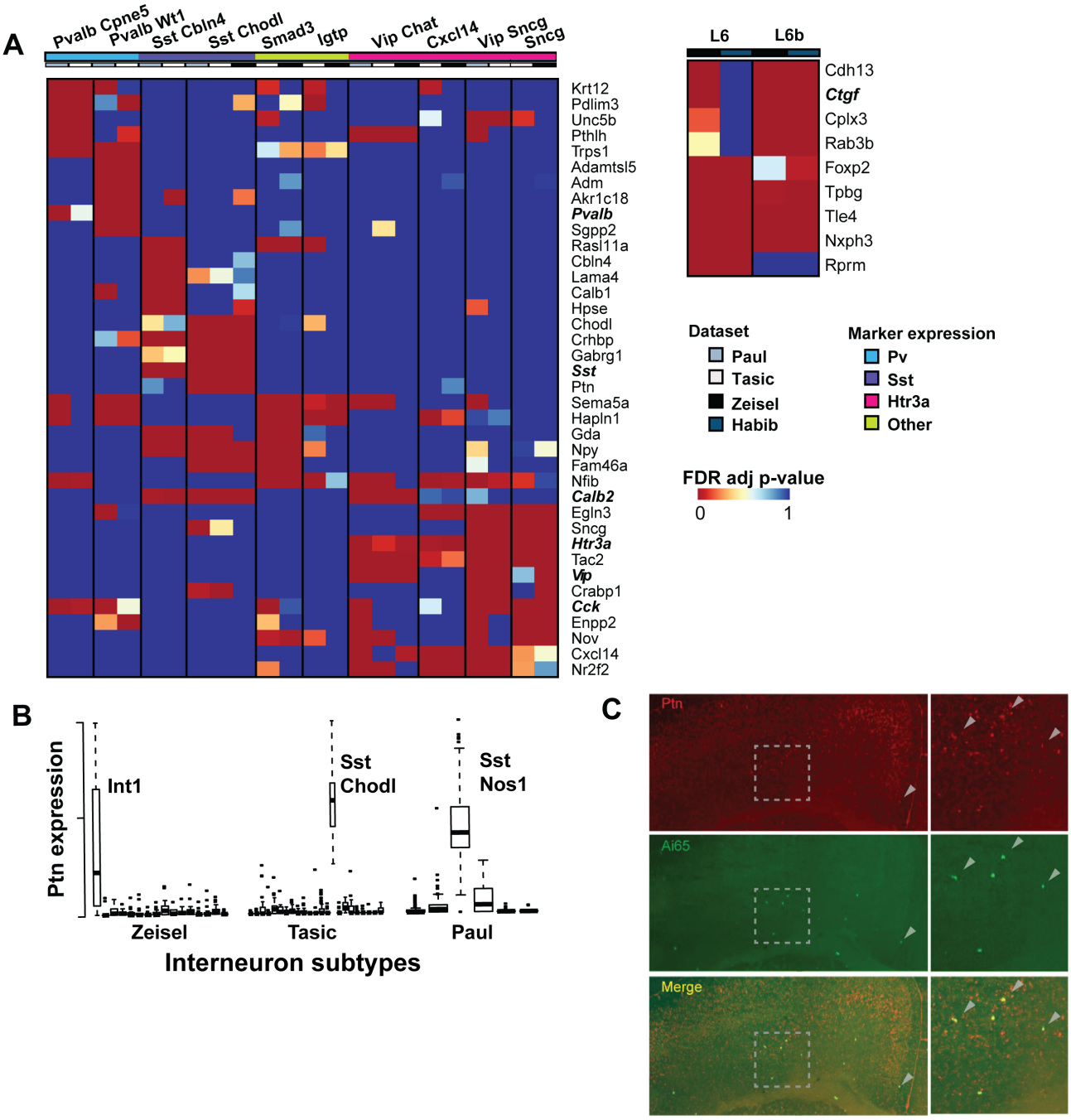


Figure 4 – Replicated subtypes show consistent differential expression
A – (Top) Heatmap of FDR adjusted p-values of top differentially expressed genes among replicated interneuron subtypes (NB only ten subtypes are shown as no differentially expressed genes were found for the Ndnf Car4 subtype). Subtype names are listed at the top of the columns and are labeled as in Tasic *et al*. Many genes are commonly differentially expressed among multiple subtypes, but combinatorial patterns distinguish them. (Right) Heatmap of FDR adjusted p-values of top differentially expressed genes among replicated pyramidal neuron subtypes. (NB only the two with overlapping differential expression are shown). Subtypes are labeled by layer. **B** – Standardized Ptn expression is plotted across the three experiments, where each box represents an interneuron subtype. High, but variable expression is observed across the three Sst Chodl types. **C** – Fluorescent double in-situ of Ai14/tdTomato driven by Sst-Flp and Nos-Cre expression (green) and Ptn (red). Dotted box indicates the area shown in higher magnification on the right, arrowheads point to cells that express both transcripts.