Meta-analytic markers reveal a generalizable description of cortical cell types

Stephan Fischer[1] and Jesse Gillis[1,2,*]

[1]Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY 11724, USA

[2]Cold Spring Harbor Laboratory, Watson School of Biological Sciences, Cold Spring Harbor, NY, USA

*Correspondence: jgillis@cshl.edu

Abstract

Our understanding of neuronal cell types has advanced considerably with the publication of single cell atlases, mapping hundreds of putative cell types across brain areas. Characterizing these cell types is a key step towards understanding how the brain works and dysfunctions. Marker genes play an essential role for experimental validation and computational analyses such as physiological characterization through pathway enrichment, annotation, and deconvolution. However, a framework for quantifying marker replicability and picking replicable markers is currently lacking, particularly for cell types at the finest resolution. Here, using high quality data from the Brain Initiative Cell Census Network (BICCN), we systematically investigate marker replicability for the proposed BICCN cell types. We show that, due to dataset-specific noise, we need to combine 5 datasets to obtain robust differentially expressed (DE) genes, particularly for rare populations and lowly expressed genes. We find that two standard DE statistics, fold change and AUROC, can be combined to select the best markers, ranging from highly sensitive to highly specific. Finally, we show that a meta-analytic selection of markers improves performance in downstream

computational tasks: cluster separability, annotation and deconvolution. Based on these tasks, we estimate that 10 to 200 meta-analytic markers provide optimal performance, offering an interpretable and robust description of cell types. Replicable marker lists condense single cell atlases into generalizable and accessible information about cell types, opening avenues for multiple downstream applications, including cell type annotation, selection of gene panels and bulk data deconvolution.

Keywords

Cell types, single-cell RNA sequencing, replicability, marker genes, cell type taxonomy.

Introduction

Recent atlas efforts based on single cell technologies have led to comprehensive cell type taxonomies that include a multitude of novel cell types (Tasic et al. 2018; Zeisel et al. 2018; Schaum et al. 2018; Packer et al. 2019; Cao et al. 2020). The discovery of new cell types and novel biological heterogeneity served as a foundation for promising avenues for the understanding of tissue homeostasis and disease. However, to develop downstream applications and experiments, an actionable description of cell types is required that extends beyond taxonomic classification. While sporadic post-hoc markers are published alongside taxonomies, the replicability of these markers is rarely assessed. Here, we systematically evaluate marker replicability and propose unprecedented lists of replicable markers (or meta-markers) for neuronal cell types by selecting an optimal number of robustly upregulated genes across a compendium of brain datasets.

Given the rapid progression in the number and size of single-cell datasets (Svensson et al. 2018), making atlases easily accessible is an increasingly difficult challenge. Cell type centroids provide an efficient summary of active gene expression programs (Zeisel et al. 2018), but they are subject to batch effects (Tung et al. 2017) and discard expression variability. While integrative methods have

been successful at mitigating batch effects for the joint analysis of a small groups of datasets (Butler et al. 2018; Haghverdi et al. 2018; Welch et al. 2019; Korsunsky et al. 2019; Lin et al. 2019) and the transfer of cell type annotations (Kiselev et al. 2018; Stuart et al. 2019), the abstract embedding of cell types is costly, as well as difficult to interpret and to extract for downstream applications. In contrast, markers provide an interpretable common denominator that does not involve data re-analysis or complex mathematical transformations; they are commonly used for functional characterization (Mancarci et al. 2017), cell type annotation (Poulin et al. 2016; Johnson and Walsh 2017; Pliner et al. 2019; Zhang et al. 2019), deconvolution of bulk data (Wang et al. 2019; Newman et al. 2019; Patrick et al. 2020) and spatial data (Qian et al. 2020), selection of representative gene panels (Moffitt et al. 2018), cross-species comparisons (Tosches et al. 2018; Hodge et al. 2019; Krienen et al. 2019; Bakken et al. 2020), and mapping of organoids to in vivo progenitors (Velasco et al. 2019; Bhaduri et al. 2020). For many of these applications, the strength of individual markers is limited by the lack of conservation (Bakken et al. 2020) and the sporadic expression in individual cells (Kharchenko et al. 2014; Risso et al. 2018; Hicks et al. 2018; Chen and Zhou 2018). Moving past individual markers to small lists is done sporadically to capture combinatorial relationships or improve power, but has not yet exploited the full power of scRNA-seq data. In specific, because cell types are encoded in a low-dimensional expression space (Crow and Gillis 2018), we hypothesize that they can be captured with high resolution and generalizable definitions using redundant and robust marker lists. These lists can then easily be compared and combined across datasets for downstream analyses.

The problem of finding generalizable descriptions of cell types has a long history in the brain, famously illustrated by Ramon y Cajal's morphology-based descriptions (RAMON Y CAJAL 1904). More recently, the Petilla convention emphasized the need to describe neurons according to a multi-modal taxonomy, including morphology, electrophysiology, connectivity and transcriptomics (Ascoli et al. 2008). Single cell data, while only covering one aspect of this multi-modal description, have enabled unprecedented wide and deep sampling of brain cells, with current taxonomies containing

71 several hundred cell types (Tasic et al. 2018; Zeisel et al. 2018). They thus offer a chance to assess
72 the robustness of transcriptomic cell types, but current cell types are usually defined based on data
73 from a single lab and a single computational method, while an ideal description should be
74 community-based and method-independent (Yuste et al. 2020). With the recent publication of
75 several single-cell compendia by the Brain Initiative Cell Census Network (BICCN)(Yao et al. 2020a,
76 2020b), the brain offers a unique opportunity to characterize marker-based descriptions.

77 In this manuscript, we systematically assess the replicability of markers for BICCN cell types. We
78 identify robust markers (meta-markers) across a compendium of 7 brain single cell datasets
79 containing a total of 482,712 cells from the BICCN, one of the most complex and comprehensive cell
80 type taxonomies to date. The assessment procedure is based on two simple steps: (i) identify
81 markers from single datasets, (ii) obtain a list of meta-markers by selecting replicable markers. The
82 compendium samples from 6 single-cell and single-nuclei technologies, resulting in meta-markers
83 that are robust to the varying sensitivity and contamination levels of these technologies. We further
84 investigate the ability of markers to recapitulate cell types at various levels of granularity. We define
85 two simple performance axes, intuitively representing coverage (being expressed in all cells of
86 interest) and signal-to-noise ratio (being expressed exclusively in cells of interest), that can be
87 efficiently summarized using standard differential expression statistics. While individual meta-
88 markers only imperfectly capture cell types, we find that aggregating 10 to 200 meta-markers leads
89 to optimal performance in downstream computational analyses, such as cell type annotation and
90 deconvolution. Remarkably, these marker-based descriptions, derived from the primary motor
91 cortex, generalize to other cortical brain regions, enabling accurate annotation of individual cells.
92 Robust meta-markers thus provide a simple and actionable description of BICCN cell types, which we
93 make available as high-quality marker lists (Sup. Data 1-3) ranging from the lowest resolution
94 (excitatory neurons, inhibitory neurons, non-neurons) to the finest resolution defined by the BICCN
95 (85 neuronal cell types).

Results

The ideal marker gene fulfills two criteria: (1) it is expressed in all cells of the population of interest, providing high coverage, (2) it is not expressed in background cells, providing a high signal-to-noise ratio (Fig. 1a). In recently published atlases, it is often unclear how strongly and robustly the proposed markers fulfill these criteria, particularly at high clustering resolution. To investigate replicability of marker strength, our basic strategy is to look for simple statistics that can be robustly averaged across datasets and correctly capture coverage and signal-to-noise. We focused on a BICCN neuron atlas containing 7 datasets with 482,712 cells, organized into a hierarchy of 116 cell types in 3 levels of increasing resolution: classes, subclasses and clusters (Yao et al. 2020a)(Table 1).

Dataset	Brain regions	Assay	Technology	# cells	# cell types	# genes detected	# UMIs / reads
scSS	MOp	Cell	SmartSeq	6,288	61	9,420	1,750,664
snSS	MOp	Nucleus	SmartSeq	6,171	46	4,363	613,762
scCv2	MOp	Cell	10X v2	122,641	90	4,594	12,697
snCv2	MOp	Nucleus	10X v2	76,525	43	1,716	3,145
snCv3M	MOp	Nucleus	10X v3	159,738	113	4,237	12,060
scCv3	MOp	Cell	10X v3	71,183	78	7,282	46,148
snCv3Z	MOp	Nucleus	10X v3	40,166	67	3,445	16,088
AUD	AUD	Cell	10X v2	71,670	203	3,969	10,105
Isocortex-Hippocampus	21	Cell	SmartSeq	827 to 16,318	13 to 183	6,099 to 9,006	488,099 to 2,016,775
Isocortex-Hippocampus	19	Cell	10X v2	18,307 to 216,203	166 to 263	2,874 to 4,944	6,102 to 15,272

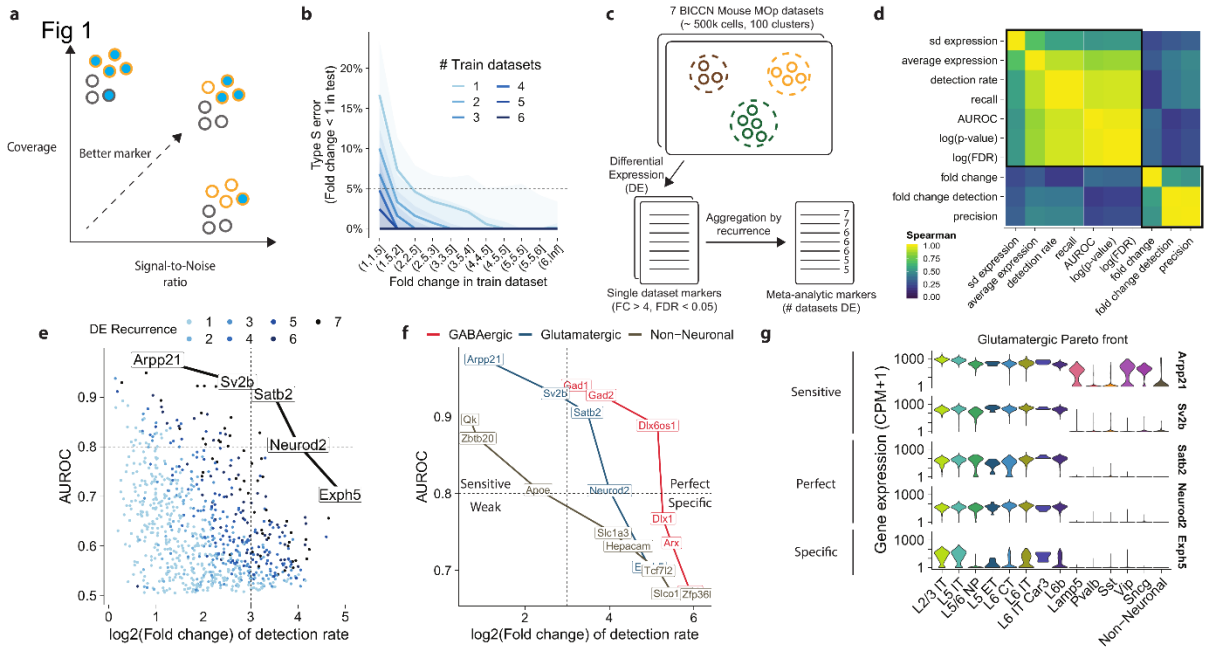
Table 1. List of Brain Initiative Cell Census Network (BICCN) datasets used in this study. *All datasets are from mouse. MOp corresponds to the primary motor cortex, AUD to the auditory cortex. The "# genes detected" column contains the median number of genes detected per cell. The "# UMI / reads" column contains either the median number of reads per cell (for SmartSeq datasets) or the median number of UMIs per cell (for 10X datasets).*

Meta-analytic markers are highly replicable

We started by investigating the replicability of standard differential expression (DE) statistics across BICCN datasets. Previous experiments in microarray and bulk RNAseq data by the MAQC (Shi et al. 2006) and SEQC (Consortium et al. 2014) consortia established that a fold change (FC) threshold between 2 and 4 was necessary to obtain replicable DE genes. We wondered if a similar threshold would hold for single cell RNAseq and how aggregation across datasets would improve the threshold

116 for fold change (FC) and the area under the receiver-operator curve (AUROC), a statistic routinely
117 used to compute the statistical significance of DE.

118 To assess the replicability of FC, we quantified how often one would draw inconsistent conclusions
119 about a significant DE gene being upregulated (type S error (Gelman and Carlin 2014)). For example,
120 given that I observed a gene with FC=2 (strongly upregulated), what is the probability that my gene
121 will have a FC<1 (downregulated) in an independent experiment? When FC was estimated from a
122 single dataset, as is routine in published studies, we found that a threshold of FC>4 was necessary to
123 call a gene reliably upregulated (type S error < 5%, Fig. 1b), in line with MAQC/SEQC conclusions. In
124 contrast, estimating FC from a higher number of datasets dramatically improved replicability: for 2
125 datasets the 5% error threshold is reached at FC > 2, for 3 datasets at FC > 1.5. Surprisingly, for more
126 than 5 datasets, our results suggest that thresholding becomes unnecessary: a gene that was
127 detected as upregulated in 5 independent datasets was almost always upregulated in the
128 2 remaining datasets, even at low effect size (FC~1). Moreover, for a single dataset, only the top 10
129 upregulated genes were replicable, while the top 1000 genes are reliably upregulated when
130 aggregating across 6 datasets (Sup. Fig. 1b). We observed similar trends for AUROCs. Based on a
131 single dataset, the replicability threshold was AUROC>0.65, yielding 100 reliably upregulated genes.
132 Aggregating six datasets, no replicability threshold was needed and we could identify more than
133 5000 reliably upregulated genes (Sup. Fig. 1a,c). The impact of dataset aggregation was particularly
134 dramatic for small clusters and lowly expressed genes (Sup. Fig. 1d-g); for 5/85 neuron clusters,
135 fewer than 5 of the top 10 single dataset markers (based on fold change) were reliably upregulated.



136

137 ***Figure 1. The meta-analytic Pareto front of markers: a trade-off between coverage and signal-to-noise ratio. a*** *Ideal*
138 *markers have high coverage (high expression in cells of interest) and high signal-to-noise ratio (relatively low expression in*
139 *background cells).* ***b*** *Fraction of genes inconsistently detected as upregulated (type S error) depending on the fold change in*
140 *the training dataset. Colors indicate the number of datasets used to estimate the fold change (geometric mean).* ***c***
141 *Schematic of extraction of meta-analytic markers: differentially expressed (DE) genes are computed independently in each*
142 *dataset, meta-markers are selected based on the number of times they were DE across datasets.* ***d*** *Spearman correlation of*
143 *standard DE statistics for putative markers (averaged over datasets). We highlight two independent groups of statistics that*
144 *can serve as a proxy for coverage and signal-to-noise ratio.* ***e*** *Recurrent DE genes in glutamatergic neurons, using AUROC as*
145 *a proxy for coverage and fold change of detection rate as a proxy for signal-to-noise ratio. Gene names and lines highlight*
146 *the Pareto front of markers, which offer optimal trade-off between signal-to-noise and coverage.* ***f*** *Pareto fronts for*
147 *neuronal classes (glutamatergic neurons, GABAergic neurons and non-neuronal cells) in the coverage/signal-to-noise space.*
148 *We subdivide markers as perfect (high coverage and signal-to-noise), specific, sensitive, or weak (low coverage and signal-*
149 *to-noise).* ***g*** *Illustration of sensitive (high target expression, some background expression), perfect (high target expression,*
150 *no background expression) and specific (low target expression, no background expression) markers along the glutamatergic*
151 *Pareto front.*

152 No individual marker offers high coverage and signal-to-noise ratio

153 Having established that DE statistics are replicable in aggregate, we assessed a range of existing
154 statistics and found they strongly clustered into two groups, corresponding to definitions for
155 coverage and signal-to-noise ratio (Fig. 1c,d). The first block of statistics contained average gene
156 expression and intuitively mapped to the notion of coverage; it also included the DE p-value and the
157 detection rate, which are strongly indicative of genes that are broadly expressed. The second block
158 contained the fold change and the fold change of detection rate and intuitively mapped to the
159 notion of signal-to-noise ratio. The lack of correlation between the two blocks indicates that there is
160 trade-off, genes have a “choice” between favoring coverage or signal-to-noise ratio. Note that this is

161 broadly consistent with long-standing heuristic practice of considering both p-value and fold change
162 in bulk DE through volcano plots (Cui and Churchill 2003; Goedhart and Luijsterburg 2020). In the
163 following, we use the area under the receiver-operator characteristic curve (AUROC) as our proxy for
164 coverage (as used in Seurat's ROC test (Stuart et al. 2019) or LIGER's marker detection (Welch et al.
165 2019; Liu et al. 2020)), fold change of the detection rate (FCd) as our proxy for signal-to-noise when
166 we consider individual markers (as used in M3Drop (Andrews and Hemberg 2019)), and fold change
167 (FC) as our proxy for signal-to-noise when we consider marker lists (as used in the traditional
168 Volcano plot (Cui and Churchill 2003)).

169 In a FC/AUROC representation, genes offering a trade-off from best signal-to-noise marker to highest
170 coverage marker form a Pareto front of markers (Fig. 1e). The Pareto front representation offers a
171 rapid visualization of the strength of markers that can be associated with any given cell type. Based
172 on our exploration of the datasets, we subdivided markers as perfect (high coverage, AUROC > 0.8,
173 high signal-to-noise, FCd > 8), specific (high signal-to-noise), sensitive (high coverage) or weak (DE,
174 but low coverage and low signal-to-noise). As expected, the Pareto fronts associated with
175 Glutamatergic and GABAergic cells contain perfect markers (Fig. 1f) that identify these populations
176 with high confidence across all technologies sampled, such as *Gad1* for GABAergic cells and *Neurod2*
177 for Glutamatergic cells. In contrast, there is no perfect marker for non-neuronal cells: their Pareto
178 front only includes highly sensitive markers such as *Qk* (highly expressed in non-neurons but also
179 expressed in neurons) and highly specific markers such as the *Slco1c1* transporter (high signal-to-
180 noise, but not covering all non-neurons), consistent with the heterogeneous nature of non-neurons
181 and the need to use several markers in conjunction (Fig. 1f). Remarkably, the Pareto fronts were
182 composed of perfectly recurring genes, i.e. genes that are reliably DE across all datasets (Fig. 1e, FC >
183 4, FDR < 0.05). Conversely, this implies that markers selected based on recurrence (number of
184 datasets where they are reliably DE) naturally range from highly sensitive to highly specific. In
185 contrast, high AUROC markers have high sensitivity but low specificity.

186 To illustrate that the chosen statistics and thresholds offer an intuitive understanding of coverage
187 and signal-to-noise ratio, we plotted the expression of markers along the Glutamatergic Pareto front
188 in one of the BICCN datasets (Fig. 1g). Highly sensitive markers (*Arpp21* and *Sv2b*, AUROC > 0.8) are
189 expressed in all Glutamatergic cells at high levels but are also expressed in background cells (e.g.,
190 high expression of *Arpp21* in the *Vip*, *Sncg* and *Lamp5* cell types). The highly specific marker (*Exph5*,
191 FCd > 8) is expressed almost exclusively in glutamatergic cells, but not in all cells, indicating high
192 drop-out propensity or cell-type specific expression (e.g., it is almost not expressed in L5/6 NP).
193 Finally, the perfect markers (*Satb2* and *Neurod2*) cover almost all cells of interest and have very
194 limited background expression. To further investigate if our simple metrics capture known marker
195 genes, we investigated the Pareto fronts of inhibitory subclasses as defined by the BICCN. We found
196 that all classical markers were on the Pareto front (Sup. Fig. 1h), classified as perfect markers (*Pvalb*,
197 *Lamp5*) or highly sensitive markers (*Sst*, *Vip*), with the notable exception of *Sncg*, which was only
198 imperfectly detected in most datasets (low coverage, high signal-to-noise). A look at the Pareto front
199 of the *Sncg* population suggests that multiple genes would offer better coverage than *Sncg* while
200 preserving a high signal-to-noise ratio, in particular *Cadps2*, *Frem1* and *Megf10* (Sup. Fig. 1i), but
201 that all markers tend to have some background expression in the *Vip Serpinf1* cell type. For
202 Glutamatergic subclasses, the Pareto fronts suggested that all subclasses have perfect markers,
203 except for IT subclasses, consistent with previous observations of gradient-like properties (Tasic et
204 al. 2018; Yao et al. 2020a, 2020b) (Sup. Fig. 1j,k).

205 The FC/AUROC plot rapidly informs about the maximal strength of markers that can be expected for
206 any given cell type. In contrast to the Volcano plot, which is based on one effect size and one
207 significance statistic (Goedhart and Luijsterburg 2020), it relies on two effect sizes. Because we
208 obtain replicable statistics by combining values over multiple datasets, we remove the need to
209 visualize significance and obtain a plot with two interpretable dimensions of marker strength: signal-
210 to-noise ratio and coverage. Typically, for each population, we suggest building the FC/AUROC plot
211 across at least 5 datasets, identify genes on or next to the Pareto front, visualize their expression

212 across datasets to appreciate the optimal coverage/signal-to-noise trade-off, then select the best
213 marker(s) for the application at hand.

214 The strength of individual markers decreases with finer cell type resolutions

215 The BICCN defined three levels of cell types: classes (such as glutamatergic neurons), subclasses
216 (such as PV+ interneurons), and clusters (such as Chandelier cells)(Fig. 2a). While classes and
217 subclasses had been previously experimentally characterized and showed strong statistical
218 robustness across datasets, clusters obtained from independent datasets were more elusive (Yao et
219 al. 2020a). To further characterize how distinct cell types are, we evaluated the number of replicable
220 markers with increasing clustering resolution. We controlled for the increasing number of cell types
221 by using a hierarchical approach, for example we compare a cluster to clusters from the same
222 subclass only (Fig. 2a).

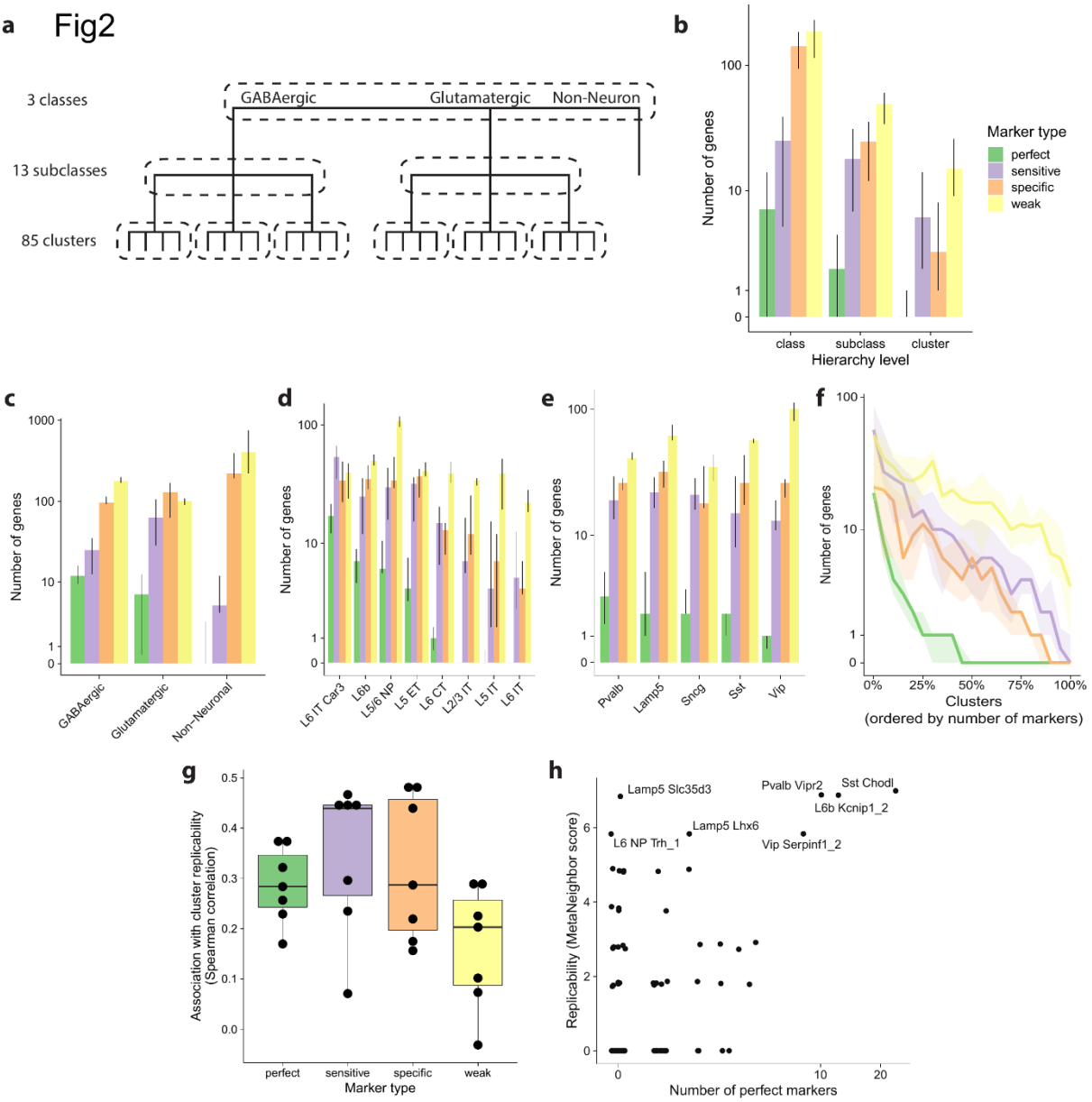


Figure 2. Markers are associated with higher cluster replicability, but become rare at finer resolutions. ***a*** *Schematic of the BICCN taxonomy. Markers are selected hierarchically: each cluster is only compared to its direct neighbors in the hierarchy (dashed lines).* ***b*** *Number of reliable markers (FC>4, FDR<0.05) along the BICCN cell type hierarchy, according to marker type: perfect (AUROC > 0.8 and FCd > 8), specific (AUROC > 0.8), sensitive (FCd > 8) and weak (FDR < 0.05).* ***c*** *Number of markers of each type for BICCN classes, error bars are interquartile range across datasets.* ***d*** *Same as c for Glutamatergic subclasses.* ***e*** *Same as c for GABAergic subclasses.* ***f*** *Number of markers of each type for BICCN clusters, with cell types ordered according to number of markers. Ribbons indicate interquartile range across datasets.* ***g*** *Association between number of makers and cross-dataset MetaNeighbor replicability score at the cluster level (Spearman correlation, one dot per BICCN dataset).* ***h*** *Illustration of association of replicability (MetaNeighbor score) and number of markers in the scCv2 dataset.*

To investigate how the number and quality of markers depends on the cell type hierarchy, we extracted all reliable markers (FC>4, FDR<0.05) and classified them as perfect (AUROC > 0.8 and FCd > 8), specific (AUROC > 0.8), sensitive (FCd > 8) and weak (FDR < 0.05). We observed an overall decrease in the median number of markers when going from coarse to finer resolution (397 total

238 markers at the class level, 108 at the subclass level, 35 at the cluster level), confirming that the signal
239 that separates neighboring populations becomes increasingly weaker (Fig. 2b). We found that all
240 classes and subclasses had at least one perfect marker except for non-neurons and IT subclasses (Fig.
241 2b-e). In contrast, only around 50% of clusters had a perfect marker (Fig. 2f, Sup. Fig. 2a-d). This
242 proportion dropped to 25% with the additional requirement that the marker should be robust across
243 all technologies (Sup. Fig. 2a). Strikingly, a handful of clusters had extremely strong support, totaling
244 close to 50 perfect markers in some of the datasets. Upon closer investigation, these clusters
245 corresponded to experimentally identified populations, such as the long-projecting interneurons
246 (Tasic et al. 2018; Paul et al. 2017)(*Sst Chodl*, up to 43 perfect markers) or Chandelier cells (Paul et al.
247 2017; Tasic et al. 2018)(*Pvalb Vipr2*, up to 20 perfect markers), suggesting that for these cell types,
248 experimentally characterized differences in morphology and physiology are reflected by a high
249 number of marker genes. Reassuringly, almost all clusters had at least one specific marker,
250 suggesting the presence of unique characteristics (Fig. 2f, Sup. Fig. 2b).

251 While more data are needed to experimentally validate cell types, we wondered whether the
252 number of markers would be predictive of computational replicability. Intuitively, a higher number
253 of markers indicates unique aspects in a population's transcriptional program, which should increase
254 its identifiability across datasets. We assessed cluster replicability using MetaNeighbor, which tests
255 the consistency of cell types across datasets using a neighbor voting framework: intuitively, if two
256 clusters represent the same cell type, they will preferentially vote for each other (see Methods). We
257 found that cluster replicability was indeed associated with the number of markers (rho=0.4, Fig 2g).
258 To understand why replicability and number of markers are only partially associated, we further
259 investigated the relationship in the best-powered datasets. We noted that, while a high number of
260 markers was associated with higher replicability, a low number of markers did not imply low
261 replicability (Fig 2h). Some clusters, such as *Lamp5 Slc35d3*, are found independently in all BICCN
262 datasets and match with high statistical confidence (MetaNeighbor replicability > 0.7), despite the
263 absence of strong markers. However, we noted that these clusters usually had a high number of

264 specific markers (Sup. Fig. 2e). Conversely, we found some instances of clusters with low replicability
265 and high number of markers (e.g. *Pvalb Nkx2.1*, Sup. Fig. 2f) but, upon further investigation, all
266 identified "markers" were stress-related genes likely to be artefacts of the extraction protocol.
267 Overall, the imperfect association of markers and replicability suggests that individual markers only
268 provide a partial view of cell type identity, which is encoded broadly across the transcriptome.

269 Meta-marker aggregation enables near-optimal cell type descriptions

270 Our previous results suggest that, at the finest level of resolution, single markers are not sufficient to
271 unambiguously identify cell types (only ~10 genes with AUROC > 0.8 at the subclass level, Fig. 3a).
272 These results are consistent with the ideas that markers are affected by dropout and that clustering
273 procedures capture information from the full transcriptome. We next tested if cell type identity can
274 be efficiently characterized by redundant marker lists. In particular, we ask how many markers
275 contribute to make cell types more unique, and how the selection of replicable markers improves
276 cell type characterization.

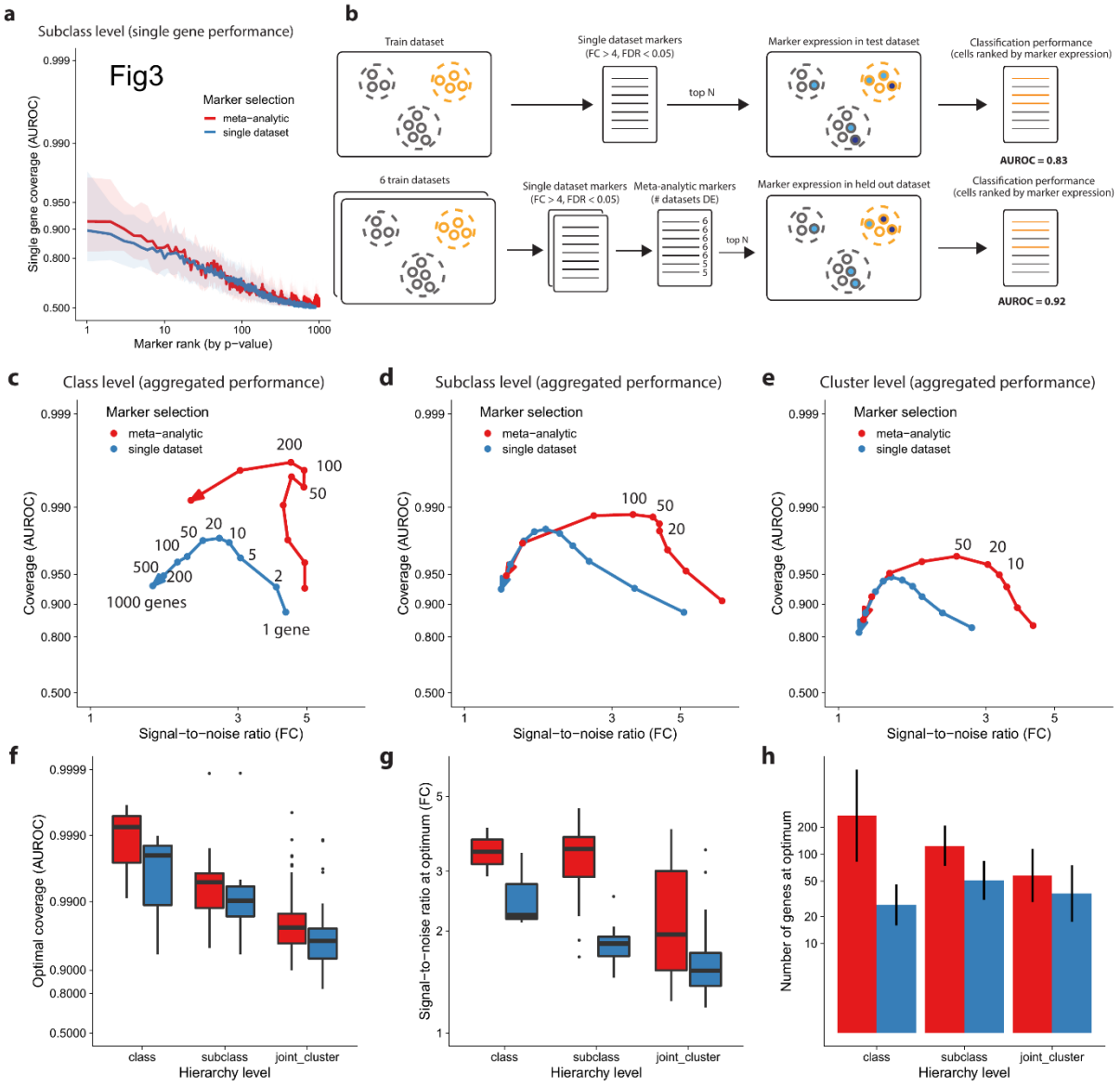


Figure 3. Meta-analytic aggregation of markers considerably improves the coverage/signal-to-noise trade-off. a *Cell type classification performance of single markers as a function of marker rank.* ***b*** *Schematic of classification task using cross-dataset validation for markers from single datasets and meta-analytic markers.* ***c*** *Cell type classification performance of aggregated markers (average expression) with increasing number of markers. Performance is plotted as a parametric curve in a coverage (AUROC), signal-to-noise (fold change) space similar to Figure 1. The arrow points toward an increasing number of markers, the numbers next to the dots show the number of genes at which performance was measured (shown in full as an example for one arrow, otherwise highlighting optimal performance).* ***d*** *Same as c at the subclass level.* ***e*** *Same as b at the cluster level.* ***f-h*** *Coverage (f), signal-to-noise ratio (g) and average number of genes (h) at optimal coverage as a function of hierarchy level. Boxplots show median and interquartile range (f,g), bar plot shows mean and standard deviation (h) across datasets. In all panels, colors indicate whether markers were prioritized according to a single dataset or using the meta-analytic approach.*

To study how the number of markers affects cell type identifiability, we framed gene aggregation as a classification task (Fig. 3b). How well can we predict cell type identity for the average expression of an increasing number of markers? We first created ranked marker lists for each dataset by ranking genes according to their AUROC. To test the effect of meta-analytic marker selection, we used cross-

293 dataset validation: we computed marker replicability across 6 datasets and predicted cell types on
294 the held-out dataset. To rank meta-analytic markers, we used two criteria: first, the number of
295 datasets in which they were reliable DE (FC>4, FDR<0.05), second, the average AUROC. To predict
296 cells that belong to a given cell type, we ranked cells based on the average expression of the top N
297 markers for that cell type. We visualized performance in the FC/AUROC space, displaying
298 classification results as a trade-off between coverage (AUROC) and signal-to-noise ratio (FC). We
299 found that marker aggregation improved cell type identification at all levels of the hierarchy,
300 independently of the marker prioritization strategy (Fig. 3c-e). Coverage reached an optimum
301 between 10 and 200 genes (Fig. 3c-e), at the cost of a slightly lower signal-to-noise ratio (class, FC=6
302 to 6, subclass, 6 to 5, cluster, 5 to 3). Optimal performance was reached between 50 and 200 genes
303 for classes, 20 to 100 genes for subclasses, and 10 to 50 genes for clusters.

304 Meta-analytic markers systematically outperformed single dataset marker genes in terms of
305 coverage (Fig. 3f, class, AUROC=0.92 to 0.99, subclass, 0.9 to 0.99, cluster 0.85 to 0.95), signal-to-
306 noise ratio (Fig. 3g) and number of relevant genes (Fig. 3h). In other words, the best candidates in a
307 single dataset by a single metric are "too good to be true". The gain in signal-to-noise ratio is
308 particularly apparent at the cluster and subclass levels (Fig 3d-e), suggesting that the meta-analytic
309 approach successfully extracts and combines lowly expressed markers. We checked that all results
310 were robust to another marker prioritization strategy, where we ranked genes by fold change
311 instead of AUROC (Sup. Fig. 3a-c).

312 We further investigated how the performance was distributed within hierarchy levels and across
313 datasets (Sup. Fig. 3d-o). The overall classification performance (AUROC) increased with dataset
314 depth (Sup. Fig. 3d). More surprisingly, the signal-to-noise ratio was approximately constant across
315 datasets (Sup. Fig. 3h) and the number of relevant markers was slightly lower for high depth
316 datasets (Sup. Fig. 3l). Classification performance was high for all classes and subclasses (median
317 AUROC > 0.99, median FC > 3, Sup. Fig. 3e,f,i,j), with the notable exception of L5 IT and L6 IT (AUROC

318 < 0.99, FC < 3). The classification performance had a wide variance at the cluster level (AUROC
319 ranging from 0.9 to 1, FC ranging from 1.5 to 8, Sup. Fig. 3g,k), 32/85 cell types had a low signal-to-
320 noise ratio (median FC < 2, Sup. Fig. 3g). Finally, we found that the ideal number of markers ranged
321 from 10 to 200 and was remarkably consistent within each hierarchy level (Sup. Fig. 3l-o).

322 Meta-marker enrichment is robust across datasets

323 Automatic annotation of cell types typically involves two steps: (a) prioritize cells that are most likely
324 to belong to a given cell type, (b) annotate cells that exceed a pre-specified threshold condition. The
325 threshold indicates that there is enough evidence to proceed with the annotation, for example
326 preventing misannotation when a cell type is missing in the reference dataset. In the previous
327 assessment, we showed that meta-analytic marker lists successfully prioritize cells, without explicit
328 consideration for correct thresholding. We wondered whether marker expression was sufficiently
329 consistent to be compatible with a simple thresholding method: a cell belongs to a given cell type
330 when its marker expression exceeds the same pre-specified value for each test dataset.

331 For each dataset in the compendium, we computed the annotation performance at various
332 threshold values (Fig. 4a). For example, in the *Pvalb* subclass, meta-analytic markers had a high
333 maximal performance (F1opt>0.9) across all datasets (Fig. 4c). Additionally, the maximal
334 performance had a distinctive plateau, indicating that a large range of thresholds had almost
335 equivalent performance, as expected from the meta-markers' tendency to preserve a high signal-to-
336 noise ratio. To visualize how well optimal thresholds aligned across datasets, we defined the
337 plateauing region as the thresholds that had at least 90% of the maximal performance (Fig. 4b).
338 While there was a large plateau in all datasets, the plateaus did not align well, suggesting
339 normalization issues (Fig. 4c). As a result, a meta-analytic threshold leads to good performance in
340 most datasets, but fails in dataset with extreme properties, such as snCv2 (nuclei, 10X v2, low depth)
341 or scCv3 (cells, 10X v3, high depth).

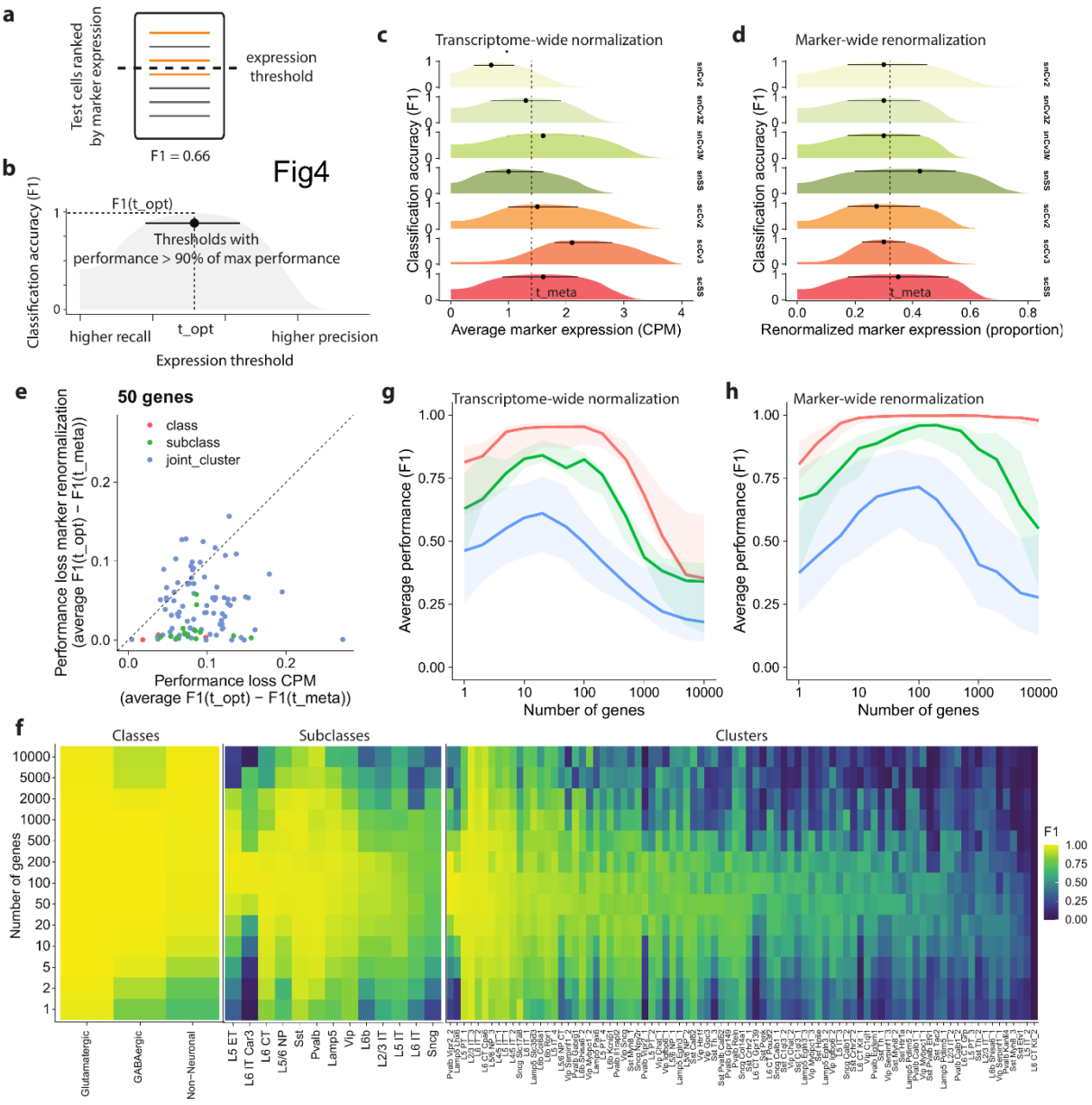


Figure 4. Aggregated expression of meta-analytic markers enables robust identification of cell types. a *Schematic of threshold-based classification. The initial steps of the procedure are identical to Figure 3.* ***b*** *Illustration of statistics measuring cell type annotation performance at various thresholds. For a given dataset, cell type and number of genes (scSS dataset, Pvalb subclass, 100 genes in the illustration), F1opt is the score obtained by picking the single best threshold, indicated by a dot. The line indicates the range of near-optimal thresholds (leading to a performance higher than 0.9*F1opt).* ***c*** *Comparison of near-optimal expression thresholds across datasets (for Pvalb subclass and 100 genes). The position of the dotted line (F1meta) is obtained by averaging optimal expression thresholds across datasets.* ***d*** *Similar to b, but defining optimal thresholds based on proportion of marker expression instead of expression.* ***e*** *For each cell type, we show how much performance is lost by switching from a dataset-specific threshold (F1opt) to a single meta-analytic threshold (F1meta) for the two types of thresholds (CPM expression, marker expression proportion). Colors indicate hierarchy level, the dashed line is the identity line (performance loss is identical for the two types of thresholds).* ***f*** *For each hierarchy level, heatmap detailing classification performance for each cell type as a function of the number of genes.* ***g-h*** *Average performance as a function of the number of genes using the optimal meta-analytic expression threshold (g) or optimal marker expression proportion threshold (h). Ribbons show interquartile range across populations and test datasets.*

To overcome the normalization discordance, we reasoned that the normalization issues are mainly driven by non-marker genes. Instead of considering marker expression for each cell type

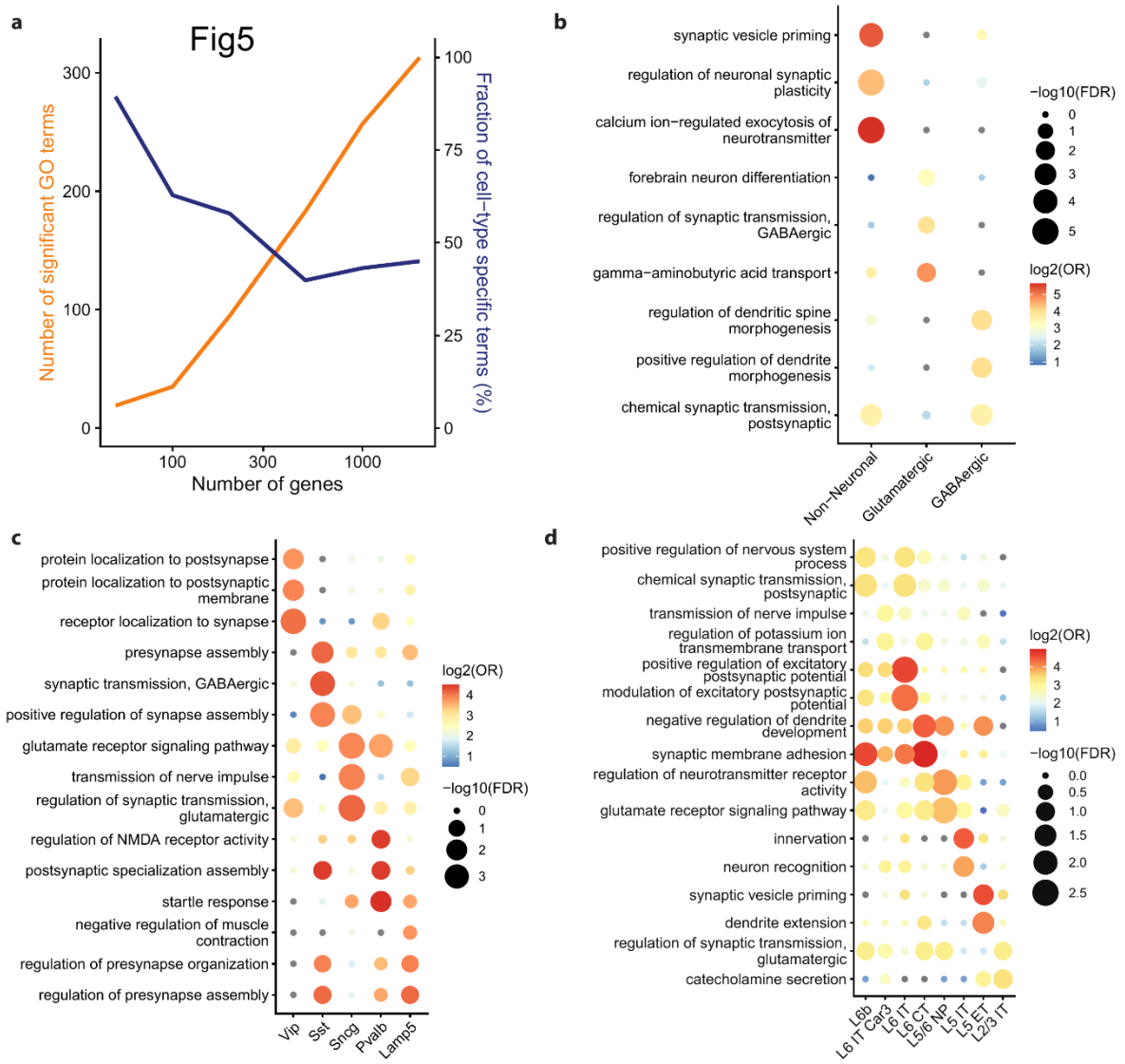
359 independently, we divided marker expression by the total marker expression (across all putative cell
360 types). After this change, plateaus of optimal thresholds aligned across all datasets (Fig. 4d),
361 suggesting that marker lists have preserved relative contributions in all datasets. To assess the utility
362 of marker-wide renormalization, we directly compare how much performance is lost by switching
363 from dataset-specific thresholds (optimal threshold per dataset) to a consensus threshold. The
364 decrease in performance was systematically lower with marker-wide renormalization for classes and
365 subclasses and was generally lower for clusters (Fig. 4e).

366 We compared the performance achieved for transcriptome-wide and marker-wide normalization as
367 a function of the number of markers (Fig. 4g-h, Sup. Fig. 4a-b) and within each hierarchical level (Fig.
368 4f, Sup. Fig. 4c-e). Both methods reached high classification performance at the class and subclass
369 level (optimal average F1 > 0.75, Fig. 4g-h), but the average performance was considerably lower at
370 the cluster level. Marker-wide normalization yielded substantially higher classification performance
371 (ΔF1 ~ 0.1) and reached peak performance by successfully integrating a higher number of genes (50-
372 500 markers, Fig. 4g-h). Performance was distributed unequally within hierarchy level, in particular
373 for subclasses and clusters (Fig. 4f). Almost all subclasses reached optimal performance around 100-
374 200 markers with a high performance (F1 > 0.75), with the exception of L5 IT, L6 IT and *Sncg*. At the
375 cluster level, the performance degraded substantially: peak performance was attained around 50-
376 100 markers, with only 43/85 of cell types reaching high performance (F1 > 0.75). All these trends
377 were consistent with results obtained using transcriptome-wide normalization, with overall higher
378 annotation performance (Sup. Fig. 4c-e).

379 Meta-markers are enriched for genes involved in synaptic regulation and 380 development

381 We next wondered if top meta-markers were enriched for specific biological processes. We
382 performed gene set enrichment analysis for the top markers in each cell type against Gene Ontology

383 (GO) terms from the Biological Process (BP) ontology. To focus on specific processes, we only
384 queried terms containing between 20 and 100 genes. For each cell type, we extracted the top 3
385 enriched terms based on the False Discovery Rate (FDR) from the hypergeometric test. The best
386 balance between number of enriched terms and cell type specificity was achieved for the top 100
387 markers for both classes and subclasses (Fig. 5a, Sup. Fig. 5a).



388

389 ***Figure 5. The top 100 meta-markers are enriched for specific synaptic processes. a*** *Total number of significantly enriched*
390 *GO terms (orange) and fraction of significant GO terms that are enriched in a unique cell type (blue) for BICCN subclasses*
391 *when an increasing number of meta-markers are considered.* ***b*** *Top 3 enriched Gene Ontology (GO) terms for the top 100*
392 *meta-markers for each BICCN class. For each dot, the size reflects the False Discovery Rate (FDR), the color reflects the Odds*
393 *Ratio (OR) of the enrichment test (hypergeometric test).* ***c*** *Same as* ***b*** *for the top 100 meta-markers for BICCN GABAergic*
394 *subclasses.* ***d*** *Same as* ***b*** *for the top 100 meta-markers for BICCN Glutamatergic subclasses (only top 2 terms per cell type*
395 *are shown).*

396 We found strong enrichment for synaptic properties for all cell types. At the class level, non-
397 neuronal markers were enriched for synaptic support functions, such as "Regulation of neuronal
398 synaptic plasticity" (Fig. 5b). Glutamatergic neurons were enriched for synaptic regulation (such as
399 the regulation of GABAergic transport), while GABAergic neurons were enriched for gene sets
400 related to the regulation of spine and dendrite formation. At the subclass level (Fig. 5c,d), GABAergic
401 neurons were most distinguishable based on synaptic sub-properties, such as localization to synapse
402 (*Vip*), synapse assembly (*Sst*, *Lamp5*), or glutamate transmission regulation (*Sncg*). Glutamatergic
403 subclasses showed a similar enrichment of synaptic sub-properties, including various aspects of
404 potential regulation and synaptic transmission (L6b, L6 IT, L2/3 IT, L5/6 NP), as well as synaptic
405 development (L6 CT, L5 IT, L5 ET). We further confirmed that these findings were consistent with the
406 enrichment of the top 200 markers, which also highlighted gene sets involved in synaptic regulation
407 and development (Sup. Fig. 5b-d). These results suggest that meta-markers define a plausible
408 biological subspace revealing cell type differences in terms of synaptic properties.

409 Meta-markers improve deconvolution performance

410 One of the primary purposes to which cell atlas data may eventually be put is deconvolution of bulk
411 data where cell composition is likely related to the condition of interest (e.g., disease). Single cell
412 data have been routinely used to increase deconvolution performance in recently developed tools
413 (Tsoucas et al. 2019; Wang et al. 2019; Newman et al. 2019; Dong et al.), but performance remains
414 plagued by batch effects and cell type similarity (Newman et al. 2019; Huang et al. 2020; Cobos et al.
415 2020). The role of marker genes in deconvolution remains particularly unclear: a recent benchmark
416 suggests that the quality of markers is more important than the deconvolution method (Cobos et al.
417 2020), in most studies the influence of the number of markers is only partially assessed (Newman et
418 al. 2019; Hunt and Gagnon-Bartsch 2019). Our annotation assessment suggested that cell types are
419 best captured with 10 to 200 meta-analytic markers; deconvolution is a natural place to test this
420 heuristic.

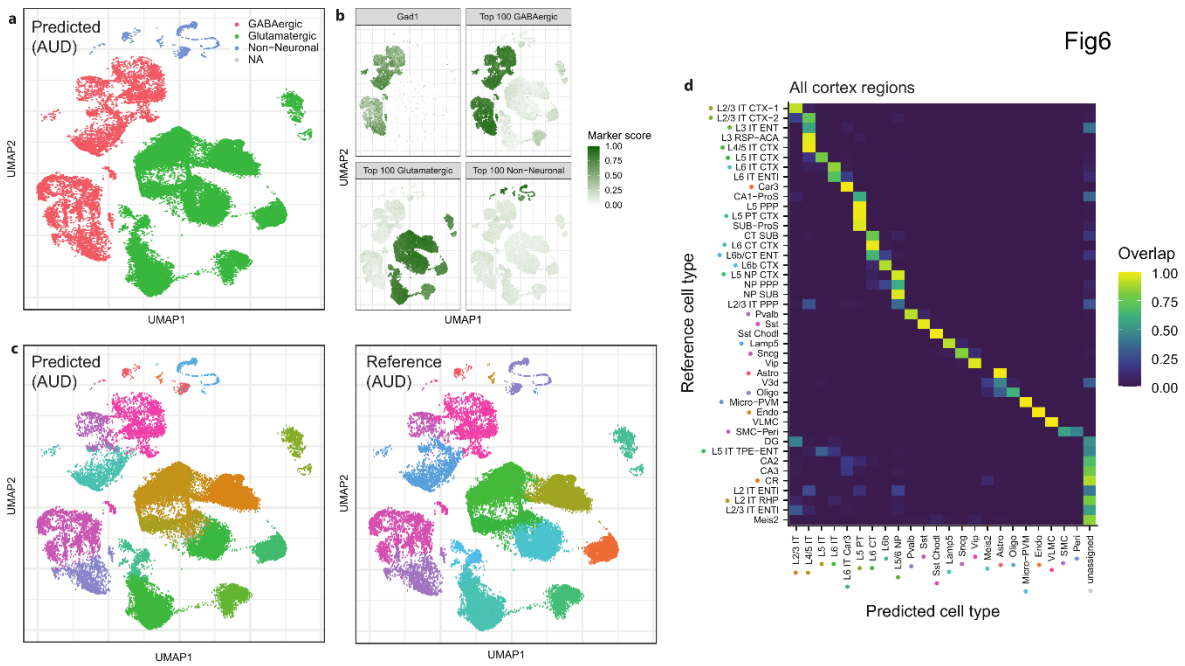
421 To measure the number of genes that yield maximal deconvolution performance, we generated
422 thousands of pseudo-bulk datasets with known mixing proportions from each of the BICCN datasets
423 (Sup. Fig. 6a). As in previous experiments, we directly compared the performance of markers
424 extracted from single datasets and performance of meta-analytic markers. We initially compared
425 two tasks: (a) within-dataset cross-validation, where cell type profiles are learned from a training
426 fold and tested on a held-out set from the same dataset, (b) cross-dataset-validation, where profiles
427 are learned on one dataset and tested in another dataset. Within-dataset cross-validation proved to
428 be a simple task, yielding extremely high performance (Pearson ~ 1, not shown). In contrast, cross-
429 dataset-validation showed only modest performance (Pearson ranging from 0 to 1, Sup. Fig. 6b),
430 highlighting the difficulty of the deconvolution task. Because deconvolution applications typically
431 involve different datasets, we focused our analyses on cross-dataset validation.

432 Deconvolution performance rapidly degraded along the neuron hierarchy (Sup. Fig. 6b), ranging
433 from almost perfect performance for classes (Pearson ~ 1) to low performance for clusters (Pearson
434 < 0.5). Classes were easily learnable across all tasks (Sup. Fig. 6c), even using random genes,
435 suggesting that at this level of the hierarchy, cell types are strongly uncorrelated and can easily be
436 separated along the first principal component. At the subclass level, the performance of random
437 genes remained close to 0, suggesting stronger covariation compared to classes (Sup. Fig. 6d). Meta-
438 analytic markers yielded more robust deconvolution performance, with performance increasing up
439 to 100 genes, while markers from single datasets prioritized only around 10 informative genes. The
440 trend was similar at the cluster level, but with lower overall performance (Sup. Fig. 6e). Meta-
441 analytic markers again proved to be more robust, prioritizing around 50 informative genes compared
442 to 10-20 from single datasets. Overall, our results suggest that, in conjunction with batch effects, the
443 increasing covariation of cell types at finer resolution significantly complicates the deconvolution
444 task. The prioritization of a large number of robust marker genes is an important first step towards
445 successful deconvolution.

446 ## Meta-markers reveal a generalizable description of cell types

447 We have previously shown that meta-marker signatures generalize across laboratories and
448 technologies. We next asked how well they generalize across the cortex by predicting cell types in a
449 BICCN dataset combining multiple cortical and hippocampal brain regions (Yao et al. 2020b). To
450 understand how easily meta-markers generalize, we used a straightforward annotation method:
451 assign cells to the cell type with the highest average meta-marker expression. For simplicity, we
452 considered the same number of meta-markers for all cell types: 100 at the class and subclass level,
453 50 at the cluster level.

454 We started by predicting cell types in the auditory cortex sub-dataset, containing 71,670 cells
455 annotated to 203 cell types. We chose to focus on the auditory cortex because of its large number of
456 cells, and to investigate the generalizability of cell types derived from a motor area (MOp) to a
457 sensory area. While inhibitory cell types are expected to generalize, excitatory cell types have been
458 shown to have divergent patterns (Tasic et al. 2018). At the class level, top meta-markers enabled
459 perfect classification down to every single cell (Fig. 6a). Assignments can be traced back to meta-
460 marker scores, as well as individual genes (Fig. 6b). Consistent with our previous points, the
461 GABAergic score is uniformly high across all cells labeled as GABAergic. In contrast, the expression of
462 the single best marker, *Gad1*, is more variable in GABAergic cells and displays sporadic expression in
463 non-GABAergic cells.



464

465 ***Figure 6. Meta-analytic markers from the primary motor cortex (MOp) generalize to other cortical regions. a*** *Example of*
466 *class-level predictions in the auditory cortex (AUD), where cells are embedded in UMAP space and colored according to*
467 *predictions based on the top 100 meta-markers for MOp classes. Cells are assessed independently and remain unassigned*
468 *(NA) if the marker enrichment score is lower than 1.5 for all classes.* ***b*** *Marker scores (re-normalized between 0 and 1) used*
469 *to determine cell type assignments. The first subpanel shows the score obtained from a single GABAergic marker, the three*
470 *other panels show the scores obtained by combining the top 100 meta-markers for each class.* ***c*** *Subclass-level predictions in*
471 *the auditory cortex based on the top 100 MOp meta-markers (left) and reference labels (right). Cells remain unassigned*
472 *(NA) if the marker enrichment score is lower than 1.5 for all subclasses. See panel* ***d*** *for color legend (some reference cell*
473 *types are absent from AUD).* ***d*** *Confusion matrix showing the concordance of subclass-level predictions based on the top*
474 *100 meta-markers with reference cell types across 40 brain areas. Cells are unassigned if the marker enrichment is lower*
475 *than 1.5 for all subclasses.*

476 Remarkably, at the subclass level, meta-markers enabled similarly strong cell type assignments, as
477 suggested by the uniform coloring of clusters in UMAP space (Fig. 6c). Note that the assignments
478 occur in each cell independently, without knowledge about clusters or expression profiles of
479 neighboring cells, highlighting the consistency of meta-marker expression. This procedure allowed
480 the identification of rare cell types, even when only one or two cells were present in the dataset (e.g.
481 smooth muscle cells and pericytes, Sup. Fig. 7b). The predicted assignments corresponded almost
482 perfectly to the reference cell types (Fig. 6c). The main exception were deep layers IT cell types, in
483 particular one group of L5 IT cells tended to be assigned as L4/5 IT or L6 IT (Sup. Fig. 7c). Finally,
484 cluster-level predictions also proved extraordinarily consistent, with smooth transitions between cell
485 types that mapped with auditory cortex reference annotations (Sup. Fig. 8).

486 To further highlight high-quality predictions, we quantified assignment confidence using meta-
487 marker enrichment (observed expression over expected expression) as a “QC” metric. In the
488 auditory cortex, we found that a threshold of 1.5 offered an optimal trade-off between annotation
489 recall and precision (Sup. Fig. 9a-c). Raising the threshold to 2 further selected high-confidence calls,
490 yielding higher precision for slightly lower recall. Interestingly, cells that became unassigned were
491 mostly located in regions where predictions and reference disagreed: deep IT layers, and inhibitory
492 neurons bridging medial ganglionic eminence (MGE) and caudal ganglionic eminence (CGE)
493 subclasses (Sup. Fig. 7a). Meta-marker enrichment thus offers a good proxy for prediction quality,
494 enabling to identify cells with a high-confidence cell type assignment.

495 Next, we systematically quantified the agreement of meta-marker based predictions and reference
496 annotations across all brain regions and 43 consensus subclasses. We found exceptionally good
497 agreement, with all reference subclasses mapping to exactly one predicted MOp subclass (Fig. 6d).
498 All MOp subclasses matched strongly with their “natural” counterparts in the reference dataset,
499 such as “L2/3 IT” with “L2/3 IT CTX-1”. Remarkably, reference cell types absent in MOp (such as
500 hippocampal cell types) were mostly labeled as “unassigned”, suggesting that meta-marker
501 signatures correctly avoid labeling unseen cell types. This trend became particularly obvious for cells
502 with marker enrichment > 2 (Sup. Fig. 9d), where all unseen cell types became “unassigned”, while
503 conserving high matching scores between shared cell types.

504 Discussion

505 By assessing marker replicability across 7 datasets, we selected robust markers and identified the
506 optimal number of markers to define a cell type. We identified highly replicable markers for 85 cell
507 types from the mouse primary motor cortex (Sup. Data 1-3). This meta-analytic strategy proved
508 particularly important for rare populations and lowly expressing genes (Fig. 1). Compared to
509 previous efforts (Tasic et al. 2018; Mancarci et al. 2017; Yao et al. 2020a), we identified a high

510 number of robust markers at high cell type resolution: at the BICCN cluster level, cell types were best
511 characterized by 10-200 meta-analytic markers, a two-fold increase of reliable markers compared to
512 markers selected from single datasets (Fig. 3). Interestingly, we found that only 50% of clusters had
513 strong markers (Fig. 2), but that some of the clusters lacking strong markers (e.g. *Lamp5 Slc35d3*)
514 were consistently identified in all BICCN datasets, suggesting broad encoding of their identity and
515 highlighting the need of extended marker signatures.

516 We found that the simple aggregation of marker expression enabled the annotation of individual
517 cells (Fig. 4), suggesting that careful feature selection is enough to provide a rough definition of cell
518 types. Remarkably, marker lists derived from a single cortical region generalized with high accuracy
519 to other cortical regions without any methodological fine-tuning (Fig. 5). By introducing redundant
520 information about cell types, meta-analytic markers dramatically increased cell type separability (Fig.
521 3). However, adding more markers is only beneficial if they are cell type-specific. As a result, we
522 established that the ideal number of markers decreases with cell type resolution: 200 genes to
523 separate classes (lowest resolution, e.g. GABAergic neurons), 100 genes for subclasses (e.g., *Pvalb*
524 interneurons) and 50 genes for clusters (highest resolution, e.g., Chandelier cells).

525 By combining datasets that were generated using different technologies, the markers we propose
526 are likely to generalize well with respect to this axis of variation. Moreover, we show that our
527 marker descriptions generalize to other cortical regions, despite all "training" datasets sampling
528 from the same cortical region. However, the data used in this study were obtained from adult mice
529 with limited genetic background and grown in lab conditions. As a result, it remains unclear how well
530 the marker descriptions would generalize across development or biological conditions. On the other
531 hand, as our approach relies on a simple procedure, marker lists can easily be extended to
532 incorporate new sources of variation, such as additional brain regions, species or biological
533 conditions. On a similar note, markers depend on one particular annotation effort, but we can
534 expect the neuron taxonomy to evolve with additional data, in particular the fine-resolution clusters.

535 Our framework, available as an R package, allows to rapidly evaluate the consistency of marker
536 expression for new cell type annotations.

537 To highlight the replicability of marker descriptions, the manuscript relies on simple methods, but
538 marker lists can easily be combined with more sophisticated methods for marker selection or cell
539 type assignment. For example, experimental applications routinely require either a few specific
540 markers to target one cell type (Huang 2014) or a panel of hundreds of markers to jointly separate
541 all cell types (Moffitt et al. 2018). Marker lists can be combined with methods to select concise sets
542 of markers (Asp et al. 2019; Zhang et al. 2020; Dumitrascu et al. 2019) by filtering candidates that
543 are likely to generalize. Similarly, development studies (Hobert 2008; Huang 2014; Kessaris et al.
544 2014; Lodato and Arlotta 2015; Mayer et al. 2018; Tosches et al. 2018) indicate that neural lineages
545 are marked by the specific onset and offset of key transcription factors (TFs), but the expression of
546 these key TFs may not be maintained at later stages or only at low levels. Since our approach is
547 powered to identify lowly expressed markers, it can be combined with time series data to help
548 identify replicable lineage-specific genes.

549 This study focused on the neuron hierarchy, but our strategy generalizes to other tissues. In order to
550 encourage broader adoption, we have made our code available as a package and in the vignette we
551 show how our analyses and guidelines can be similarly applied to a pancreas compendium. We
552 chose to focus on the BICCN dataset because of its complexity (85 neuronal cell types),
553 comprehensiveness (~500,000 cells with latest sequencing technologies) and diversity (6
554 technologies used). Our results suggest that, in the brain, there is a clear separation at the top two
555 levels of the hierarchy (3 classes, 13 subclasses), but that the molecular signature of half the clusters
556 remains unclear. We expect that similar conclusions can be drawn for other tissues, such as blood,
557 where there is a similar hierarchical organization of cell types. The main difficulty is to identify
558 replicable cell types across datasets, which may be challenging during development or complex
559 differentiation processes, such as hematopoiesis.

560 The selection of replicable markers from single cell atlases is a promising avenue for several
561 applications, including cell type annotation, selection of gene panels and bulk data deconvolution. It
562 reduces rich information to a prioritized list that is simple to use and to refine. New computational
563 methods will benefit from highly condensed prior information about genes in the cell type space,
564 without having to train on large reference datasets. Finally, as new datasets are generated, marker
565 lists will become increasingly robust to new sources of variation, leading to higher downstream
566 performance across a diverse array of tasks.

567 Methods

568 Datasets

569 We downloaded the mouse primary cortex (MOp) BICCN datasets and cell type annotations from the
570 NeMO archive (http://data.nemoarchive.org) according to preprint instructions (Yao et al. 2020a).
571 We considered the 7 transcriptomic datasets from the mouse primary cortex: single cell Smart-Seq
572 (scSS), single nucleus Smart-Seq (snSS), single cell Chromium v2 (scCv2), single nucleus Chromium v2
573 (snCv2), single cell Chromium v3 (scCv3), single nucleus Chromium v3 from the Macosko and Zeng
574 labs (scCv3M and scCv3Z, respectively)(Table 1). We kept all cells with "class" annotated as
575 "Glutamatergic", "GABAergic" or "Non-Neuronal" and kept genes that were common to all datasets,
576 arriving at a total of 482,712 cells and 24,140 genes. We normalized counts to counts per millions
577 (CPM). For cell types, we considered five levels of annotations provided by the BICCN: "class",
578 "subclass", "cluster", "joint_subclass" and "joint_cluster". "subclass" and "cluster" labels were
579 obtained by clustering and annotating the datasets independently, while "joint_subclass" and
580 "joint_cluster" labels were obtained through joint clustering and annotation. Throughout the
581 manuscript, we use "joint_cluster" labels when we need common annotations across datasets,
582 otherwise, we use "cluster" labels. To map "subclass" labels across datasets, we used the
583 independent clustering, but mapped all clusters to one of the following names: "L2/3 IT", "L5 ET",

584 “L5 IT”, “L5/6 NP”, “L6 CT”, “L6 IT”, “L6 IT Car3”, “L6b”, “*Lamp5*”, “*Pvalb*”, “*Sncg*”, “*Sst*”, “*Vip*”. In the
585 last section (generalizability of meta-markers), we use the “joint_subclass” annotation instead,
586 because it explicitly includes the distinction between L4/5 IT and L5 IT cells.

587 The BICCN isocortex and hippocampus dataset was downloaded from the NeMO archive
588 (http://data.nemoarchive.org)(Yao et al. 2020b). The full dataset contains 1,646,439 cells annotated
589 to 379 cell types. Due its size, it was separated into sub-datasets corresponding to individually
590 sequenced brain regions (as annotated in the “region_label” metadata column), resulting in 19 brain
591 regions sequenced with 10X v3, 21 brain regions sequenced with SmartSeq (Table 1). We subset all
592 datasets to a common set of 24,140 genes. Preprocessing was similar to the MOp datasets: we kept
593 all cells with “class” annotated as “Glutamatergic”, “GABAergic” or “Non-Neuronal” and normalized
594 counts to counts per million (CPM) for SmartSeq datasets or counts per 10,000 (CP10K) for 10X
595 datasets.

596 Meta-analytic hierarchical differential expression statistics

597 For each cell type, we computed DE statistics independently in each dataset using MetaMarkers’
598 “compute_markers” function. We compared a cell type to neighboring cell types in the BICCN
599 taxonomy by setting the “group_labels” parameter. For example, the “GABAergic” class contains the
600 “*Pvalb*”, “*Sst*”, “*Sncg*”, “*Lamp5*” and “*Vip*” subclasses. By stratifying analysis by classes, DE statistics
601 for “*Pvalb*” were obtained by comparing “*Pvalb*” cells to all cells that are either “*Sst*”, “*Sncg*”,
602 “*Lamp5*” or “*Vip*”, but ignoring cells from other classes (excitatory neurons and glia). At the cluster
603 level, analysis is stratified by subclasses, e.g., *Pvalb* subtypes are compared to other *Pvalb* subtypes
604 only.

605 For each dataset, “compute_markers” returns a table of standard statistics. Let x_{ij}be the expression
606 of gene i in cell j (normalized to CPM in all the manuscript), let C be the cells belonging to the cell
607 type of interest, and $\overline{C}$ be all background cells. All statistics are computed for each gene

independently, so we will drop the subscript i in the following. The fold change (FC) is computed as the ratio of average expression between the cell type of interest and background cells, $FC = \overline{\{x_j\}_{j\in C}}/\overline{\{x_j\}_{j\in \overline{C}}}$. Statistical significance is based on the ROC test. First we compute the AUROC according to the following formula (derived from the Mann-Whitney U statistic):

$$AUROC = \frac{1}{NP}\left(\sum_{j\in C} r_j - \frac{P(P+1)}{2}\right),$$

where $P = |C|$ are the number of positives (cells from the cell type of interest), $N = |\overline{C}|$ are the number of negatives (background cells), and r_j are the ranks of positives (obtained after ranking all cells according to the gene's expression value). P-values are computed under a normal approximation of the AUROC with continuity and tie correction according to the following formulas:

$$z = (AUROC - 0.5) / \sigma; \quad \sigma = \sqrt{\frac{NP}{12}(P + 1 - T)}; \quad T = \sum_{i=1}^{k} \frac{{t_i}^3 - t_i}{(N+P)(N+P+1)};$$

where z follows a standard normal distribution under the null hypothesis that positives and negatives are from the same population, σ is the analytical standard deviation of AUROC, T is a tie correction formula where k is the number of distinct expression values and t_i is the number of cells that share the same expression value with index i. P-values are converted to False Discovery Rates (FDR) according to the Benjamini-Hochberg procedure. For exhaustivity, we considered four additional statistics related to binarized gene expression: gene detection rate, fold change of detection rate (FCd), recall and precision. Gene detection rate is the fraction of cells in the population of interest that express the gene of interest, $dr_C = |\{x_j > 0\}_{j\in C}| \,/\, |C|$. FCd= $dr_C/dr_{\overline{C}}$ is the ratio of gene detection rates in the population of interest over the background population. Recall is identical to gene detection rate (seen from a classification perspective). Precision= $|\{x_j > 0\}_{j\in C}| \,/\, |\{x_j > 0\}_{j\in C\cup\overline{C}}|$ is the fraction of cells expressing the gene of interest that belong to the population of interest. All operations are vectorized across genes and cell types to allow rapid marker extraction and aggregation across datasets.

631 We combined statistics across datasets using MetaMarkers' "make_meta_markers" function, which
632 averages the above statistics across datasets for all cell types. "make_meta_markers" uses the
633 arithmetic mean by default, and uses the geometric mean for the following statistics: FC, FCd,
634 expression. To define DE recurrence, we used the number of datasets where a gene is reliably DE
635 ("fdr_threshold=0.05", "fc_threshold=4"). Throughout the manuscript, we considered a gene to be
636 DE if it had a FC>1 and an FDR<0.05, and reliably DE if FC>4 and FDR<0.05.

637 Reliable fold change and AUROC thresholds

638 To establish the reliability of FC, we picked all combinations of training datasets and extracted genes
639 that were significantly upregulated in all training datasets (AUROC>0.5, FDR<0.05, average FC>1).
640 Then, for each gene, we looked up the held out datasets and counted how often the gene remained
641 upregulated (FC>1) or was detected as downregulated (FC<=1). We summarized the results as a type
642 S error, the fraction of held out datasets where the gene was detected as downregulated. Formally,
643 let$G$be the set of genes that are consistently upregulated across training datasets d_1, …, d_m. Let d'_1,
644 …, d'_nbe the held-out test datasets. For a given cell type, the average type S error is defined as:

645 $$e = \frac{1}{n|G|}\left|\{FC_{gd'_i} < 1\}_{g\in G, i\in\{1..n\}}\right|,$$

646 where $FC_{gd'_i}$is the fold change of gene g in test dataset d'_i. We computed the type S error across all
647 combinations of cell types and training datasets. To establish the reliability of AUROC, we followed a
648 similar procedure, replacing the FC<1 condition by AUROC>0.5.

649 MetaNeighbor cell type replicability score

650 To compute the association between the number of markers and cell type replicability, we
651 computed cell type similarity using MetaNeighbor by following the procedure described in (Yao et al.
652 2020a). Briefly, MetaNeighbor uses a neighbor voting framework to match cell types from a train
653 dataset to a test dataset, where the matching strength is quantified as an AUROC. First, we use the

654 “MetaNeighborUS” function to create a graph where each node is a cell type and each edge is the
655 matching strength (directed from train dataset to test dataset). By applying the
656 “extractMetaClusters” function, we keep only edges that correspond to high confidence reciprocal
657 matches (1-vs-best AUROC > 0.7 both ways). After this step, we are left with groups of connected
658 cell types that we call “meta-clusters”. The replicability score is the number of datasets spanned by
659 the meta-cluster, e.g. a cell type has a score of 6 if it is connected to cell types from 5 other datasets.
660 For visualization purposes, we created jittering by adding the average AUROC across the meta-
661 cluster to the replicability score. To avoid overfitting, we considered the “cluster” annotations from
662 the BICCN, which were obtained by clustering and annotating the datasets independently.

663 Marker-based cell type classification

664 To quantify the ability of a list of markers to identify a cell type, we framed the problem as a
665 hierarchical classification task where we predict cell type labels from gene expression. First, for each
666 cell, we computed a prediction score by averaging expression profiles across markers. Let x_{ij} be the
667 CPM-normalized expression of gene $i$in cell j, and M_c be a set of marker genes for cell type c. For
668 each cell j, we compute the marker score is:

669 $$S_j(c) = \frac{1}{|M_c|} \sum_{i \in M_c} log2(x_{ij} + 1)$$

670 This score is efficiently implemented by MetaMarker’s “score_cells” function. To obtain marker-wide
671 renormalized scores, we compute the above score for a series of cell types $c_1, .., c_n$ then, for each
672 cell type, we compute:

673 $$S'_j(c) = S_j(c) \,/\, \frac{1}{n} \sum_{i=1}^{n} S_j(c_i)$$

674 To compute classification performance, we labeled cells from the cell type of interest as positives
675 and cells from cell types sharing the same parent class or subclass as negatives (similar to DE

676 statistics computation, see “Meta-analytic hierarchical differential expression statistics”). Intuitively,
677 we are looking whether positives (cells from the cell type of interest) have high prediction scores
678 (marker scores). We summarized the prediction accuracy as an AUROC (in the threshold-free case)
679 and F1 (harmonic mean of precision and recall, in the thresholding case). To avoid circularity, we
680 always made predictions on held out datasets. For markers from a single dataset, predictions were
681 averaged across the 6 remaining datasets; for meta-analytic markers, we picked markers on all
682 combinations of 6 datasets and predicted cell types in the remaining dataset. We obtained
683 classification scores for individual populations of neurons by averaging over every combination of
684 train and test datasets.

685 Gene ontology enrichment of meta-markers

686 Gene ontology terms and mouse annotations were downloaded using the org.Mm.eg.db and GO.db
687 R packages. To focus on specific cell processes, we further selected terms from the “Biological
688 Process” ontology containing between 20 and 100 gene annotations. Gene set enrichment was
689 computed using the hypergeometric test, based on R’s “phyper” function and the Maximum
690 Likelihood Estimate (MLE) of the sample odds ratio (OR).

691 Marker-based deconvolution

692 To investigate the impact of marker selection on deconvolution, we applied deconvolution in a
693 hierarchical framework similar to DE computation and cell type classification. We applied Non-
694 Negative Least Square (NNLS) deconvolution (Abbas et al. 2009) using the nnls R package, which was
695 shown to be both efficient and accurate according to multiple recent benchmarks (Patrick et al.
696 2020; Cobos et al. 2020). Briefly, we inferred cell type proportions from the following equation:

697 $$T = C.P$$

698 where T is a bulk expression matrix (genes x sample), in our case pseudo-bulk matrices extracted
699 from each test dataset, C is a cell type signature matrix (genes x cell type), P is the estimated cell

700 type proportion matrix (cell type x sample). To test all combinations of train and test datasets, we
701 split each dataset in half by assigning each cell randomly to a test or train fold. From each train fold,
702 we built signature matrices by averaging unnormalized expression profiles for each cell type. From
703 each test fold, we built 1000 pseudo-bulks containing 1000 cells. To generate pseudo-bulks with
704 highly variable cell type proportions, we started by drawing target cell type proportions for each
705 pseudo-bulk, in a procedure similar to (Cobos et al. 2020). We sampled target proportions for each
706 cell type from a uniform distribution, normalized proportions to 1, then converted to a target
707 number of cells which we sampled with replacement, then averaged the unnormalized counts.

708 Given a set of markers (obtained from a single dataset or meta-analytically across all datasets except
709 the test dataset), a train dataset (signature matrix) and a test dataset (1000 pseudo-bulks), we
710 performed NNLS deconvolution by subsetting the signature matrix C and pseudo-bulks T to the set of
711 markers. We computed deconvolution performance as the Pearson correlation between theoretical
712 cell type proportions and the predicted cell type proportions (one value per pseudo-bulk). For
713 computational efficiency, we only tested one group of populations at the "joint cluster" level. We
714 chose to focus on the Lamp5 populations, as it contained 8 populations that were well represented
715 across all datasets (range 14 to 3016 cells per single population, 257 cells on average).

716 Note that, because of the difficulty of matching UMI counts with full-length read counts (Newman et
717 al. 2019), we only considered train-test combinations within similar technologies (one pair of Smart-
718 seq datasets, 10 pairs of 10X datasets). To control for globally encoded differences in expression
719 profiles (correlating with the first principal component), we created random marker sets by picking
720 genes that were expression-matched with meta-analytic markers (decile-matched).

721 Generation of robust meta-marker sets

722 We generated meta-marker sets for each cell type in the MOp hierarchy (Sup. Data 1-3), using the
723 "class", "joint_subclass" and "joint_cluster" annotation levels (see "Meta-analytic hierarchical

724 differential expression statistics"). We kept meta-markers that were either strongly DE (FC>4,
725 FDR<0.05) in at least one dataset or had a meta-analytic FC > 2. We ranked the remaining markers
726 by recurrence, then by AUROC, and selected the top 100 genes (top 50 genes for clusters). If fewer
727 than 100 markers remained, we selected all remaining markers.

728 Cell type annotation of the BICCN isocortex and hippocampus datasets

729 We annotated cells in the isocortex and hippocampus datasets using our robust marker lists (see
730 "Generation of robust marker lists for all BICCN MOp cell types"). To annotate cell types, we adopted
731 a hierarchical cell type annotation procedure. We classified each brain region independently,
732 starting from the log-normalized count matrix. First, we obtained marker scores (average meta-
733 marker expression, see "Marker-based cell type classification") for all cells by running MetaMarker's
734 "score_cells" function. Then, marker scores were converted into cell type predictions using
735 MetaMarker's "assign_cell" function, which finds the marker set with the highest score and returns
736 several QC metric, including the highest score and the marker enrichment (observed score divided
737 by expected score, under the assumption that all marker sets have equal expression). The
738 "assign_cell" function takes two parameters: marker scores and group-level assignments. For
739 subclasses, we provided class-level predictions as the group assignments; for clusters, we provided
740 subclass-level predictions as the group assignments. To filter out cells with unclear assignments, we
741 labeled cells that had a marker enrichment below 1.5 (unless otherwise indicated in the text) as
742 "unassigned".

743 Data access

744 The datasets analyzed during the current study are available in the NeMO archive
745 (https://nemoarchive.org/) at https://assets.nemoarchive.org/dat-ch1nqb7. The full meta-marker
746 lists for the BICCN cell types and optimal number of markers are available on FigShare at

https://doi.org/10.6084/m9.figshare.13348064. The code for MetaMarkers is freely available as an R package on Github at https://github.com/gillislab/MetaMarkers.

Competing interest statement

The authors declare that they have no competing financial interests.

Acknowledgments

JG was supported by NIH grants R01MH113005 and R01LM012736. SF was supported by NIH grant U19MH114821.

Author contributions

SF and JG designed the experiments, performed the data analysis and wrote the paper. All authors read and approved the final manuscript.

Supplemental Material

Supplemental Figure 1. a-c Type S error as a function of AUROC in train datasets (a), marker rank by fold change (b) and marker rank by AUROC (c). The dashed line indicates a type S error of 5%, ribbons around lines indicate variability across cell types and test datasets. **d-g** Type S error as a function of AUROC (d) or FC (e-g) in train dataset, with facets showing variability across hierarchy level (d,e), average cell type size (f) and average gene expression (g). **h** Pareto fronts in FC/AUROC space for inhibitory subclasses. Arrows point to the main historical marker for each subclass. **i** Expression of genes on the Sncg Pareto front across BICCN inhibitory clusters. **j** Pareto fronts in FC/AUROC space for excitatory subclasses. **k** Expression of genes on the L5 ET Pareto front across BICCN excitatory clusters.

767 **Supplemental Figure 2. a-d** Number of perfect markers (a), specific markers (b), sensitive markers
768 (c), and weak markers (d) for BICCN clusters, with cell types ordered according to number of
769 markers, colored according to the dataset used to compute markers. **e-f** MetaNeighbor replicability
770 as a function of the number of specific markers in the scCv2 dataset (e) and the number of perfect
771 markers in the snCv3M dataset (f).

772 **Supplemental Figure 3. a-c** Parametric curve in FC/AUROC space showing evolution of classification
773 performance with an increasing number of marker genes at the class (a), subclass (b) and joint
774 cluster (c) level. **d-g** Breakdown of optimal AUROC performance (meta-analytic markers) as a
775 function of dataset depth, colored by hierarchy level (d), for individual classes, showing variability
776 across test datasets (e), for individual subclasses, showing variability across test datasets (f), for
777 individual clusters, showing variability across test datasets (g). **h-k** Same as d-g with signal-to-noise
778 ratio (FC) at optimal performance instead of AUROC. **l-o** Same as d-g with number of genes at
779 optimal performance instead of AUROC.

780 **Supplemental Figure 4. a-b** Summary of optimal classification performance (F1) across hierarchy
781 levels with transcriptome-wide normalization (a) and marker-wide renormalization (b). Variability is
782 shown across cell types and test datasets. **c-e** Heatmap detailing classification performance for each
783 cell type as a function of the number of genes at the class (c), subclass (d) and cluster (e) level.

784 **Supplemental Figure 5. Top 200 meta-markers show strong, but less specific, enrichment for**
785 **synaptic processes. a** Total number of significantly enriched GO terms (orange) and fraction of
786 significant GO terms that are enriched in a unique cell type (blue) for BICCN classes when an
787 increasing number of meta-markers are considered. **b** Top 3 enriched Gene Ontology (GO) terms for
788 the top 200 meta-markers for each BICCN class. For each dot, the size reflects the False Discovery
789 Rate (FDR), the color reflects the Odds Ratio (OR) of the enrichment test (hypergeometric test). **c**
790 Same as **b** for the top 200 meta-markers for BICCN GABAergic subclasses. **d** Same as **b** for the top
791 200 meta-markers for BICCN Glutamatergic subclasses (only top 2 terms are shown).

792 **Supplemental Figure 6. Meta-analytic markers improve deconvolution performance at every level**
793 **of the hierarchy. a** Schematic of deconvolution task. **b** Summary of deconvolution performance
794 (Pearson's r) at each hierarchy level with 100 markers per cell type. Colors show 3 marker
795 prioritization strategies (single dataset markers, meta-analytic markers or expression-level matched
796 random genes). **c-e** Deconvolution performance (Pearson correlation of true and estimated cell type
797 proportions) for 3 marker prioritization strategies at the class level (c), the subclass level (d), and the
798 cluster level (e). Colors as b.

799 **Supplemental Figure 7. Focus on subclass-level predictions in the auditory cortex. a** Subclass-level
800 predictions in the auditory cortex based on the top 100 meta-markers. Cells remain unassigned (NA)
801 if the enrichment score is lower than 2 for all subclasses. **b** Subclass-level predictions for non-
802 neurons in the auditory cortex based on the top 100 meta-markers (left) and reference labels (right).
803 **c** Subclass-level predictions for Intra-Telencephalic (IT) excitatory neurons in the auditory cortex
804 based on the top 100 meta-markers (left) and reference labels (right).

805 **Supplemental Figure 8. Cluster-level predictions in the auditory cortex. a-c** Cluster-level predictions
806 for Lamp5 inhibitory neurons (a), Near-Projecting (NP) excitatory neurons (b) and layer 2/3 Intra-
807 Telencephalic (IT) excitatory neurons (c) in the auditory cortex based on the top 100 meta-markers
808 (left) and reference labels (right). In all panels, cells remain unassigned (NA) if the enrichment score
809 is lower than 1.5 for all clusters.

810 **Supplemental Figure 9. The marker enrichment score provides robust separability of cell types in**
811 **other cortical regions. a** Marker enrichment scores based on the top 100 meta-markers for the
812 3 BICCN classes in the auditory cortex. The facets are organized according to reference cell types
813 (from the auditory cortex), the x-axis according to meta-markers sets (for the motor cortex). **b** Same
814 as **a** for BICCN inhbitory subclasses. **c** Same as **a** for BICCN excitatory subclasses. **d** Confusion matrix
815 showing the concordance of subclass-level predictions based on the top 100 meta-markers with

816 reference cell types across 40 brain areas. Cells are unassigned if the marker enrichment is lower
817 than 2 for all subclasses.

818 **Supplemental Data 1. Class-level markers.** Top 100 robust markers for BICCN cell types at the class
819 level in CSV format (.csv).

820 **Supplemental Data 2. Subclass-level markers.** Top 100 robust markers for BICCN cell types at the
821 subclass level in CSV format (.csv).

822 **Supplemental Data 3. Cluster-level markers.** Top 50 robust markers for BICCN cell types at the
823 cluster level in CSV format (.csv).

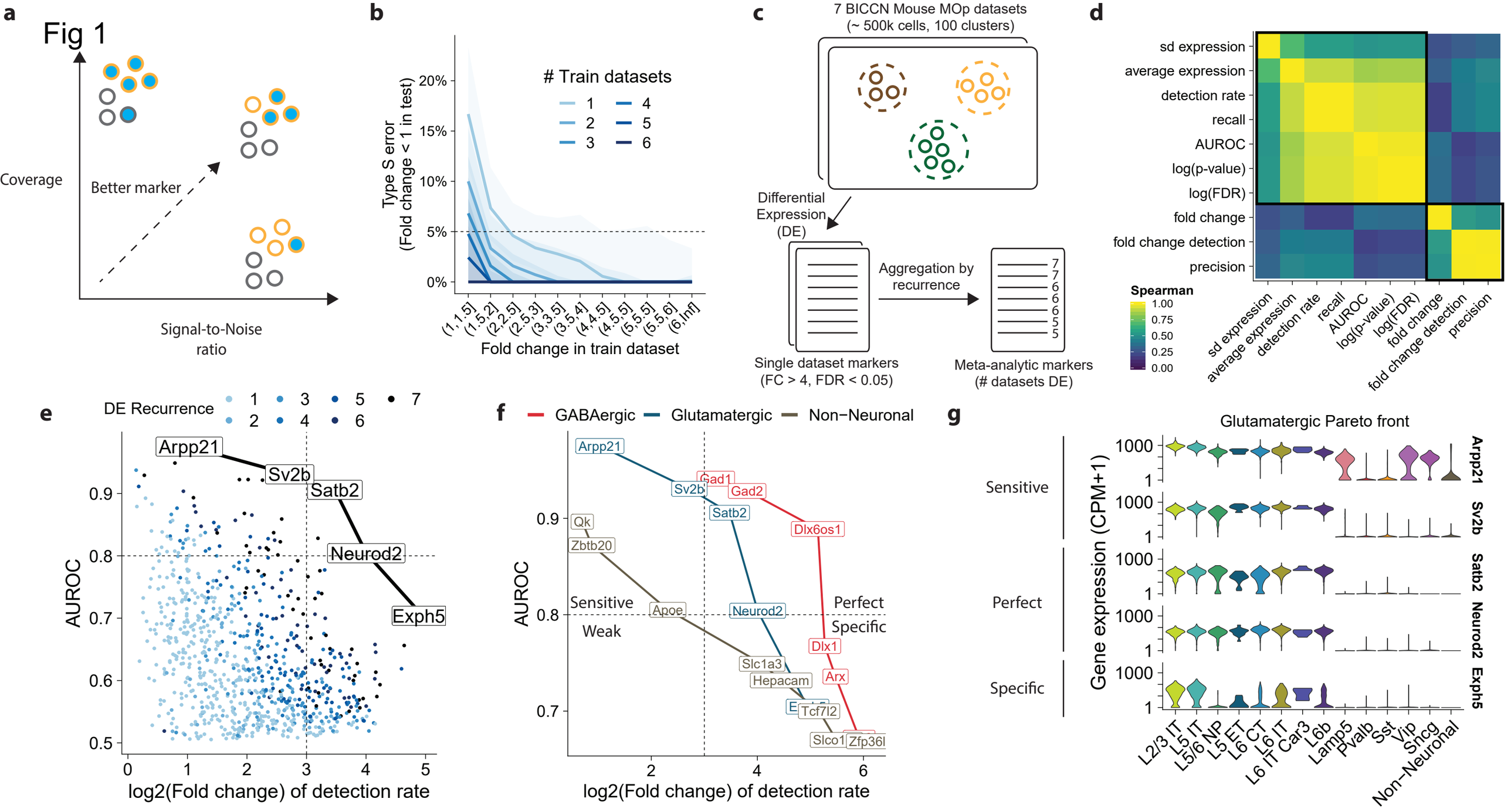
824 References

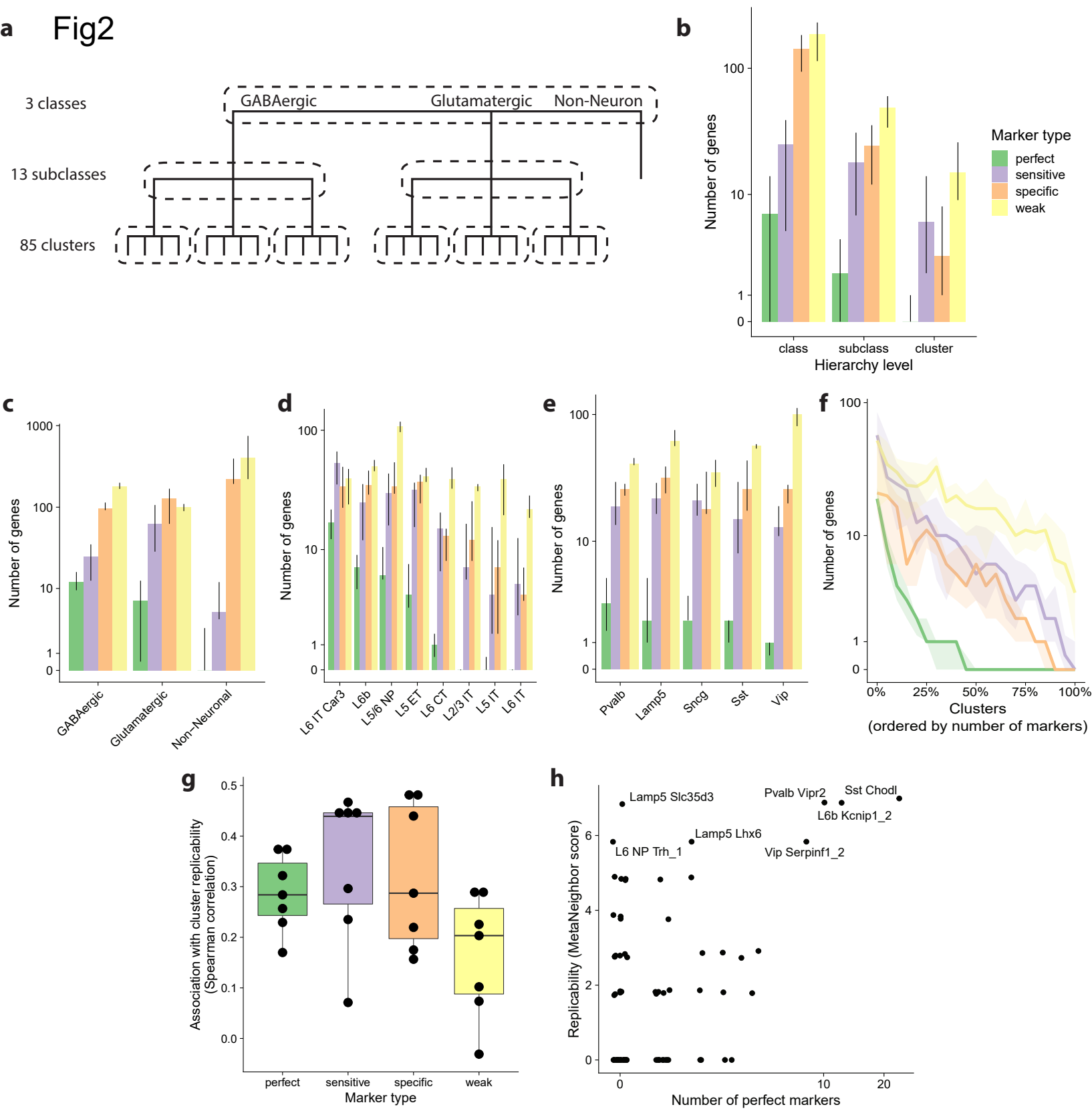
825 Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. 2009. Deconvolution of Blood
826 Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus.
827 *PLOS ONE* **4**: e6098.
828 Andrews TS, Hemberg M. 2019. M3Drop: dropout-based feature selection for scRNASeq.
829 *Bioinformatics* **35**: 2865–2867.
830 Ascoli GA, Alonso-Nanclares L, Anderson SA, Barrionuevo G, Benavides-Piccione R, Burkhalter A,
831 Buzsáki G, Cauli B, DeFelipe J, Fairén A, et al. 2008. Petilla terminology: nomenclature of
832 features of GABAergic interneurons of the cerebral cortex. *Nat Rev Neurosci* **9**: 557–568.
833 Asp M, Giacomello S, Larsson L, Wu C, Fürth D, Qian X, Wärdell E, Custodio J, Reimegård J, Salmén F,
834 et al. 2019. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing
835 Human Heart. *Cell* **179**: 1647-1660.e19.
836 Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, Crow M, Hodge RD, Krienen FM,
837 Sorensen SA, et al. 2020. Evolution of cellular diversity in primary motor cortex of human,
838 marmoset monkey, and mouse. *bioRxiv* 2020.03.31.016972.
839 Bhaduri A, Andrews MG, Mancia Leon W, Jung D, Shin D, Allen D, Jung D, Schmunk G, Haeussler M,
840 Salma J, et al. 2020. Cell stress in cortical organoids impairs molecular subtype specification.
841 *Nature*.
842 Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data
843 across different conditions, technologies, and species. *Nat Biotechnol*.
844 http://www.nature.com/doifinder/10.1038/nbt.4096 (Accessed May 7, 2018).
845 Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA, Blecher-Gonen R,
846 Zhang F, et al. 2020. A human cell atlas of fetal gene expression. *Science* **370**.
847 https://science.sciencemag.org/content/370/6518/eaba7721 (Accessed December 1, 2020).
848 Chen M, Zhou X. 2018. VIPER: variability-preserving imputation for accurate gene expression
849 recovery in single-cell RNA sequencing studies. *Genome Biol* **19**: 196.
850 Cobos FA, Alquicira-Hernandez J, Powell J, Mestdagh P, Preter KD. 2020. Comprehensive
851 benchmarking of computational deconvolution of transcriptomics data. *bioRxiv*
852 2020.01.10.897116.
853 Consortium S-I, Su Z, Łabaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, Wang C, Schroth GP,

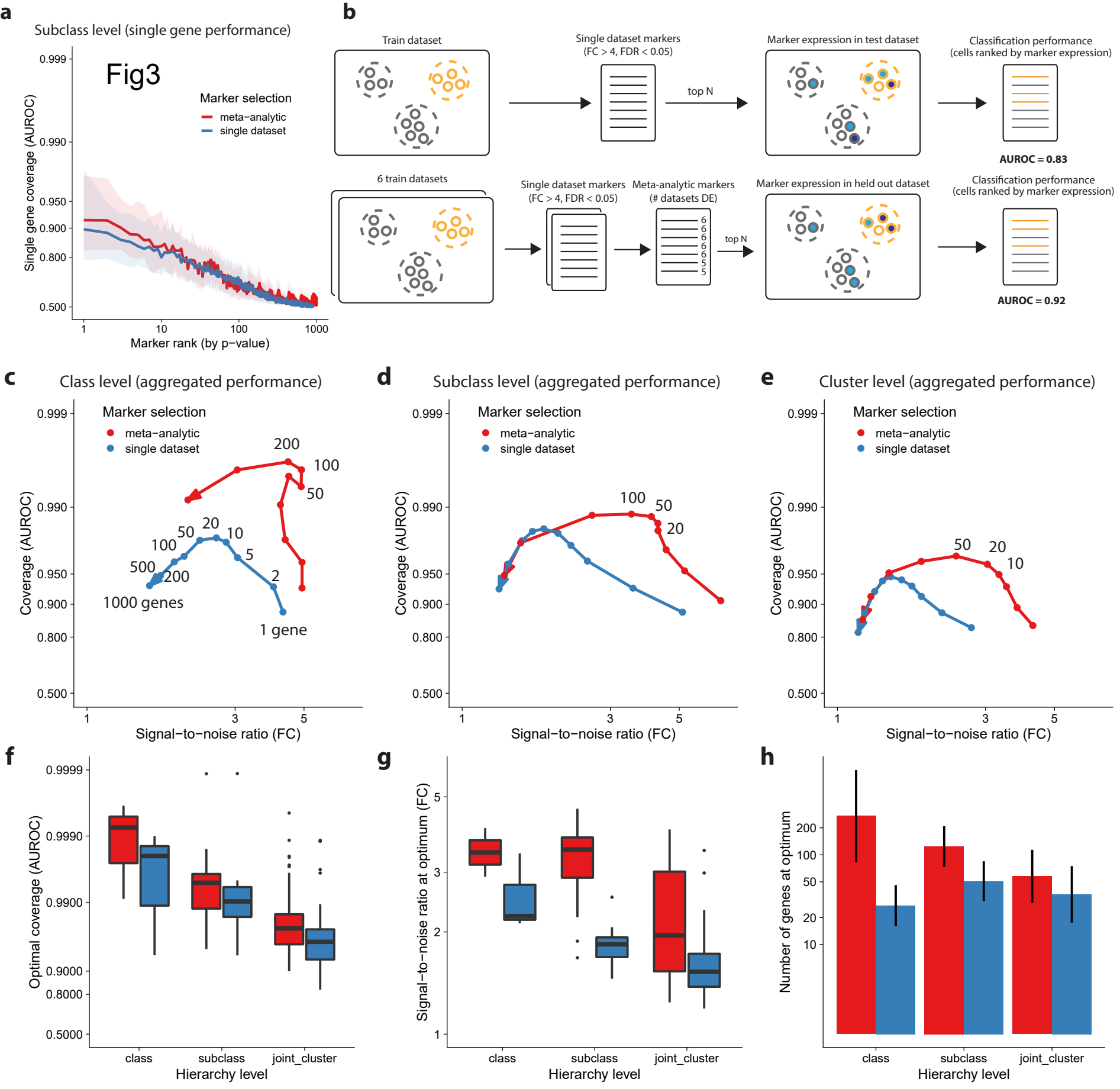
854 Setterquist RA, et al. 2014. A comprehensive assessment of RNA-seq accuracy,
855 reproducibility and information content by the Sequencing Quality Control Consortium. *Nat*
856 *Biotechnol* **32**: 903–914.
857 Crow M, Gillis J. 2018. Co-expression in Single-Cell Analysis: Saving Grace or Original Sin? *Trends*
858 *Genet* **34**: 823–831.
859 Cui X, Churchill GA. 2003. Statistical tests for differential expression in cDNA microarray experiments.
860 *Genome Biol* **4**: 210.
861 Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, Jiang Y. SCDC: bulk gene expression
862 deconvolution by multiple single-cell RNA sequencing references. *Brief Bioinform*.
863 https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz166/5699815 (Accessed
864 August 31, 2020).
865 Dumitrascu B, Villar S, Mixon DG, Engelhardt BE. 2019. Optimal gene selection for cell type
866 discrimination in single cell analyses. *bioRxiv* 599654.
867 Gelman A, Carlin J. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M
868 (Magnitude) Errors. *Perspect Psychol Sci* **9**: 641–651.
869 Goedhart J, Luijsterburg MS. 2020. VolcaNoseR – a web app for creating, exploring, labeling and
870 sharing volcano plots. *bioRxiv* 2020.05.07.082263.
871 Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing
872 data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**: 421–427.
873 Hicks SC, Townes FW, Teng M, Irizarry RA. 2018. Missing data and technical variability in single-cell
874 RNA-sequencing experiments. *Biostatistics* **19**: 562–578.
875 Hobert O. 2008. Regulatory logic of neuronal diversity: Terminal selector genes and selector motifs.
876 *Proc Natl Acad Sci* **105**: 20067–20071.
877 Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N,
878 Penn O, et al. 2019. Conserved cell types with divergent features in human versus mouse
879 cortex. *Nature* **573**: 61–68.
880 Huang Q, Liu Y, Du Y, Garmire LX. 2020. Evaluation of Cell Type Annotation R Packages on Single Cell
881 RNA-seq Data. *bioRxiv* 827139.
882 Huang ZJ. 2014. Toward a Genetic Dissection of Cortical Circuits in the Mouse. *Neuron* **83**: 1284–
883 1302.
884 Hunt GJ, Gagnon-Bartsch JA. 2019. The Role of Scale in the Estimation of Cell-type Proportions.
885 *bioRxiv* 857805.
886 Johnson MB, Walsh CA. 2017. Cerebral cortical neuron diversity and development at single-cell
887 resolution. *Curr Opin Neurobiol* **42**: 9–16.
888 Kessaris N, Magno L, Rubin AN, Oliveira MG. 2014. Genetic programs controlling cortical interneuron
889 fate. *Curr Opin Neurobiol* **26**: 79–87.
890 Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential
891 expression analysis. *Nat Methods* **11**: 740–742.
892 Kiselev VY, Yiu A, Hemberg M. 2018. scmap: projection of single-cell RNA-seq data across data sets.
893 *Nat Methods* **15**: 359–362.
894 Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P,
895 Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with
896 Harmony. *Nat Methods* **16**: 1289–1296.
897 Krienen FM, Goldman M, Zhang Q, Rosario R del, Florio M, Machold R, Saunders A, Levandowski K,
898 Zaniewski H, Schuman B, et al. 2019. Innovations in Primate Interneuron Repertoire. *bioRxiv*
899 709501.
900 Lin Y, Ghazanfar S, Wang KYX, Gagnon-Bartsch JA, Lo KK, Su X, Han Z-G, Ormerod JT, Speed TP, Yang
901 P, et al. 2019. scMerge leverages factor analysis, stable expression, and pseudoreplication to
902 merge multiple single-cell RNA-seq datasets. *Proc Natl Acad Sci* **116**: 9775–9784.
903 Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. 2020. Jointly Defining Cell Types from
904 Multiple Single-Cell Datasets Using LIGER. *bioRxiv* 2020.04.07.029546.

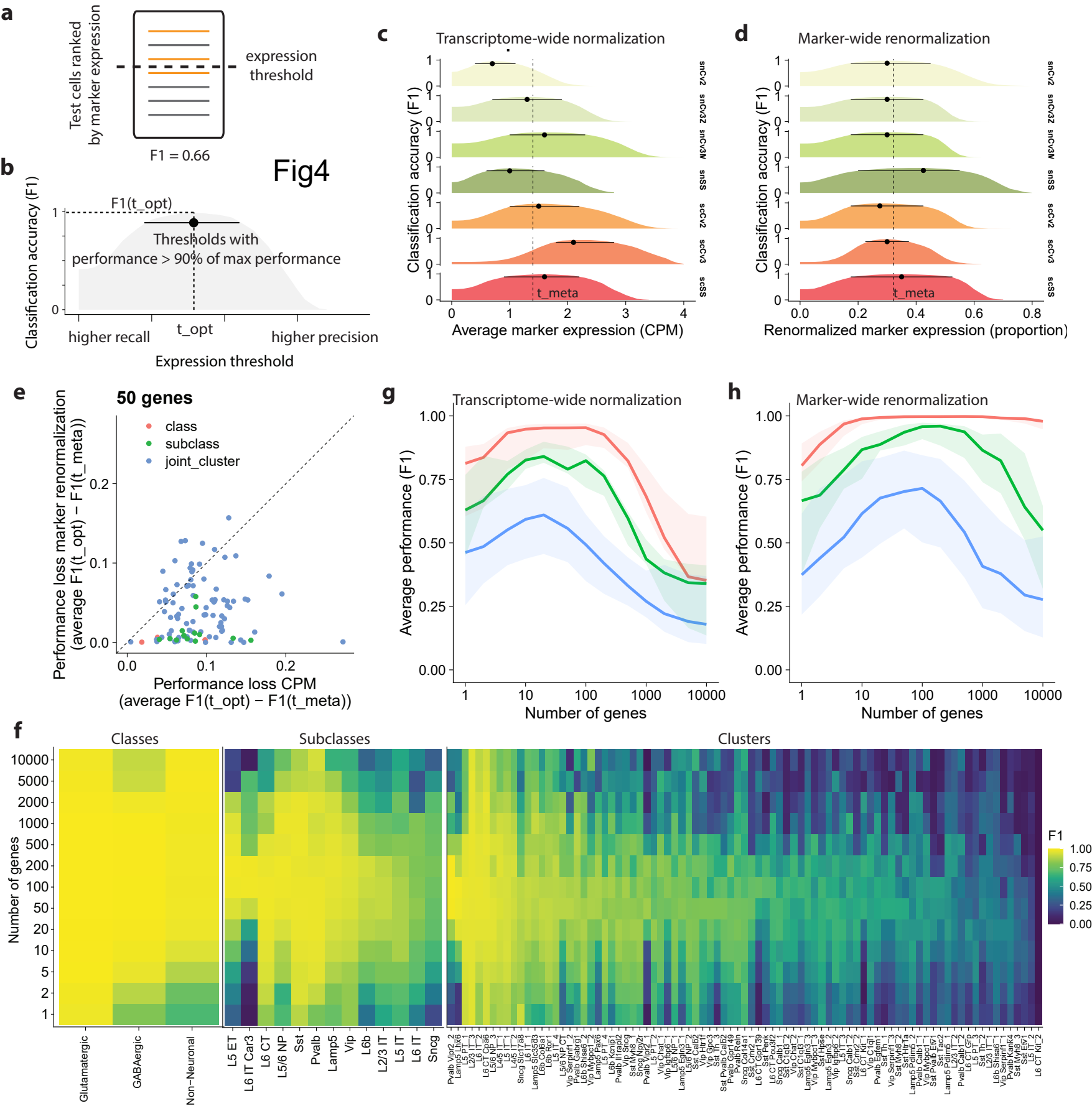
905 Lodato S, Arlotta P. 2015. Generating Neuronal Diversity in the Mammalian Cerebral Cortex. *Annu*
906 *Rev Cell Dev Biol* **31**: 699–720.
907 Mancarci BO, Toker L, Tripathy SJ, Li B, Rocco B, Sibille E, Pavlidis P. 2017. Cross-Laboratory Analysis
908 of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data.
909 *eneuro* **4**: ENEURO.0212-17.2017.
910 Mayer C, Hafemeister C, Bandler RC, Machold R, Batista Brito R, Jaglin X, Allaway K, Butler A, Fishell
911 G, Satija R. 2018. Developmental diversification of cortical inhibitory interneurons. *Nature*
912 **555**: 457–462.
913 Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J,
914 Regev A, Dulac C, et al. 2018. Molecular, spatial and functional single-cell profiling of the
915 hypothalamic preoptic region. *Science* **362**: eaau5324.
916 Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS,
917 Luca BA, Steiner D, et al. 2019. Determining cell type abundance and expression from bulk
918 tissues with digital cytometry. *Nat Biotechnol 2019* 1.
919 Packer JS, Zhu Q, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, Stefanik D, Tan K, Trapnell C, Kim
920 J, et al. 2019. A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell
921 resolution. *Science* **365**. https://science.sciencemag.org/content/365/6459/eaax1971
922 (Accessed December 8, 2020).
923 Patrick E, Taga M, Ergun A, Ng B, Casazza W, Cimpean M, Yung C, Schneider JA, Bennett DA, Gaiteri
924 C, et al. 2020. Deconvolving the contributions of cell-type heterogeneity on cortical gene
925 expression. *PLOS Comput Biol* **16**: e1008120.
926 Paul A, Crow M, Raudales R, He M, Gillis J, Huang ZJ. 2017. Transcriptional Architecture of Synaptic
927 Communication Delineates GABAergic Neuron Identity. *Cell* **171**: 522-539.e20.
928 Pliner HA, Shendure J, Trapnell C. 2019. Supervised classification enables rapid annotation of cell
929 atlases. *Nat Methods* **16**: 983–986.
930 Poulin J-F, Tasic B, Hjerling-Leffler J, Trimarchi JM, Awatramani R. 2016. Disentangling neural cell
931 diversity using single-cell transcriptomics. *Nat Neurosci* **19**: 1131–1141.
932 Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, Hjerling-Leffler J,
933 Nilsson M. 2020. Probabilistic cell typing enables fine mapping of closely related cell types in
934 situ. *Nat Methods* **17**: 101–106.
935 RAMON Y CAJAL S. 1904. Textura del Sistema Nervioso del Hombre y de los Vertebrados, tomo II,
936 primera parte. *Imprenta Libr Nicolas Moya Madr Repr Graf Vidal Leuka Alicante 1992* 399–
937 402.
938 Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. 2018. A general and flexible method for signal
939 extraction from single-cell RNA-seq data. *Nat Commun* **9**: 284.
940 Schaum N, Karkanias J, Neff NF, May AP, Quake SR, Wyss-Coray T, Darmanis S, Batson J, Botvinnik O,
941 Chen MB, et al. 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris.
942 *Nature* **562**: 367–372.
943 Shi L, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F,
944 Kawasaki ES, et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and
945 intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**: 1151–
946 1161.
947 Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P,
948 Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888-1902.e21.
949 Svensson V, Vento-Tormo R, Teichmann SA. 2018. Exponential scaling of single-cell RNA-seq in the
950 past decade. *Nat Protoc* **13**: 599–604.
951 Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, Goldy J, Garren E, Economo MN,
952 Viswanathan S, et al. 2018. Shared and distinct transcriptomic cell types across neocortical
953 areas. *Nature* **563**: 72–78.
954 Tosches MA, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G, Laurent G. 2018. Evolution of
955 pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in

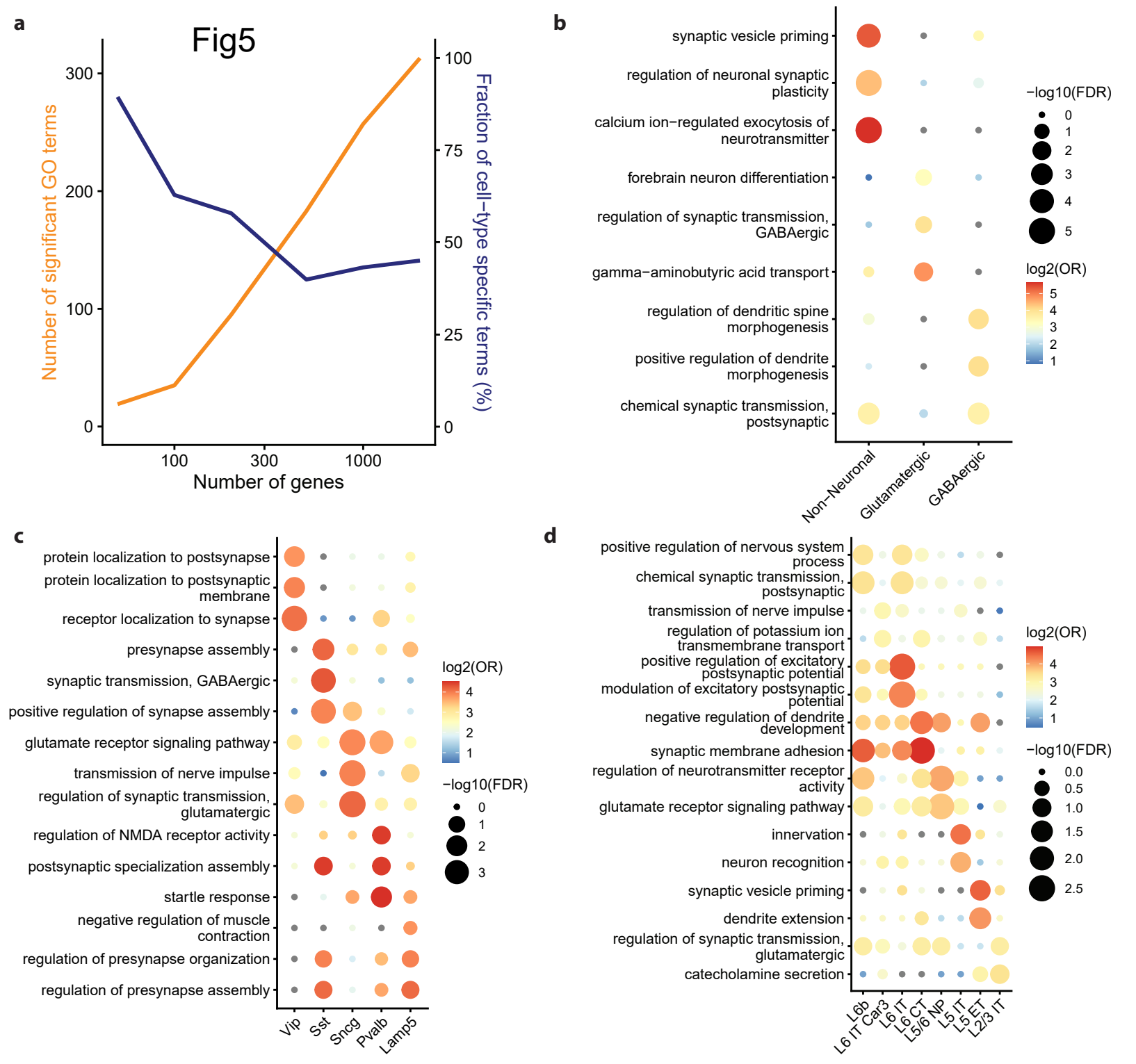
956 reptiles. *Science* **360**: 881–888.
957 Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan G-C. 2019. Accurate estimation of cell-type
958 composition from gene expression data. *Nat Commun* **10**: 2975.
959 Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, Gilad Y. 2017. Batch effects and
960 the effective design of single-cell gene expression studies. *Sci Rep* **7**: 39921.
961 Velasco S, Kedaigle AJ, Simmons SK, Nash A, Rocha M, Quadrato G, Paulsen B, Nguyen L, Adiconis X,
962 Regev A, et al. 2019. Individual brain organoids reproducibly form cell diversity of the human
963 cerebral cortex. *Nature* 1.
964 Wang X, Park J, Susztak K, Zhang NR, Li M. 2019. Bulk tissue cell type deconvolution with multi-
965 subject single-cell expression reference. *Nat Commun* **10**: 1–9.
966 Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. 2019. Single-Cell Multi-omic
967 Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**: 1873-1887.e17.
968 Yao Z, Liu H, Xie F, Fischer S, Booeshaghi AS, Adkins RS, Aldridge AI, Ament SA, Pinto-Duarte A,
969 Bartlett A, et al. 2020a. An integrated transcriptomic and epigenomic atlas of mouse primary
970 motor cortex cell types. *bioRxiv* 2020.02.29.970558.
971 Yao Z, Nguyen TN, Velthoven CTJ van, Goldy J, Sedeno-Cortes AE, Baftizadeh F, Bertagnolli D, Casper
972 T, Crichton K, Ding S-L, et al. 2020b. A taxonomy of transcriptomic cell types across the
973 isocortex and hippocampal formation. *bioRxiv* 2020.03.30.015214.
974 Yuste R, Hawrylycz M, Aalling N, Aguilar-Valles A, Arendt D, Arnedillo RA, Ascoli GA, Bielza C,
975 Bokharaie V, Bergmann TB, et al. 2020. A community-based transcriptomics classification
976 and nomenclature of neocortical cell types. *Nat Neurosci* 1–13.
977 Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E,
978 Borm LE, La Manno G, et al. 2018. Molecular Architecture of the Mouse Nervous System. *Cell*
979 **174**: 999-1014.e22.
980 Zhang M, Eichhorn SW, Zingg B, Yao Z, Zeng H, Dong H, Zhuang X. 2020. Molecular, spatial and
981 projection diversity of neurons in primary motor cortex revealed by in situ single-cell
982 transcriptomics. *bioRxiv* 2020.06.04.105700.
983 Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, Mahrt E, Guo W, Stawiski EW, Modrusan Z, et al.
984 2019. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. *Genes*
985 **10**: 531.
986


a
Fig 1
Coverage
Better marker
Signal-to-Noise ratio
b
Train datasets
1
2
3
4
5
6
Type S error (Fold change < 1 in test)
20%
15%
10%
5%
0%
(1,1.5]
(1.5,2]
(2,2.5]
(2.5,3]
(3,3.5]
(3.5,4]
(4,4.5]
(4.5,5]
(5,5.5]
(5.5,6]
(6,Inf]
Fold change in train dataset
c
7 BICCN Mouse MOp datasets
(~ 500k cells, 100 clusters)
Differential Expression (DE)
Aggregation by recurrence
Single dataset markers
(FC > 4, FDR < 0.05)
Meta-analytic markers
(# datasets DE)
d
sd expression
average expression
detection rate
recall
AUROC
log(p-value)
log(FDR)
fold change
fold change detection
precision
Spearman
1.00
0.75
0.50
0.25
0.00
e
DE Recurrence
1
2
3
4
5
6
7
Arpp21
Sv2b
Satb2
Neurod2
Exph5
AUROC
log2(Fold change) of detection rate
f
GABAergic
Glutamatergic
Non−Neuronal
Arpp21
Sv2b
Satb2
Gad1
Gad2
Dlx6os1
Qk
Zbtb20
Apoe
Neurod2
Sensitive
Weak
Perfect
Specific
Dlx1
Arx
Slc1a3
Hepacam
Tcf7l2
Slco1
Zfp36l
AUROC
log2(Fold change) of detection rate
g
Glutamatergic Pareto front
Sensitive
Perfect
Specific
Gene expression (CPM+1)
Arpp21
Sv2b
Satb2
Neurod2
Exph5
L2/3 IT
L5 IT
L5/6 NP
L5 ET
L6 CT
L6 IT
L6 IT Car3
L6b
Lamp5
Pvalb
Sst
Vip
Sncg
Non−Neuronal


a
Fig2
3 classes
13 subclasses
85 clusters
GABAergic
Glutamatergic
Non-Neuron
b
Number of genes
Hierarchy level
class
subclass
cluster
Marker type
perfect
sensitive
specific
weak
c
Number of genes
GABAergic
Glutamatergic
Non-Neuronal
d
Number of genes
L6 IT Car3
L6b
L5/6 NP
L5 ET
L6 CT
L2/3 IT
L5 IT
L6 IT
e
Number of genes
Pvalb
Lamp5
Sncg
Sst
Vip
f
Number of genes
Clusters
(ordered by number of markers)
g
Association with cluster replicability
(Spearman correlation)
perfect
sensitive
specific
weak
Marker type
h
Replicability (MetaNeighbor score)
Number of perfect markers
Lamp5 Slc35d3
Pvalb Vipr2
Sst Chodl
L6b Kcnip1_2
Lamp5 Lhx6
L6 NP Trh_1
Vip Serpinf1_2


a
Subclass level (single gene performance)
Fig3
Marker selection
meta−analytic
single dataset
Single gene coverage (AUROC)
Marker rank (by p−value)
b
Train dataset
Single dataset markers
(FC > 4, FDR < 0.05)
top N
Marker expression in test dataset
Classification performance
(cells ranked by marker expression)
AUROC = 0.83
6 train datasets
Single dataset markers
(FC > 4, FDR < 0.05)
Meta-analytic markers
(# datasets DE)
top N
Marker expression in held out dataset
Classification performance
(cells ranked by marker expression)
AUROC = 0.92
c
Class level (aggregated performance)
Marker selection
meta−analytic
single dataset
Coverage (AUROC)
Signal−to−noise ratio (FC)
1000 genes
1 gene
d
Subclass level (aggregated performance)
Marker selection
meta−analytic
single dataset
Coverage (AUROC)
Signal−to−noise ratio (FC)
e
Cluster level (aggregated performance)
Marker selection
meta−analytic
single dataset
Coverage (AUROC)
Signal−to−noise ratio (FC)
f
Optimal coverage (AUROC)
class
subclass
joint_cluster
Hierarchy level
g
Signal−to−noise ratio at optimum (FC)
class
subclass
joint_cluster
Hierarchy level
h
Number of genes at optimum
class
subclass
joint_cluster
Hierarchy level


a
Test cells ranked by marker expression
expression threshold
F1 = 0.66
Fig4
b
F1(t_opt)
Thresholds with performance > 90% of max performance
Classification accuracy (F1)
higher recall
t_opt
higher precision
Expression threshold
c
Transcriptome-wide normalization
Classification accuracy (F1)
snCv2
snCv3Z
snCv3N
snSS
scCv2
scCv3
scSS
t_meta
Average marker expression (CPM)
d
Marker-wide renormalization
Classification accuracy (F1)
Renormalized marker expression (proportion)
e
50 genes
class
subclass
joint_cluster
Performance loss marker renormalization (average F1(t_opt) − F1(t_meta))
Performance loss CPM (average F1(t_opt) − F1(t_meta))
g
Transcriptome-wide normalization
Average performance (F1)
Number of genes
h
Marker-wide renormalization
Average performance (F1)
Number of genes
f
Classes
Subclasses
Clusters
Number of genes
F1
Glutamatergic
GABAergic
Non-Neuronal
L5 ET
L6 IT Car3
L6 CT
L5/6 NP
Sst
Pvalb
Lamp5
Vip
L6b
L2/3 IT
L5 IT
L6 IT
Sncg


a
Fig5
Number of significant GO terms
Fraction of cell−type specific terms (%)
Number of genes
b
synaptic vesicle priming
regulation of neuronal synaptic plasticity
calcium ion−regulated exocytosis of neurotransmitter
forebrain neuron differentiation
regulation of synaptic transmission, GABAergic
gamma−aminobutyric acid transport
regulation of dendritic spine morphogenesis
positive regulation of dendrite morphogenesis
chemical synaptic transmission, postsynaptic
Non−Neuronal
Glutamatergic
GABAergic
−log10(FDR)
log2(OR)
c
protein localization to postsynapse
protein localization to postsynaptic membrane
receptor localization to synapse
presynapse assembly
synaptic transmission, GABAergic
positive regulation of synapse assembly
glutamate receptor signaling pathway
transmission of nerve impulse
regulation of synaptic transmission, glutamatergic
regulation of NMDA receptor activity
postsynaptic specialization assembly
startle response
negative regulation of muscle contraction
regulation of presynapse organization
regulation of presynapse assembly
Vip
Sst
Sncg
Pvalb
Lamp5
log2(OR)
−log10(FDR)
d
positive regulation of nervous system process
chemical synaptic transmission, postsynaptic
transmission of nerve impulse
regulation of potassium ion transmembrane transport
positive regulation of excitatory postsynaptic potential
modulation of excitatory postsynaptic potential
negative regulation of dendrite development
synaptic membrane adhesion
regulation of neurotransmitter receptor activity
glutamate receptor signaling pathway
innervation
neuron recognition
synaptic vesicle priming
dendrite extension
regulation of synaptic transmission, glutamatergic
catecholamine secretion
L6b
L6 IT Car3
L6 IT
L6 CT
L5/6 NP
L5 IT
L5 ET
L2/3 IT
log2(OR)
−log10(FDR)

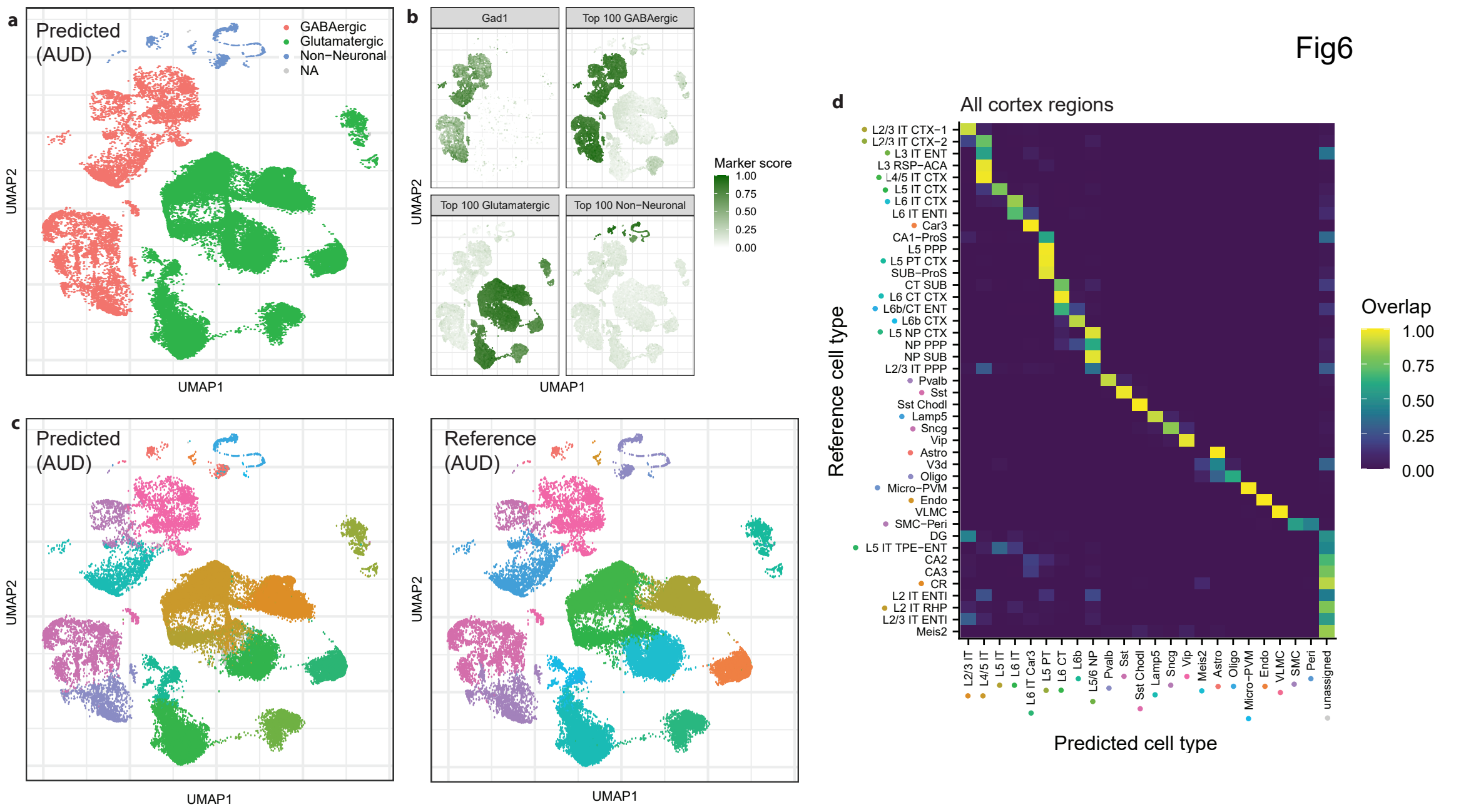


Fig6
a
Predicted (AUD)
GABAergic
Glutamatergic
Non−Neuronal
NA
UMAP2
UMAP1
b
Gad1
Top 100 GABAergic
Top 100 Glutamatergic
Top 100 Non−Neuronal
Marker score
1.00
0.75
0.50
0.25
0.00
UMAP2
UMAP1
c
Predicted (AUD)
Reference (AUD)
UMAP2
UMAP1
d
All cortex regions
Reference cell type
Predicted cell type
Overlap
1.00
0.75
0.50
0.25
0.00
L2/3 IT CTX−1
L2/3 IT CTX−2
L3 IT ENT
L3 RSP−ACA
L4/5 IT CTX
L5 IT CTX
L6 IT CTX
L6 IT ENTI
Car3
CA1−ProS
L5 PPP
L5 PT CTX
SUB−ProS
CT SUB
L6 CT CTX
L6b/CT ENT
L6b CTX
L5 NP CTX
NP PPP
NP SUB
L2/3 IT PPP
Pvalb
Sst
Sst Chodl
Lamp5
Sncg
Vip
Astro
V3d
Oligo
Micro−PVM
Endo
VLMC
SMC−Peri
DG
L5 IT TPE−ENT
CA2
CA3
CR
L2 IT ENTI
L2 IT RHP
L2/3 IT ENTI
Meis2
L2/3 IT
L4/5 IT
L5 IT
L6 IT
L6 IT Car3
L5 PT
L6 CT
L6b
L5/6 NP
Pvalb
Sst
Sst Chodl
Lamp5
Sncg
Vip
Meis2
Astro
Oligo
Micro−PVM
Endo
VLMC
SMC
Peri
unassigned