Title:

Gene panel selection for targeted spatial transcriptomics

Yida Zhang[1,2], Viktor Petukhov[1], Evan Biederstedt[1], Richard Que[3], Kun Zhang[3,4], Peter V. Kharchenko[1,4,*]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.

[2]Department of Neurobiology, Duke University, Durham, NC, USA.

[3]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA.

[4]San Diego Institute of Science, Altos Labs, San Diego, CA, USA.

[*]Corresponding author: peter_kharchenko@hms.harvard.edu

Abstract

Targeted spatial transcriptomics hold particular promise in analysis of complex tissues. Most such methods, however, measure only a limited panel of transcripts, which need to be selected in advance to inform on the cell types or processes being studied. A limitation of existing gene selection methods is that they rely on scRNA-seq data, ignoring platform effects between technologies. Here we describe gpsFISH, a computational method to perform gene selection through optimizing detection of known cell types. By modeling and adjusting for platform effects, gpsFISH outperforms other methods. Furthermore, gpsFISH can incorporate cell type hierarchies and custom gene preferences to accommodate diverse design requirements.

Key words: gene panel selection, targeted spatial transcriptomics, single cell RNA sequencing, platform effect, cell type hierarchy

Background

The building block of complex tissues is the diverse range of cell types [1–4]. Knowing the identity and spatial location of cells from different cell types is the key for understanding how they communicate with each other to carry out specific functions and how diseases emerge when this complex network of interactions goes awry [5–11]. Single-cell RNA sequencing (scRNA-seq) provides a powerful tool to study the identity of cell types and cell states [12–17]. However, the spatial information is lost due to cell disassociation during library preparation. Recent advances in spatial transcriptomics technologies have overcome this limitation by providing ways to quantify gene expression while keeping the spatial information of cells, leading to more comprehensive and detailed understanding of diseases and normal functions [18–23].

Based on the number of transcripts that can be probed, spatial transcriptomics technologies can be broadly categorized as (1) targeted, measuring a limited panel of transcripts; and (2) untargeted, capturing all transcripts from the transcriptome. Targeted spatial transcriptomics include in situ hybridization (ISH)-based [24–28] and most in situ sequencing (ISS)-based methods [29–32]. Untargeted spatial transcriptomics include next-generation sequencing (NGS)-based methods [33–39].

47 Compared to untargeted spatial transcriptomics, targeted spatial transcriptomics
48 can achieve high sensitivity and subcellular resolution. However, their targeted
49 nature requires a panel of genes (from a few hundred to thousand) to be selected in
50 advance to recognize cell types or processes relevant to the tissue being studied.
51
52 Gene selection methods are used to design gene panels. They can be classified into
53 two major categories based on their gene selection objectives. One category with an
54 imputation-based objective aims to select genes based on their ability to capture as
55 much of transcriptional variation in the scRNA-seq data as possible. Examples range
56 from simply selecting highly-variable genes to more advanced methods like L1000
57 [40], geneBasis [41], and SCMER [42]. Specifically, L1000 identified the optimal set
58 of 'landmark' transcripts that construct a reduced representations of the
59 transcriptome. geneBasis finds genes that can yield a *k*-nearest neighbor graph that
60 is similar to the "true" graph constructed using the entire transcriptome. SCMER
61 aims to select genes that preserve the manifold of scRNA-seq data. Another category
62 of gene selection method with a classification-based objective selects genes given
63 their ability to reconstruct cell classifications or relationships. Examples range from
64 selecting differentially expressed genes (DEGs) to more advanced methods like
65 scGeneFit [43], RankCorr [44], and NS-Forest [45]. scGeneFit selects marker genes
66 that jointly optimize cell type recovery using a label-aware compressive
67 classification method. RankCorr is a rank-based one-vs-all feature selection method
68 that selects marker genes for each cell type based on a sparsity parameter that
69 controls the number of marker genes selected per cell type. NS-Forest is a machine

70 learning-based marker gene selection algorithm that uses the nonlinear attributes of
71 random forest feature selection and a binary expression scoring approach to select
72 the minimal combination of marker genes that captures the cell type identity in
73 scRNA-seq data. All these methods can be used to design gene panels for targeted
74 spatial transcriptomics technologies.
75
76 A key limitation of current gene selection methods is that they select genes purely
77 based on scRNA-seq data without considering potential differences between scRNA-
78 seq and the target spatial transcriptomics technologies. Such platform effects
79 include systematic differences in capture efficiency of genes between platforms
80 caused by technology-dependent factors, including detection technique and library
81 preparation chemistry. Platform effects have been previously noted when
82 comparing gene expression measurements from single-cell and single-nucleus RNA-
83 seq on the same biological sample [46]. Platform effects also exist between scRNA-
84 seq and spatial transcriptomics technologies [47–49], posing a challenge when
85 transferring cell type information from scRNA-seq to spatial transcriptomics
86 technologies. When selecting gene panels using scRNA-seq data, such platform-
87 specific distortions can lead to reduced performance of selected gene panels in the
88 resulting spatial measurements.
89
90 Besides platform effects, there are other complications involved in gene panel
91 selection. First, current classification-based gene selection methods [43–45] treat
92 cell types as equally distinct. However, cell types are organized in a hierarchical

93 manner with cell subpopulations belonging to the same broad cell type more similar
94 to each other than subpopulations belonging to different broad cell types [50–56].
95 Depending on the biological questions and capabilities of the assays, a gene
96 selection method could optimize for fidelity at lower cell type resolution, or place
97 increased emphasis on certain subgroups of cell types. More generally, this is not
98 only useful for selecting genes that inform on cell types but can also be extended to
99 selecting genes for other biological entities with a hierarchical structure, e.g., gene
100 ontology and pathways [57,58]. Second, both imputation-based and classification-
101 based gene selection methods select genes solely based on a pre-defined objective
102 function. However, in practice of gene panel design for targeted spatial
103 transcriptomics, there can be other criteria contributing to the gene selection.
104 Examples range from technical factors, such as ability to design probes for targeting
105 certain transcripts, to biological factors such as preferences for certain pathways or
106 marker genes commonly used in the literature. A framework that takes such
107 orthogonal preferences into consideration can be helpful in practice.
108
109 To address these challenges, we developed gpsFISH, a classification-based gene
110 selection method that models and adjusts for the platform effects between scRNA-
111 seq and targeted spatial transcriptomics technologies, yielding more informative
112 gene panels and better cell type classification compared to previously published
113 classification-based gene selection methods. In addition, gpsFISH provides options
114 to account for cell type hierarchy and gene-specific custom preferences during gene

panel design, offering flexible and finer control of cell type granularity and gene selection for different biological questions.

Results

Platform effects between scRNA-seq and targeted spatial transcriptomics

Even molecule counting assays carry inherent detection biases, posing challenges for joint analysis of multiple assays, such as scRNA-seq and spatially-resolved counts [47–49]. Indeed, we observed a systematic difference of transcript detection rate across platforms (**Fig. 1A-D**), which distorts the resulting transcriptional profile estimates. Consequently, a panel of genes selected based on scRNA-seq that works well on distinguishing cell types may not achieve similar level of performance when measured by targeted spatial transcriptomics.

To address this challenge, we estimate the level of gene expression distortion in targeted spatial transcriptomics data relative to scRNA-seq and from the same tissue using a Bayesian model (**Fig. S1, Methods**). Bayesian inference estimates the posterior distribution of distortion magnitudes, which will be used to predict the potential distortion levels for genes that have not yet been observed in a given assay. Specifically, we assume platform effects are on a per gene basis. γ_i and c_i represent gene specific multiplicative and additive platform effect for each gene i, respectively. These distortion parameters are assumed to follow two normal distributions with μ_γ, μ_c as mean and σ_γ, σ_c as standard deviation, respectively. The posterior distribution of σ_γ and σ_c can be considered as a generalized description of

138 the magnitude of multiplicative and additive platform effects. We can use them to
139 sample the magnitudes of gene specific multiplicative and additive platform
140 distortions for unobserved genes. The model is fitted for a given pair of scRNA-seq
141 and targeted spatial transcriptomics platforms to account for the differences
142 between them.
143
144 To check the extent to which the model is able to capture platform biases, we used
145 three paired scRNA-seq and targeted spatial transcriptomics datasets: scRNA-seq
146 and MERFISH data from mouse hypothalamic preoptic region [24] (Moffit dataset),
147 scRNA-seq and osmFISH data from mouse cortex [26,59] (Codeluppi dataset), and
148 scRNA-seq and DARTFISH data from healthy human kidney [60] (Zhang dataset)
149 (**Methods, Table S1**). Fitting a model for each pair of datasets, we then performed
150 posterior predictive check, i.e., we simulated spatial transcriptomics measurements
151 from scRNA-seq data using the fitted Bayesian model (**Methods**). Comparisons of
152 the distribution of simulated and observed spatial transcriptomics measurements
153 demonstrated that the Bayesian model can accurately recapitulate the platform
154 effects from different pairs of technologies (**Fig. 1E, Fig. S2A-C**). The posterior
155 distributions of σ_γ and σ_c (**Fig. 1F-G**) on the three datasets showed distinct levels of
156 additive and multiplicative platform effects, indicating the need to account for
157 platform-specific properties during gene panel selection.
158
159 **Gene panel selection using genetic algorithm**
160 To take platform distortions into account during selection of the gene panels, we use

161 the platform specific Bayesian model to simulate spatial transcriptomics
162 measurements with distortions (**Methods**). The gene panels are optimized for their
163 ability to recover cell type labels from such simulated spatial measurements, rather
164 than the original scRNA-seq measurements. Such an approach is intended to
165 provide a more accurate estimation of panel performance in a real spatially-
166 resolved measurement. Instead of selecting top-performing genes, gpsFISH
167 optimizes the entire gene panel in its combined ability to recover cell type labels. To
168 optimize within this combinatorial gene space, gpsFISH uses genetic algorithm
169 optimizer [61,62] (**Fig. 2, Methods**).
170
171 Within each iteration of optimization, multiple cross validations of classification are
172 performed for each proposed gene panel. To avoid biasing towards a specific
173 realization of spatial transcriptomics distortions, gpsFISH performs the platform
174 simulations separately in each cross validation. As a result, gene panels that are
175 more robust to unexpected platform distortions will be favored. This gene panel
176 selection framework ensures the evaluation is reflective of the gene panel's real
177 classification performance when measured by specific targeted spatial
178 transcriptomics technologies.
179
180 We first tested gpsFISH on the mouse hypothalamic scRNA-seq data (Moffitt
181 dataset) with simulated platform effect by optimizing a 200 gene panel to
182 distinguish "level 1" cell type annotation, which includes 12 broadly defined cell
183 types (**Fig 3A**). Most of the cells are correctly classified, yielding an overall accuracy

184 of 0.983 and high area under the receiver-operator curve (AUC) across all cell types
185 (**Fig 3B, C**). The optimized gene panel selected with considering platform effect was
186 also more successful in separating the 12 cell types on the resulting UMAP
187 embedding compared to the gene panel selected without considering platform effect
188 (**Fig. 3D, E**).
189
190 To quantify performance of different methods we simulated spatial transcriptomics
191 data from scRNA-seq, separating training and test sets (**Methods**). Simulations
192 were performed both with and without distortions in order to evaluate how taking
193 platform effects into account impacts gene panel performance. We also compared
194 gpsFISH with two previously published classification-based gene selection methods:
195 RankCorr and scGenefit. Both methods rely on the scRNA-seq expression profiles
196 without considering platform effects. RankCorr is a rank-based one-vs-all feature
197 selection method that selects marker genes for each cell type given a sparsity
198 parameter, which controls the number of marker genes selected per cell type. We
199 tuned this parameter to make sure the panels generated using RankCorr have the
200 same size (200 genes). scGenefit selects gene markers that jointly optimize cell label
201 recovery using label-aware compressive classification methods. As control, we
202 provided a naïve way to simulate spatial transcriptomics measurements without
203 platform effects (**Methods**). In addition, we also generated a panel of randomly
204 selected genes as baseline.
205
206 The objective of gpsFISH optimization is to achieve high quality cell type

207 classification on the spatial transcriptomics data. This entails two tasks: (1)
208 selecting a good gene panel, and (2) using the gene panel for accurate cell type
209 classification. In practice, while the design of an initial gene panel may rely on the
210 scRNA-seq data, optimization of subsequent panels can take advantage of the probe-
211 specific distortions that have already been observed in earlier measurements.
212 Similarly, as more and more spatial transcriptomics data are generated, when
213 classifying cell types in a newly generated spatially-resolved measurement, it is
214 likely that some partial annotations may already be available for that platform
215 either on the current or previously acquired datasets. Regardless of the cell type
216 granularity of partial annotation, it contains gene-specific platform effect
217 information of genes in the spatial transcriptomics data, which can be estimated
218 using our Bayesian model to improve cell type classification. Following this logic, we
219 used two benchmark strategies, which evaluate the impact of platform effects on the
220 two tasks (**Methods**). Both strategies share the same general framework in which a
221 classifier is trained on the training data with gene expression profiles for all cell
222 types, and then applied onto the testing data for cell type classification. The
223 difference is how the two strategies incorporate partial annotation into the training
224 data when available. Specifically, for the first strategy, evaluation with platform
225 effect re-estimation (**Methods**), platform effects are estimated from the partial
226 annotation data and incorporated into the training data for all gene selection
227 methods. Since under this strategy the gene panels from all methods are evaluated
228 in the same manner, it is useful in evaluating the impact of platform effect on the
229 first task, i.e., selecting a good gene panel. In contrast, under the second evaluation

230 strategy, evaluation without platform effect re-estimation (**Methods**), only gpsFISH
231 panels are evaluated with platform effect estimation as described above (**Methods**),
232 illustrating the impact of platform effects on both tasks.
233
234 Evaluation with platform effect re-estimation on the Moffit dataset using naïve
235 Bayes as the classifier shows gpsFISH outperforms the control with naïve simulation
236 and other gene selection methods (**Fig. 4A**), indicating that taking platform effects
237 into consideration leads to more informative gene panels. Similar results were
238 observed for the Zhang and Codeluppi dataset (**Fig. S3A and S3C**) and using random
239 forest as classifier (**Fig. S4A-C**). From the normalized confusion matrix of the gene
240 panel selected by gpsFISH with hierarchical tree on the left showing the relationship
241 between cell types (**Fig. 4C**) we can see that most of the misclassifications are
242 within the complex subpopulations of inhibitory and excitatory neurons.
243
244 A larger performance improvement of gpsFISH over other gene selection methods is
245 observed using evaluation without platform effect re-estimation, especially when
246 the level of partial annotation is low (**Fig. 4B, Fig. S3B and S3D, Fig. S4D-F**),
247 indicating that considering platform effects can lead to more accurate cell type
248 classification.
249
250 Overall, the comparison results show that gpsFISH outperforms other gene selection
251 methods and considering platform effects can result in more informative gene
252 panels and better cell type classification.

253

254 **Redundancy in gene space across independent gene panel optimizations**

255 **enables incorporation of customized preferences**

256 Independent panel optimizations performed multiple times (10) for each of the

257 three datasets showed high level of redundancy in the gene space (**Fig. 5A**).

258 Specifically, despite similar levels of overall performance, the overlap between

259 independently optimized 200-gene panels was around 85, 65, and 35 genes, and

260 more than 20%, 30%, and 45% of the genes showed up in only one of the 10

261 optimized gene panels for the Zhang, Moffit, and Codeluppi datasets, respectively

262 (**Fig. S5A-C**). We observed similar level of redundancy even when the optimization

263 was performed for a more granular "level 2" cell type annotations (46, 87, and 47

264 cell types for Zhang, Moffit, and Codeluppi dataset, **Fig. S5D**). The ability to achieve

265 similar level of performance with different gene sets suggests that the panels can be

266 further optimized to accommodate secondary criteria, such as inclusion of pre-

267 selected genes, emphasis on genes with specific features or from specific pathways,

268 etc.

269

270 gpsFISH allows to incorporate secondary preferences during gene panel

271 optimization, by specifying custom gene weights. To illustrate how panel

272 redundancy can be used to incorporate secondary preferences with little impact on

273 the classification performance, we evaluated the ability to increase the number of

274 technical probes per gene. Specifically, many ISH-based assays, including DARTFISH,

275 can include multiple different probes to enhance detection of any given transcript.

276 The number of probes that can be designed to target each gene is determined by
277 gene-specific factors like gene length. Genes with more probes are preferred, as they
278 can be used to improve robustness and sensitivity of detection. To generate a gene
279 panel with high number of potential probes, we used the predicted number of
280 probes for each gene in the DARTFISH data (Zhang dataset) (**Methods**) as gene
281 weight during gene panel selection. Of note, we capped the probe count at 15 to
282 avoid bias towards a small portion of genes with extremely high number of probes
283 (**Fig. S6A**). This also agrees with the fact that sensitivity will saturate when we have
284 enough probes for a gene.
285
286 Following this approach, we performed 10 optimizations with and without probe
287 count gene weights on the Zhang dataset using "level 1" cell type annotations. As
288 expected, the optimizations with gene weight had slightly lower accuracy (**Fig. 5B**)
289 but achieved a significantly higher number of total probes (**Fig. 5C**). This
290 demonstrates that the redundancy of gene spaces allows one to incorporate
291 additional customized constraints/preferences based on orthogonal information to
292 design gene panels with preferred features without sacrificing the overall cell type
293 classification performance.
294
295 **Hierarchical gene selection based on cell type hierarchy**
296 Cell types are organized in a hierarchical manner with broad cell types divided into
297 more detailed subpopulations. This hierarchical relationship can be considered
298 when evaluating cell classification errors. For example, failure to distinguish two

299 closely related subtypes, such as Th1 and Th17, of cells is likely to be considered
300 less severe than mis-annotation of a Th cell into a different major cell type such as B
301 cells.
302
303 In addition to the default "flat" cell type evaluation, gpsFISH therefore, implements a
304 hierarchical classification option (**Fig. 6A, Methods**), in which correct classifications
305 or misclassification between different cell types will receive different credit/penalty
306 specified by a weighted penalty matrix according to cell type hierarchy. Using this
307 hierarchical classification framework, gpsFISH provides flexibility to customize
308 optimization based on desired level of cell type granularity.
309
310 To evaluate the effect of the hierarchical classification for gene selection, we
311 performed hierarchical gene selection at level 2 cell annotation of all three datasets.
312 Under a hierarchical penalty scheme, misclassifications of cells between different
313 level 1 categories incur a fixed penalty, whereas misclassifications within the same
314 level 1 category were given partial credit, proportional to the expression similarity
315 between the called and true subtypes (**Methods, Fig. 6B**). To quantify to what
316 extent this hierarchical classification framework reduces misclassifications across
317 broad cell types at level 1, we calculated the percentage of across broad cell type
318 mistakes over all mistakes (**Methods**). We observed that the optimized gene panels
319 using hierarchical classification tend to make significantly fewer misclassifications
320 across broad cell types at level 1 compared to flat classification (**Fig. 6C-E**),

321 indicating that cell type granularity can be controlled through the hierarchical
322 classification framework.
323

324 Discussion and conclusions

325 Accurate cell type classification is crucial for understanding the spatial relationship
326 of cells in complex tissues. We implemented gpsFISH, a method for gene panel
327 design of targeted spatial transcriptomics. By accounting for platform effects
328 between scRNA-seq and targeted spatial transcriptomics technologies, gpsFISH is
329 able to find more robust and informative gene panels and achieve better cell type
330 classification.
331
332 Different technology has different patterns of platform effects. Specifically, we
333 decomposed platform effects into two components: multiplicative and additive
334 platform effects. While the multiplicative effect has been considered in
335 deconvolution contexts (e.g., RCTD [47]), neither type of platform-specific
336 distortions have been considered by other gene selection methods. Among other
337 things, the additive platform effect enables gpsFISH to describe situations where
338 specific genes show no expression in scRNA-seq data, but is detected in spatial
339 transcriptomics data (dots forming the vertical line in **Fig. 1A** and **1B**). This
340 observation is common for osmFISH (Codeluppi dataset) and MERFISH (Moffit
341 dataset), and cannot be modelled using only multiplicative platform effect.
342

343 Comparing the three targeted spatial transcriptomics platforms, we found highest
344 levels of additive platform effects in DARTFISH, followed by osmFISH and MERFISH
345 (**Fig. 1G**). More specifically, DARTFISH had the lowest μ_c, indicating the highest level
346 of signal reduction compared to MERFISH and osmFISH (**Fig. S2E**). Signal reduction
347 increases the possibility of good marker genes from scRNA-seq losing cell type
348 specificity in spatial transcriptomics data (dots forming the horizontal line in **Fig.**
349 **1C**), which is a main scenario where platform effects affect gene panel selection.
350 Higher level of signal reduction for DARTFISH agrees with our result that the
351 performance improvement of gpsFISH over other gene selection methods is the
352 largest in the Zhang dataset compared to the other two datasets, indicating the
353 necessity to account for additive platform effects, especially for targeted spatial
354 transcriptomics technologies with higher level of signal reduction.
355
356 In addition to additive platform effect, multiplicative platform effect also contributes
357 to the systematic difference of transcripts detection rate across technologies, posing
358 a challenge when transferring cell type information from scRNA-seq to spatial
359 transcriptomics technologies. Comparison of three targeted spatial transcriptomics
360 platforms shows osmFISH has the highest level of multiplicative platform effect,
361 followed by MERFISH and then DARTFISH (**Fig. 1F**). Higher level of multiplicative
362 platform effect leads to poorer cell type classification when there is no or low level
363 of partial annotation compared to high level of partial annotation (**Fig. 4A and 4B**,
364 **Fig. S3 and S4**), especially for evaluation without platform effect re-estimation due
365 to distorted expression profiles between scRNA-seq and targeted spatial

366 transcriptomics technologies. For evaluation with platform effect re-estimation, low
367 level of partial annotation provided limited statistical power to accurately estimate
368 gene specific platform effects, thus not able to increase the classification
369 performance. This reduced performance is gone when we have more than one cell
370 type included in the partial annotation, indicating partial annotation of a few cell
371 types is enough to enhance cell type classification if multiplicative platform effects
372 are accounted for.
373
374 Redundancy across independent optimizations allows incorporation of customized
375 preferences into gene selection. However, gene weight needs to be carefully
376 specified to ensure no sacrifice on overall gene panel performance. For the result in
377 **Fig. 5B** and **5C**, we capped the number of probes for each gene at 15. For cutoffs
378 lower than 15, gene weight difference between genes are small, leading to gene
379 panels with similar performance but also similar total number of probes. However,
380 for cutoffs higher than 15, the optimization will bias towards a small group of genes
381 with high probe count, resulting in local minimum during optimization (**Fig. S6B-C**).
382 This does achieve panels with significantly higher total number of probes, but the
383 classification accuracy is dropped. This emphasizes the need to test different ways
384 for gene weight specification in order to get the expected result without sacrificing
385 performance.
386
387 Similarly, in our test of hierarchical gene selection, we specified the weighted
388 penalty matrix directly from cell type hierarchy. Although we reduced

389 misclassifications across broad cell types, the overall accuracy is slightly lower than
390 flat classification (**Fig. S7**). This shows that partial credit of misclassifications needs
391 to be given carefully, especially when there are many similar subpopulations within
392 the same broad cell type like in the Moffit dataset. In real usage, it is suggested to
393 prune the weighted penalty matrix constructed from the cell type hierarchy to
394 remove unnecessary partial credit. Gene panel selection using flat classification can
395 be run first to help adjust the weighted penalty matrix constructed using cell type
396 hierarchy. In addition, the hierarchical classification provides a generic framework
397 to fine tune emphasis of classification on certain cell types. Here we showed its
398 usage to incorporate cell type hierarchy, but it is not restricted to cell type
399 hierarchy. Customized weighted penalty matrix can be constructed using other
400 information that provides preferences towards different classifications.
401
402 A major goal of spatial transcriptomics is to understand the spatial distribution of
403 cell types and their corresponding cellular environment. gpsFISH facilitates this by
404 selecting more informative and robust gene panels and providing ways for better
405 cell type annotation. We also provide options to account for various custom
406 preferences. As more targeted spatial transcriptomics data are generated, we expect
407 that gpsFISH can facilitate the study of cellular organization of complex tissues
408 under different biological contexts.
409
410 **Methods**

411 **Datasets**

412 In our study, we used three datasets that have both scRNA-seq and targeted spatial
413 transcriptomics data from the same tissue. Information regarding the three datasets
414 is summarized in **Table S1**. Further processing details are discussed below.

415

416 ***Moffit dataset***

417 scRNA-seq data was downloaded from Gene Expression Omnibus (GEO) [63] under
418 accession code GSE113576. MERFISH data was downloaded from Dryad [64]. Of
419 note, the MERFISH data from Dryad is normalized and batch corrected. We undid
420 the volume normalization and batch correction to get the original data.

421

422 In the scRNA-seq data, we first filtered out cells annotated as "Ambiguous" and
423 "Unstable". We then used information in the supplementary Table 1 of the original
424 study to assign cell types. "Cell class (determined from clustering of all cells)"
425 column was used as level 1 cell type annotation. "Neuronal cluster (determined
426 from clustering of inhibitory or excitatory neurons)" and "Non-neuronal cluster
427 (determined from clustering of all cells)" were used as level 2 cell annotation.
428 Normalization was performed as described in the original study.

429

430 Only MERFISH data from naïve mice was used (to match scRNA-seq data). In
431 addition, we also filtered out cells annotated as "Ambiguous" and "Unstable". Fos
432 gene and five blank genes were filtered out. 135 genes imaged in the combinatorial
433 smFISH imaging were kept. Following the naming of cell types in Fig. 3D of the

434 original study, we modified the cell type annotation of MERFISH data to make it
435 consistent with the scRNA-seq data. Specifically, at level 1 cell type annotation, cells
436 annotated as "Endothelial 1", "Endothelial 2", "Endothelial 3" were merged into
437 “Endothelial”. “Astrocyte” was changed to “Astrocytes”. "OD Immature 1" and "OD
438 Immature 2" were changed to “Immature_oligodendrocyte”. "OD Mature 1", "OD
439 Mature 2", "OD Mature 3", and "OD Mature 4" were changed to
440 “Mature_oligodendrocyte”. “Pericytes” was changed to “Mural”. At cell type level 2,
441 “Endothelial 1”, “Endothelial 2”, and “Endothelial 3” were changed to
442 “Endothelial_1”, “Endothelial_2”, and “Endothelial_3”, respectively. “Ependymal” was
443 changed to “Ependymal_1”. “OD Immature 1” and “OD Immature 2” were changed to
444 “Immature_oligodendrocyte_1” and “Immature_oligodendrocyte_2”, respectively.
445 “OD Mature 1”, “OD Mature 2”, “OD Mature 3”, and “OD Mature 4” were changed to
446 “Mature_oligodendrocyte_1”, “Mature_oligodendrocyte_2”,
447 “Mature_oligodendrocyte_3”, and “Mature_oligodendrocyte_4”, respectively.
448
449 After the processing above, additional filters were applied on the raw and
450 normalized scRNA-seq data before gene panel selection. Genes with maximum cell
451 type average expression lower than 1 were filtered out. In addition, long non-coding
452 RNAs were also removed. As a result, 2886 and 5100 genes were used for gene
453 panel selection at level 1 and 2, respectively. For platform effects estimation, the
454 subset of the raw scRNA-seq and MERFISH data with cells from overlapping cell
455 types were used.
456

457 ***Codeluppi dataset***

458 scRNA-seq data was downloaded from GEO under accession code GSE60361.
459 Annotation data was downloaded from [65]. osmFISH and corresponding
460 annotation data was downloaded from [66].

461

462 For scRNA-seq data, cell labels in row 9 of the annotation were used as level 1 cell
463 type annotation, and row 11 were used as level 2 cell type annotation. However, the
464 level 1 cell type annotation is too broad (only 5 major cell types). Therefore, we
465 regenerated level 1 cell type annotation by merging similar cell types at level 2
466 following descriptions from the original study. Specifically, in generating data for
467 gene panel selection at level 1, "S1PyrDL", "S1PyrL23", "S1PyrL4", "S1PyrL5",
468 "S1PyrL5a", "S1PyrL6", S1PyrL6b", "ClauPyr" were merged into "S1_Excitatory".
469 “CA1Pyr1", "CA1Pyr2", "CA1PyrInt", "CA2Pyr2", "SubPyr” were merged into
470 "Hippocampus_Excitatory". 16 subclasses of interneurons (“Int1” to “Int16”) were
471 merged into "Interneuron". "Astro1" and "Astro2" were merged into "Astrocyte".
472 "Mgl1" and "Mgl2" were merged into "Microglia". "Pvm1" and "Pvm2" were merged
473 into "Pvm". Six subpopulations of oligodendrocytes ("Oligo1" to "Oligo6") were
474 merged into "Oligodendorcyte". "Vend1" and "Vend2" were merged into
475 "Endothelial". To make cell type labels consistent between scRNA-seq and osmFISH,
476 "Peric" was changed to "Pericyte". "Choroid" was changed to "Ventricle". "Epend"
477 was changed to "Ependymal".

478

479 To generate the data for platform effect estimation, cell type labels were modified

480 slightly differently to reflect the correspondence between cell types in scRNA-seq
481 and osmFISH as shown in Fig. 2C and Fig. 2D of the original study. Specifically, three
482 CA1 subclasses ("CA1Pyr1", "CA1Pyr2", "CA1PyrInt") were merged into
483 "Hippocampus_Excitatory". 16 subclasses of interneurons ("Int1" to "Int16") were
484 merged into "Interneuron". Two subclasses of microglia ("Mgl1" and "Mgl2") were
485 merged into "Microglia". Two subclasses of perivascular macrophages ("Pvm1" and
486 "Pvm2") were merged into "Pvm". Subclasses of S1 pyramidal cells were also
487 merged: "S1PyrL4" and "S1PyrL5a" were merged into "S1_Excitatory_L45a",
488 "S1PyrL5" and "S1PyrL6b" were merged into "S1_Excitatory_L56b". In addition, to
489 make the cell type labels consistent between scRNA-seq and osmFISH, we changed
490 "Astro1" and "Astro2" to "Astrocyte1" and "Astrocyte2", respectively. We changed
491 "Oligo6" to "Oligo_Mature", "Oligo5" to "Oligo_MF", "Oligo1", to "Oligo_COP", "Vend1"
492 to "Endothelial1", "Vend2" to "Endothelial2", "Peric" to "Pericyte", "Choroid" to
493 "Ventricle", "Ependy" to "Ependymal", "S1PyrL23" to "S1_Excitatory_L23", and
494 "S1PyrL6" to "S1_Excitatory_L6". Cell types with fewer than 50 cells were removed.
495
496 For osmFISH data, we first filtered out invalid cells based on the "Valid" column of
497 the annotation data. Then, similar to scRNA-seq data, we modified cell type labels
498 according to Fig. 2C and Fig.2D in the original study, which shows correspondence
499 between cell types in scRNA-seq and osmFISH. Specifically, "Astrocyte Gfap" was
500 changed to "Astrocyte1". "Astrocyte Mfge8" was changed to "Astrocyte2".
501 "Hippocampus" was changed to "Hippocampus_Excitatory". "pyramidal L4" was
502 changed to "S1_Excitatory_L45a". "Pyramidal L5" was changed to

503 “S1_Excitatory_L56b”. “Pyramidal L6” was changed to “S1_Excitatory_L6”.
504 “Perivascular Macrophages” was changed to “Pvm”. "Oligodendrocyte COP" was
505 changed to “Oligo_COP”. “"Oligodendrocyte Mature"” was changed to
506 “Oligo_Mature”. “Oligodendrocyte MF” was changed to “Oligo_MF”. “Endothelial 1”
507 was changed to “Endothelial1”, and “Endothelial” was changed to “Endothelial2”.
508 “Pericytes” was changed to “Pericyte”. “Vascular Smooth Muscle” was changed to
509 “Vsmc”, “C. Plexus” was changed to “Ventricle”. "Pyramidal L2-3" and "Pyramidal L2-
510 3 L5" were merged into “S1_Excitatory_L23”. "Inhibitory Cnr1", "Inhibitory CP",
511 "Inhibitory Crhbp", "Inhibitory IC", "Inhibitory Kcnip2", "Inhibitory Pthlh", and
512 "Inhibitory Vip" were merged into “Interneuron”.
513
514 scRNA-seq data was normalized using the count_normalize function in the scran
515 package. Similar to the Moffit dataset, the raw and normalized scRNA-seq were
516 further filtered before gene panel selection using the same filters.
517 6123 and 9052 genes were used for gene panel selection at level 1 and 2,
518 respectively. For platform effect estimation, the subset of the raw scRNA-seq and
519 osmFISH data with cells from overlapping cell types were used.
520
521 ***Zhang dataset***
522 Raw and normalized scRNA-seq data from kidney were obtained from [60]. They
523 were further filtered before gene panel selection using the same filters. 2920 and
524 3796 genes were used for gene panel selection at level 1 and 2, respectively.

525 The DARTFISH data is unpublished. It can be found in Zenodo [67]. We annotated
526 the cells in the DARTFISH data manually using curated marker genes (**Table S2**) at
527 subclass level (third column). For platform effect estimation, the subset of the raw
528 scRNA-seq and DARTFISH data with cells from overlapping cell types were used.
529
530 **Platform effects estimation using a Bayesian model**
531 We assume the observed number of molecules y_{ij} in the spatial transcriptomics
532 data for gene i in cell j follows a zero-inflated negative bimonial (ZINB) distribution
533 with:
534
535 $$y_{ij} \sim ZINB(\mu_{ij}, \theta_{ij}, \pi) \qquad (1)$$
536
537 where π is the zero inflation parameter which is assumed to be constant across
538 genes and cells. μ_{ij} is the mean parameter determined by a global intercept α, true
539 expression level of gene i in cell j denoted as λ_{ij}, and the cell depth (total number of
540 molecules) of cell j from spatial transcriptomics data as CD_j^{SP}:
541
542 $$\ln(\mu_{ij}) = \alpha + \ln(\lambda_{ij}) + \ln(CD_j^{SP}) \qquad (2)$$
543
544 To account for platform effects, we assume the true expression level λ_{ij} is a random
545 variable defined by:
546

547 $$\text{logit}(\lambda_{ij}) = \gamma_i \times \sqrt{x_{ij}} + c_i \qquad (3)$$

548

549 where γ_i is a gene specific coefficient representing multiplicative platform effects,

550 and c_i is a gene specific intercept representing additive platform effects. x_{ij}

551 represents the relative expression of gene i in cell j calculated from scRNA-seq data:

552

553 $$x_{ij} = \frac{c_{ij}}{\sum_{i=1}^{N} c_{ij}}$$

554

555 where c_{ij} is the number of count for gene i in cell j from the scRNA-seq data, and N

556 is the totol number of genes. When fitting the Bayesian model, in order to match

557 measurement between scRNA-seq data and targeted spatial transcriptomics data,

558 we used cell type average relative expression to replace individual cell level relative

559 expression:

560

561 $$x_{ij,j\subset k} = x_{ik} = \frac{\sum_{j=1}^{M_k} c_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M_k} c_{ij}}$$

562

563 where M_k is the number of cells in cell type k.

564

565 For the dispersion parameter θ_{ij} of the ZINB distribution, we assume it is also

566 dependent on λ_{ij}:

567

$$\ln(\theta_{ij}) = \beta + \lambda_{ij} \qquad (4)$$

where β is the intercept.

A Beta prior distribution is assumed for π. For α, β, and c_i, we assume they follow normal distribution. γ_i is assumed to follow a log-normal distribution:

$$\pi \sim \text{Beta}(1, 1)$$

$$\alpha \sim \text{Normal}(0, \sigma_\alpha)$$

$$\beta \sim \text{Normal}(0, \sigma_\beta)$$

$$c_i \sim \text{Normal}(\mu_c, \sigma_c)$$

$$\gamma_i \sim \text{LogNormal}(\mu_\gamma, \sigma_\gamma)$$

where the hyperparameters are assumed to follow Cauchy and half Cauchy distribution:

$$\mu_c, \mu_\gamma \sim \text{Cauchy}(0, 5)$$

$$\sigma_\alpha, \sigma_\beta, \sigma_c, \sigma_\gamma \sim \text{HalfCauchy}(0, 5)$$

scRNA-seq and targeted spatial transcriptomics data from overlapping genes and overlapping cell types were used as input. Additional filters were applied on the MERFISH data to reduce the totol number of cells for more efficient estimation. Specifically, cells with cell depth lower than 100 were filtered out. Cell types with

591 fewer than 1000 cells were filtered out. Then we subsampled each cell type to keep
592 at most 1000 cells for each cell type. Variational inference in Stan was used for
593 model fitting.
594
595 **Simulation of spatial transcriptomics measurements from scRNA-seq data**
596 **with platform effects**
597 We used fitted Bayesian models to simulate spatial transcriptomics measurements
598 from scRNA-seq data. Specifically, $\alpha, \beta, \pi, \mu_c, \sigma_c, \mu_\gamma, \sigma_\gamma$ were randomly sampled from
599 their estimated posterior distribution. c_i, and γ_i were randomly sampled from their
600 corresponding normal and log normal distribution for each new gene that is not
601 observed in the data used to fit the Bayesian model. If a gene is already seen during
602 fitting the Bayesian model, we can either use the empirical c_i, and γ_i estimated
603 during model fitting (used in this study) or randomly sample them from the
604 corresponding normal and log norml distribution. CD_j^{SP} was randomly sampled
605 from empirical cell depth distribution from observed targeted spatial
606 transcriptomics data. x_{ij} was calculated from scRNA-seq data. It can be cell type
607 average as we used in model fitting or calculated within each individual cell. In our
608 study, the latter was used when simulating spatial transcriptomics measurements to
609 maintain the cell level heterogenity in scRNA-seq data. Finally, the generated values
610 were plugged into equations (1), (2), (3), and (4) to generate spatial transcriptomics
611 measurements.
612

613 **Simulation of spatial transcriptomics measurements from scRNA-seq data**
614 **without platform effects (naïve simulation)**
615 We provided a naïve way to simulate spatial transcriptomics measurements without
616 platform effects as control. During the simulation without platform effects, cell
617 depth of simulated spatial transcriptomics cell were randomly sampled from the
618 empirical cell depth distribution of observed targeted spatial transcriptomics data.
619 Of note, the empirical cell depth distribution was adjusted proportionally based on
620 the ratio between relative expression of new genes for simulation and relative
621 expression of overlapping genes used in fitting the Bayesian model. After having the
622 simulated cell depth for each cell, the number of molecules for each gene within
623 each cell was sampled from a multinomial distribution with size equal to the
624 simulated cell depth and probability equal to each gene's relative expression in that
625 cell. At the end, genes were randomly selected given the probe failure rate. Then,
626 simulated molecule count of selected genes were set to 0 to reflect probe failure.
627
628 **Genetic algorithm for gene panel selection**
629 We used genetic algorithm as the framework for gene panel selection. Each
630 individual in a population is one candidate gene panel. We set the gene panel size to
631 200 genes. Each population contains 200 individuals.
632
633 The first step of genetic algorithm is to initialize a population of candidate gene
634 panels. The genes can be either randomly selected from all candidate genes or
635 selected based on their differential expression between cell types. In this study, we

636 took a hybrid approach. 95% of the 200 gene panels were initiated randomly from
637 all candidate genes to maintain population diversity. The rest 5% were initialized
638 using DEGs for each cell type. DE analysis was performed using Pagoda2. Genes
639 with AUC greater than 0.7 were considered significant.
640
641 The second step is to evaluate the fitness of each candidate gene panel in the
642 population. Here we define fitness as the average classification accuracy over 5
643 cross validations. Classification was performed on simulated spatial transcriptomics
644 measurements from scRNA-seq data. Cell type annotation from scRNA-seq data was
645 used as ground truth. The accuracy was calculated based on the original confusion
646 matrix for flat classification and weighted confusion matrix for hierarchical
647 classification. We provided two classifiers, random forest and naïve Bayes. In this
648 study we used naïve Bayes due to its fast speed and relatively similar level of
649 accuracy compared to random forest. To improve the efficiency, scRNA-seq data was
650 subsampled to reduce the number of cells for large cell types and resampled to
651 increase the number of cells for small cell types. Specifically, for level 1 cell type
652 annotation, cell type size was capped at 1500 cells. The lower bound was set as
653 1000 cells for Moffit dataset and 500 for Zhang and Codeluppi dataset. For level 2
654 cell type annotation, 250 and 500 were used as the cell type size range for Moffit
655 dataset. The range for Zhang and Codeluppi dataset was 300 and 900.
656
657 The third step is selection and mutation. The selection strategy we used is
658 tournaments. Specifically, randomly selected candidate gene panels face each other

659 1 vs. 1. The one with a higher fitness value was used as parent. In addition,
660 candidate gene panels with higher fitness values were more likely to be selected in
661 the tournaments. After having the parent gene panels, uniform crossover was
662 performed to generate the offspring gene panels. Duplicated genes after uniform
663 crossover were replaced by randomly sampled genes in the parent candidate gene
664 panels but not in the offspring gene panel. Mutation was then performed to maintain
665 gene diversity and prevent premature convergence. We set the mutation rate to 1%.
666 When gene weight was provided, genes with higher weight were (1) more likely to
667 be selected during crossover, (2) less likely to be mutated if it is already in the
668 population, (3) and more likely to be introduced into the population through
669 mutation if it is not in the current population.
670
671 Finally, the same process was repeated for the offspring population. The candidate
672 gene panel with the highest fitness value for one iteration was considered as the
673 optimal gene panel. If the iteration after it has a candidate gene panel with higher
674 fitness value, the optimal panel will be replaced by this new candidate gene panel.
675 Otherwise, the optimal panel will stay the same. The iterative process will end either
676 when it reaches a given number of iterations, or the accuracy doesn't improve more
677 than a threshold for a given number of iterations. In our study, we ran all the
678 optimizations for at least 500 iterations to ensure convergence although in all cases
679 the optimization converged a lot earlier.
680

681 If a list of pre-selected genes, e.g., canonical marker genes based on previous
682 knowledge, is provided, genes in the list will be included in each candidate gene
683 panel as well as the final optimal gene panel.
684
685 **Hierarchical classification using cell type hierarchy**
686 During genetic algorithm optimization, a weighted penalty matrix can be provided
687 to assign partial credit or extra penalty to classification between certain cell types.
688 The weighted penalty matrix is a square matrix with each row and each column
689 representing one cell type. For each value p_{ij} $(i \neq j)$ in the weighted penalty matrix,
690 if $p_{ij} > 1$, an extra penalty is given to misclassifying cells from cell type j to cell type
691 i. If $p_{ij} < 1$, a partial credit is given to misclassifying cells from cell type j to cell type
692 i. $p_{ij} = 1$ means no penalty or partial credit. In hierarchical classification, the
693 weighted penalty matrix was incorporated to the confusion matrix by element-wise
694 multiplication to provide a weighted confusion matrix. The accuracy of the weighted
695 confusion matrix was used to evaluate the fitness of candidate gene panels.
696
697 Essentially, the weighted penalty matrix can be constructed arbitrarily by user's
698 preference. In this study, we constructed the weighted penalty matrix from cell type
699 hierarchy. First, pairwise distance between cell types was calculated. Specifically,
700 average expression profile of each cell type was calculated using normalized count
701 by taking average expression of all cells in each cell type. Top 1000 genes with
702 highest standard deviation were used to calculate pairwise Pearson correlation
703 coefficient. One minus the pairwise Pearson correlation coefficient was used as

704 pairwise distance between cell types. Second, the pairwise distance matrix was
705 normalized by the largest distance so the values range from 0 to 1. Third, the
706 pairwise distance matrix was then adjusted based on cell type hierarchy.
707 Specifically, a level of cell type annotation was selected as reference. For cell types
708 below the reference level that are from the same cell type at the reference level, the
709 pairwise distance (between 0 and 1) between them was kept unchanged to reflect
710 partial credit to wrong classifications among them. For cell types below the
711 reference level that are from different cell types at the reference level, the pairwise
712 distance between them was changed to a user defined value where 1 means no extra
713 penalty and greater than 1 means extra penalty. In this study, we used 1 for no extra
714 penalty and level 1 cell type annotation was used as reference. Finally, the diagonal
715 value was changed to 1 to reflect no extra credit to correct classifications. This
716 weighted penalty matrix was used for hierarchical classification in our study.
717
718 **Calculating the percentage of across broad cell type mistakes over all mistakes**
719 We performed 5 optimizations with flat classification and hierarchical classification
720 for all three datasets, respectively. Average confusion matrix over 5 optimizations
721 for each data was calculated. After that, for each cell type, we counted the total
722 number of misclassifications and among all the misclassifications, what proportion
723 of them misclassifies cells to cell types at level 2 that don't belong to the same cell
724 type at level 1.
725
726 **Gene panel selection using RankCorr and scGeneFit**

727 The same scRNA-seq data from the three datasets after filtering were used as input.
728 For RankCorr, raw scRNA-seq data before normalization was used as suggested. The
729 lamb parameter was tuned to make sure the output marker gene list has 200 genes.
730 For scGeneFit, normalized scRNA-seq data was used by following the examples on
731 its GitHub page. Panel size was set to 200.
732
733 **Evaluation of optimized gene panel**
734 To evaluate optimized gene panels, we first simulated spatial transcriptomics
735 measurements with platform effects based on the gene panel's expression profile in
736 scRNA-seq data. Then this simulated spatial transcriptomics data was split into
737 training and testing data. The training data contains cells from a subset of cell types
738 whose cell type labels are known. This was used as the partial annotation of the
739 simulated spatial transcriptomics data. The testing data contains cells from all cell
740 types (excluding cells in the training data), which is considered as part of the
741 simulated spatial transcriptomics data that hasn't been annotated yet. We varied the
742 number of cell types in the training data from zero to all the cell types to reflect
743 different levels of partial annotation. When there was no partial annotation, scRNA-
744 seq data was used as the final training data for classifier training. When there was
745 partial annotation, information in the partial annotation was included in the final
746 training data. After that, a classifier (naïve Bayes or random forest) was trained
747 using the final training data and applied on the testing data for cell type
748 classification evaluation. Since the testing data was simulated from scRNA-seq data,
749 the cell type labels in scRNA-seq data were used as ground truth. Classification

750 accuracy was used as the metric to evaluate a gene panel. At each level of partial
751 annotation, we repeated the same calculation 10 times. To separately evaluate the
752 impact of platform effects on gene panel selection and cell type classification, within
753 the same framework described here, we designed two different strategies to
754 evaluate a gene panel by varying whether platform effect distortions that can be
755 learned from partial annotation examples are used to produce more realistic
756 training data for cell type classification.
757
758 ***Evaluation with platform effect re-estimation***
759 This evaluation strategy was designed to focus on the performance of the optimized
760 gene panels, and not on the differences in the cell type classification (evaluation)
761 stage. In this strategy, partial annotation was first used to estimate gene specific
762 platform effects using the Bayesian model. We then used these estimated gene
763 specific platform effects to simulate an updated spatial transcriptomics training
764 data, which will be combined with the partially annotated spatial transcriptomics
765 data and then used for training cell type classifiers for all the methods being
766 evaluated. Only cell types not already available in the partially annotated spatial
767 transcriptomics data were simulated. When partial annotation was available for 5 or
768 fewer cell types, the final training data combined the partially annotated and
769 simulated spatial transcriptomics training data with scRNA-seq data. When more
770 than 5 cell types were available, training was performed on the partially annotated
771 and simulated spatial transcriptomics training data only. The final training data and
772 testing data were normalized by the total number of transcripts within each cell and

773 scaled by 10000. It was then log transformed after adding 1 pseudocount. This
774 normalized training and testing data were used for classifier training and testing.
775
776 ***Evaluation without platform effect re-estimation***
777 In this evaluation strategy, only gpsFISH is able to make use of the platform effects
778 information in the partial annotation (as described above). All the other methods
779 used the partial annotation according to their own method design. Specifically, for
780 the control which used naïve simulation during gene panel selection, the empirical
781 cell depth distribution of the complete testing data was used to simulate a spatial
782 transcriptomics training data without platform effect. This simulated spatial
783 transcriptomics training data was used in the same way as described above to get
784 the final training data. For RankCorr, scGeneFit, and the random panel, since the
785 gene selection was solely based on scRNA-seq data, cells in the partial annotation
786 were directly combined with the scRNA-seq data of cell types not already available
787 in the partial annotation. The combined data were used as the final training data.
788 Same normalization was performed on the final training data and testing data
789 before classifier training and testing.
790
791 **Calculate the number of probes for each gene for the DARTFISH data**
792 During the generation of the DARTFISH data, ppDesigner [68] was used to calculate
793 the number of probes that can be designed to target each gene.
794
795 **Declarations**

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Scripts to generate data and to perform the above analysis are available in Zenodo [67].

gpsFISH's open-source code is maintained and documented on Github [69] and is publicly available under the MIT license.

Pre-fitted Bayesian models based on the Zhang, Moffit, and Codeluppi dataset respectively are deposited in Zenodo [70].

Competing interests

P.V.K. serves on the Scientific Advisory Board to Celsius Therapeutics Inc. and Biomage Inc. P.V.K. is an employee of Altos Labs.

Funding

P.V.K. and Y.Z. were supported by 5U54HL145608 grant from NIH.

Authors' contributions

P.V.K. and Y.Z. formulated the study and the overall approach. Y.Z. developed the detailed algorithms and performed the analysis with advice from P.V.K. and V.P. Y.Z. implemented the gpsFISH package with help from E.B. R.Q. and K.Z. generated the DARTFISH dataset. Y.Z. and P.V.K. drafted the manuscript.

Acknowledgements

We thank Hirak Sarkar and Teng Gao for useful discussions on the method development. We also thank Chienju Chen and Kian Kalhor for useful discussions on the DARTFISH technology and data processing. We are also grateful to Jeffrey Moffitt who provided help with obtaining the original MERFISH data before normalization and batch correction. P.V.K. serves on the Scientific Advisory Board to Celsius Therapeutics Inc. and Biomage Inc. P.V.K. is an employee of Altos Labs.

Supplementary Information

Additional file 1: Table S1.xlsx

Information of the Moffit, Codeluppi, and Zhang dataset

Additional file 2: Table S2.xlsx

Curated marker genes for the Zhang dataset

Reference

1. Arendt D. The evolution of cell types in animals: emerging principles from molecular studies. Nat Rev Genet. 2008;9:868–82.

840 2. Elmentaite R, Domínguez Conde C, Yang L, Teichmann SA. Single-cell atlases:
841 shared and tissue-specific cell types across human organs. Nat Rev Genet.
842 2022;23:395–410.

843 3. Lindeboom RGH, Regev A, Teichmann SA. Towards a Human Cell Atlas: Taking
844 Notes from the Past. Trends in Genetics. 2021;37:625–30.

845 4. Zeng H. What is a cell type and how to define it? Cell. 2022;185:2739–55.

846 5. Kölsch Y, Hahn J, Sappington A, Stemmer M, Fernandes AM, Helmbrecht TO, et al.
847 Molecular classification of zebrafish retinal ganglion cells links genes to cell types to
848 behavior. Neuron. 2021;109:645-662.e9.

849 6. Osterhout JA, Kapoor V, Eichhorn SW, Vaughn E, Moore JD, Liu D, et al. A preoptic
850 neuronal population controls fever and appetite during sickness. Nature.
851 2022;606:937–44.

852 7. Xu S, Yang H, Menon V, Lemire AL, Wang L, Henry FE, et al. Behavioral state
853 coding by molecularly defined paraventricular hypothalamic cell type ensembles.
854 Science. 2020;370:eabb2494.

855 8. Elmentaite R, Kumasaka N, Roberts K, Fleming A, Dann E, King HW, et al. Cells of
856 the human intestinal tract mapped across space and time. Nature. 2021;597:250–5.

857 9. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions
858 and communication from gene expression. Nat Rev Genet. 2021;22:71–88.

859 10. Chen W-T, Lu A, Craessaerts K, Pavie B, Sala Frigerio C, Corthout N, et al. Spatial
860 Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. Cell.
861 2020;182:976-991.e19.

862 11. Hwang WL, Jagadeesh KA, Guo JA, Hoffman HI, Yadollahpour P, Reeves JW, et al.
863 Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies
864 multicellular dynamics associated with neoadjuvant treatment. Nature Genetics
865 [Internet]. 2022; Available from: https://doi.org/10.1038/s41588-022-01134-8

866 12. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of
867 human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci USA.
868 2015;112:7285–90.

869 13. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly
870 Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter
871 Droplets. Cell. 2015;161:1202–14.

872 14. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical
873 cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016;19:335–
874 46.

875 15. Ner-Gaon H, Melchior A, Golan N, Ben-Haim Y, Shay T. JingleBells: A Repository
876 of Immune-Related Single-Cell RNA–Sequencing Datasets. JI. 2017;198:3375–9.

877 16. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res.
878 2015;25:1491–8.

879 17. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human
880 Cell Atlas. eLife. 2017;6:e27041.

881 18. Marx V. Method of the Year: spatially resolved transcriptomics. Nat Methods.
882 2021;18:9–14.

883 19. Chen R, Blosser TR, Djekidel MN, Hao J, Bhattacherjee A, Chen W, et al. Decoding
884 molecular and cellular heterogeneity of mouse nucleus accumbens. Nat Neurosci.
885 2021;24:1757–71.

886 20. Fernandez J. Molecular atlas of the adult mouse brain. SCIENCE ADVANCES.
887 2020;14.

888 21. Rao A, Barkley D, França GS, Yanai I. Exploring tissue architecture using spatial
889 transcriptomics. Nature. 2021;596:211–20.

890 22. Zhang M, Eichhorn SW, Zingg B, Yao Z, Cotter K, Zeng H, et al. Spatially resolved
891 cell atlas of the mouse primary motor cortex by MERFISH. Nature. 2021;598:137–
892 43.

893 23. Wang Y, Eddison M, Fleishman G, Weigert M, Xu S, Wang T, et al. EASI-FISH for
894 thick tissue defines lateral hypothalamus spatio-molecular organization. Cell.
895 2021;184:6361-6377.e24.

896 24. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al.
897 Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic
898 region. Science. 2018;362:eaau5324.

899 25. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly
900 multiplexed RNA profiling in single cells. Science. 2015;348:aaa6090–aaa6090.

901 26. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al.
902 Spatial organization of the somatosensory cortex revealed by osmFISH. Nat
903 Methods. 2018;15:932–5.

904 27. Shah S, Lubeck E, Zhou W, Cai L. In Situ Transcription Profiling of Single Cells
905 Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron.
906 2016;92:342–57.

907 28. Cai M, Zhang K. Spatial mapping of single cells in human cerebral cortex using
908 DARTFISH: A highly multiplexed method for in situ quantification of targeted RNA
909 transcripts. eScholarship, University of California; 2019.

910 29. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-
911 dimensional intact-tissue sequencing of single-cell transcriptional states. Science.
912 2018;361:eaat5691.

913 30. Chen X, Sun Y-C, Church GM, Lee JH, Zador AM. Efficient in situ barcode
914 sequencing using padlock probe-based BaristaSeq. Nucleic Acids Research.
915 2018;46:e22–e22.

916 31. Chen X, Sun Y-C, Zhan H, Kebschull JM, Fischer S, Matho K, et al. High-
917 Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing.
918 Cell. 2019;179:772-786.e19.

919 32. Gyllborg D, Langseth CM, Qian X, Choi E, Salas SM, Hilscher MM, et al.
920 Hybridization-based *in situ* sequencing (HybISS) for spatially resolved
921 transcriptomics in human and mouse brain tissue. Nucleic Acids Research.
922 2020;48:e112–e112.

923 33. Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Spatiotemporal transcriptomic
924 atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell.
925 2022;185:1777-1792.e21.

926 34. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al.
927 Visualization and analysis of gene expression in tissue sections by spatial
928 transcriptomics. Science. 2016;353:78–82.

929 35. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al.
930 Slide-seq: A scalable technology for measuring genome-wide expression at high
931 spatial resolution. 2019;6.

932 36. Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, et al.
933 High-definition spatial transcriptomics for in situ tissue profiling. Nat Methods.
934 2019;16:987–90.

935 37. Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, et al. High-Spatial-Resolution
936 Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell. 2020;183:1665-
937 1681.e18.

938 38. Cho C-S, Xi J, Si Y, Park S-R, Hsu J-E, Kim M, et al. Microscopic examination of
939 spatial transcriptome using Seq-Scope. Cell. 2021;184:3559-3572.e22.

940 39. Fu X, Sun L, Chen JY, Dong R, Lin Y, Palmiter RD, et al. Continuous Polony Gels for
941 Tissue Mapping with High Resolution and RNA Capture Efficiency. bioRxiv.
942 2021;2021.03.17.435795.

943 40. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next
944 Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell.
945 2017;171:1437-1452.e17.

946 41. Missarova A, Jain J, Butler A, Ghazanfar S, Stuart T, Brusko M, et al. geneBasis: an
947 iterative approach for unsupervised selection of targeted gene panels from scRNA-
948 seq. Genome Biology. 2021;22:333.

949 42. Liang S, Mohanty V, Dou J, Miao Q, Huang Y, Müftüoğlu M, et al. Single-cell
950 manifold-preserving feature selection for detecting rare cell populations. Nat
951 Comput Sci. 2021;1:374–84.

952 43. Dumitrascu B, Villar S, Mixon DG, Engelhardt BE. Optimal marker gene selection
953 for cell type discrimination in single cell analyses. Nat Commun. 2021;12:1186.

954 44. Vargo AHS, Gilbert AC. A rank-based marker selection method for high
955 throughput scRNA-seq data. BMC Bioinformatics. 2020;21:477.

956 45. Aevermann BD, Zhang Y, Novotny M, Keshk M, Bakken TE, Miller JA, et al. A
957 machine learning method for the discovery of minimum marker gene combinations
958 for cell-type identification from single-cell RNA sequencing. Genome Res.
959 2021;gr.275569.121.

960 46. Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, et al. Single-
961 nucleus and single-cell transcriptomes compared in matched cortical cell types.
962 Soriano E, editor. PLoS ONE. 2018;13:e0209648.

963 47. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust
964 decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol.
965 2022;40:517–26.

966 48. Okochi Y, Sakaguchi S, Nakae K, Kondo T, Naoki H. Model-based prediction of
967 spatial gene expression via generative linear mapping. Nat Commun. 2021;12:3731.

968 49. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández
969 Navarro J, et al. Single-cell and spatial transcriptomics enables probabilistic
970 inference of cell type topography. Commun Biol. 2020;3:565.

971 50. the FANTOM Consortium, Liang C, Forrest ARR, Wagner GP. The statistical
972 geometry of transcriptome divergence in cell-type evolution and cancer. Nat
973 Commun. 2015;6:6066.

974 51. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid
975 annotation of cell atlases. Nat Methods. 2019;16:983–6.

976 52. Tasic B. Single cell transcriptomics in neuroscience: cell classification and
977 beyond. Current Opinion in Neurobiology. 2018;50:242–9.

978 53. Zeng H, Sanes JR. Neuronal cell-type classification: challenges, opportunities and
979 the path forward. Nat Rev Neurosci. 2017;18:530–46.

980 54. Yuste R, Hawrylycz M, Aalling N, Aguilar-Valles A, Arendt D, Armañanzas R, et al.
981 A community-based transcriptomics classification and nomenclature of neocortical
982 cell types. Nat Neurosci. 2020;23:1456–68.

983 55. Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biology.
984 2005;5.

985 56. Bakken T, Cowell L, Aevermann BD, Novotny M, Hodge R, Miller JA, et al. Cell
986 type discovery and representation in the era of high-content single cell phenotyping.
987 BMC Bioinformatics. 2017;18:559.

988 57. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene
989 Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.

990 58. The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris
991 NL, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids
992 Research. 2021;49:D325–34.

993 59. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A,
994 et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-
995 seq. Science. American Association for the Advancement of Science; 2015;347:1138–
996 42.

997 60. Lake BB, Menon R, Winfree S, Hu Q, Ferreira RM, Kalhor K, et al. An atlas of
998 healthy and injured cell states and niches in the human kidney. bioRxiv.
999 2021;2021.07.28.454201.

1000 61. Goldberg DE. Genetic Algorithms in Search, Optimization and Machine Learning.
1001 1st ed. USA: Addison-Wesley Longman Publishing Co., Inc.; 1989.

1002 62. Holland JH. Adaptation in Natural and Artificial Systems: An Introductory
1003 Analysis with Applications to Biology, Control, and Artificial Intelligence [Internet].
1004 The MIT Press; 1992 [cited 2022 Jul 30]. Available from:
1005 https://doi.org/10.7551/mitpress/1090.001.0001

1006 63. Gene Expression Omnibus. www.ncbi.nlm.nih.gov/geo

1007 64. Dryad. https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248

1008 65. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A,
1009 et al. Single-cell analysis of mouse cortex. http://linnarssonlab.org/cortex/

66. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. osmFISH: Spatial organization of the somatosensory cortex revealed by cyclic smFISH. http://linnarssonlab.org/osmFISH/

67. Zhang Y, Petukhov V, Biederstedt E, Que R, Zhang K, Kharchenko PV. gpsFISH analysis code and data (Zenodo link). 2023; https://doi.org/10.5281/zenodo.7613712

68. ppDesigner: Algorithm to design Padlock Probes. http://genome-tech.ucsd.edu/public/Gen2_BSPP/ppDesigner/ppDesigner.php

69. Zhang Y, Petukhov V, Biederstedt E, Que R, Zhang K, Kharchenko PV. gpsFISH R package. GitHub. 2023; https://github.com/kharchenkolab/gpsFISH

70. Zhang Y, Petukhov V, Biederstedt E, Que R, Zhang K, Kharchenko PV. Pre-fitted Bayesian models (Zenodo link). 2023; https://doi.org/10.5281/zenodo.6946054

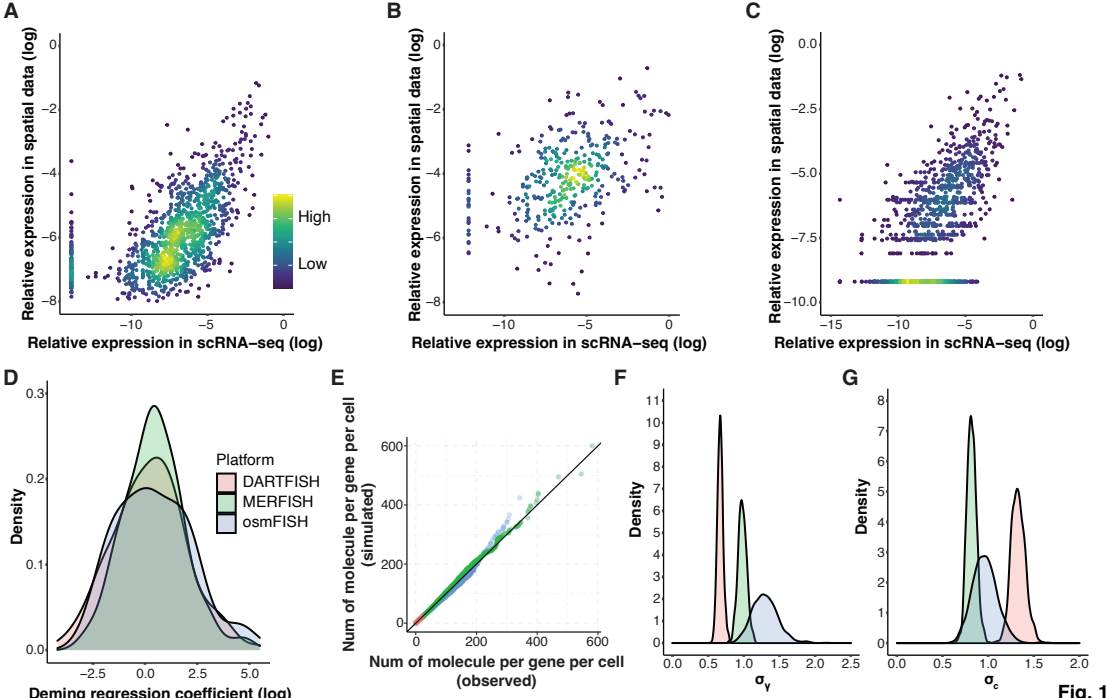

A
Relative expression in spatial data (log)
Relative expression in scRNA−seq (log)
High
Low
B
Relative expression in spatial data (log)
Relative expression in scRNA−seq (log)
C
Relative expression in spatial data (log)
Relative expression in scRNA−seq (log)
D
Density
Deming regression coefficient (log)
Platform
DARTFISH
MERFISH
osmFISH
E
Num of molecule per gene per cell (simulated)
Num of molecule per gene per cell (observed)
F
Density
σ_Y
G
Density
σ_c

Fig. 1

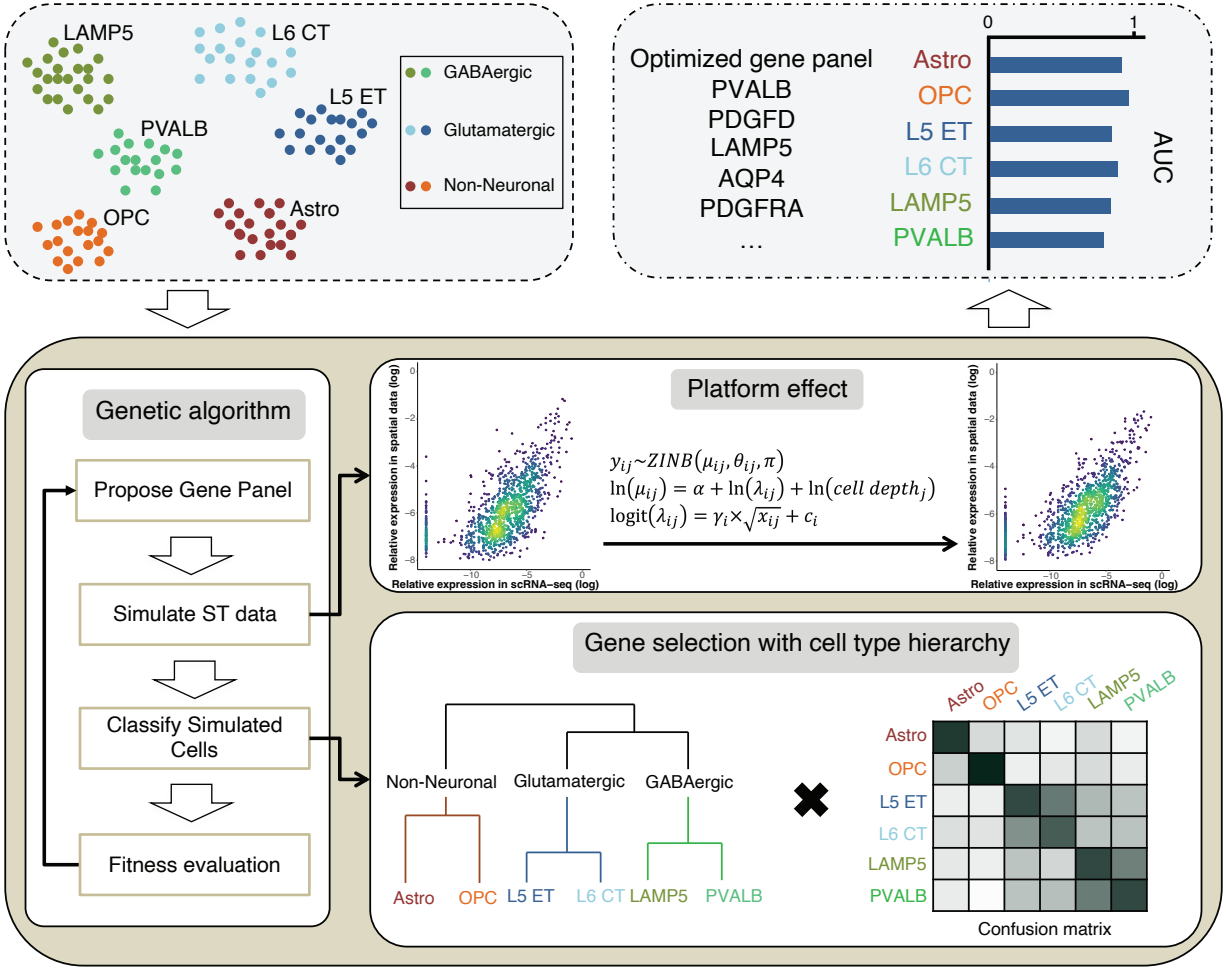

LAMP5
L6 CT
L5 ET
PVALB
OPC
Astro
GABAergic
Glutamatergic
Non-Neuronal
Optimized gene panel
PVALB
PDGFD
LAMP5
AQP4
PDGFRA
...
0
1
Astro
OPC
L5 ET
L6 CT
LAMP5
PVALB
AUC
Genetic algorithm
Propose Gene Panel
Simulate ST data
Classify Simulated Cells
Fitness evaluation
Platform effect
Relative expression in spatial data (log)
Relative expression in scRNA-seq (log)
$y_{ij} \sim ZINB(\mu_{ij}, \theta_{ij}, \pi)$
$\ln(\mu_{ij}) = \alpha + \ln(\lambda_{ij}) + \ln(cell\ depth_j)$
$\mathrm{logit}(\lambda_{ij}) = \gamma_i \times \sqrt{x_{ij}} + c_i$
Gene selection with cell type hierarchy
Non-Neuronal
Glutamatergic
GABAergic
Astro
OPC
L5 ET
L6 CT
LAMP5
PVALB
Confusion matrix

Fig. 2

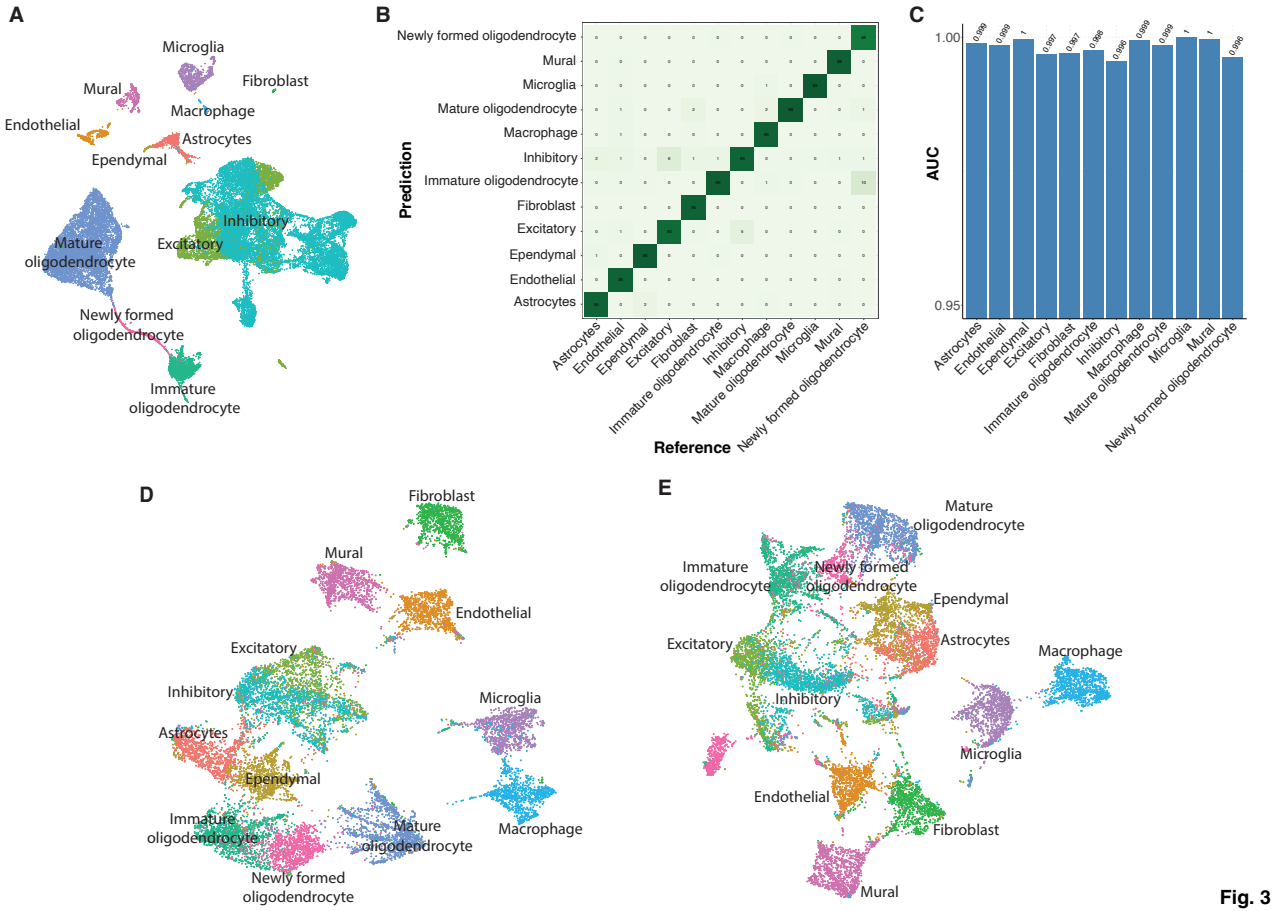

A
Microglia
Fibroblast
Mural
Macrophage
Endothelial
Astrocytes
Ependymal
Inhibitory
Mature oligodendrocyte
Excitatory
Newly formed oligodendrocyte
Immature oligodendrocyte
B
Newly formed oligodendrocyte
Mural
Microglia
Mature oligodendrocyte
Macrophage
Inhibitory
Immature oligodendrocyte
Fibroblast
Excitatory
Ependymal
Endothelial
Astrocytes
Prediction
Reference
C
AUC
1.00
0.95
D
E

Fig. 3

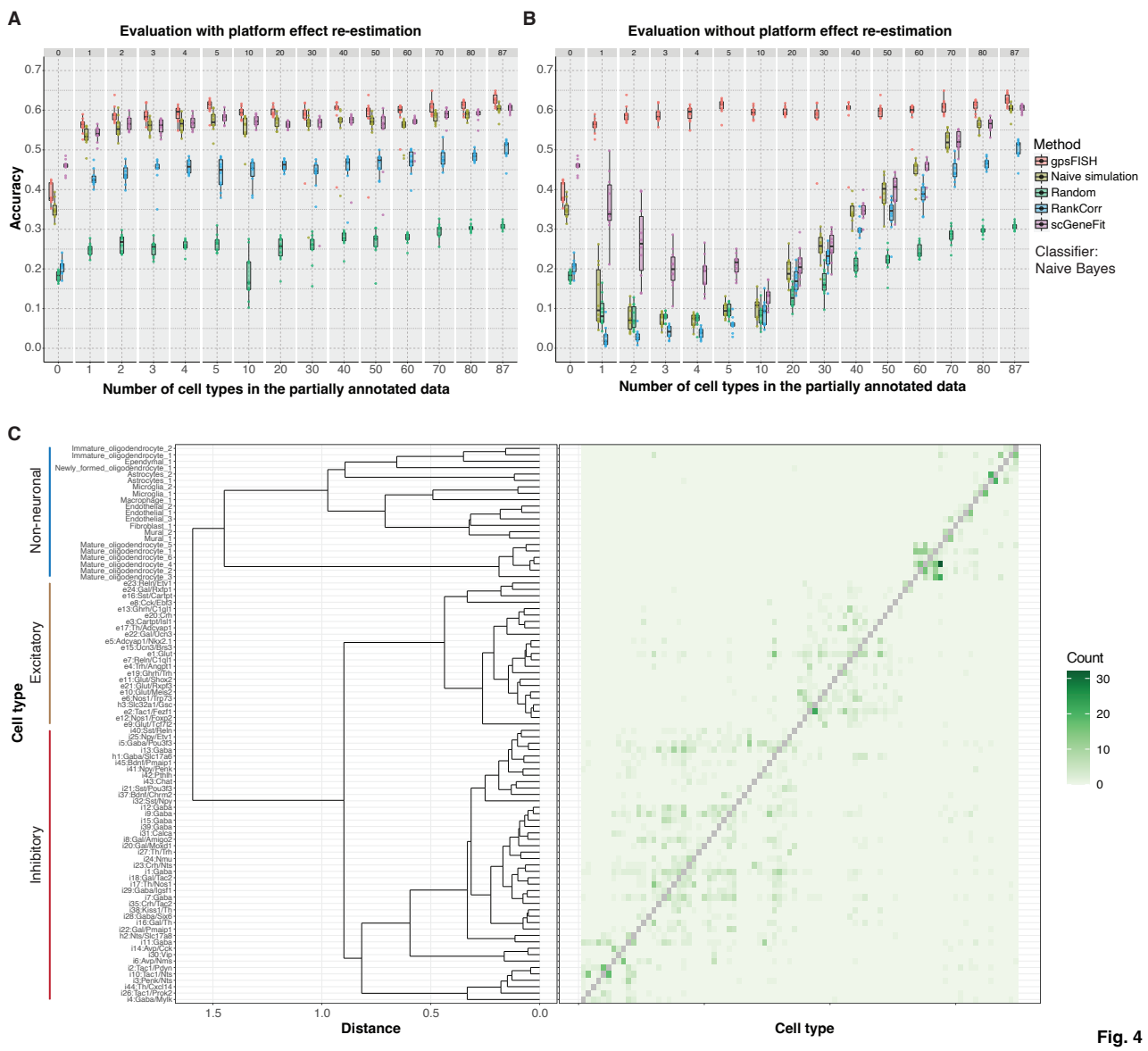

A
Evaluation with platform effect re-estimation
B
Evaluation without platform effect re-estimation
Accuracy
Number of cell types in the partially annotated data
Method
gpsFISH
Naive simulation
Random
RankCorr
scGeneFit
Classifier:
Naive Bayes
C
Non-neuronal
Excitatory
Inhibitory
Cell type
Distance
Count

Fig. 4

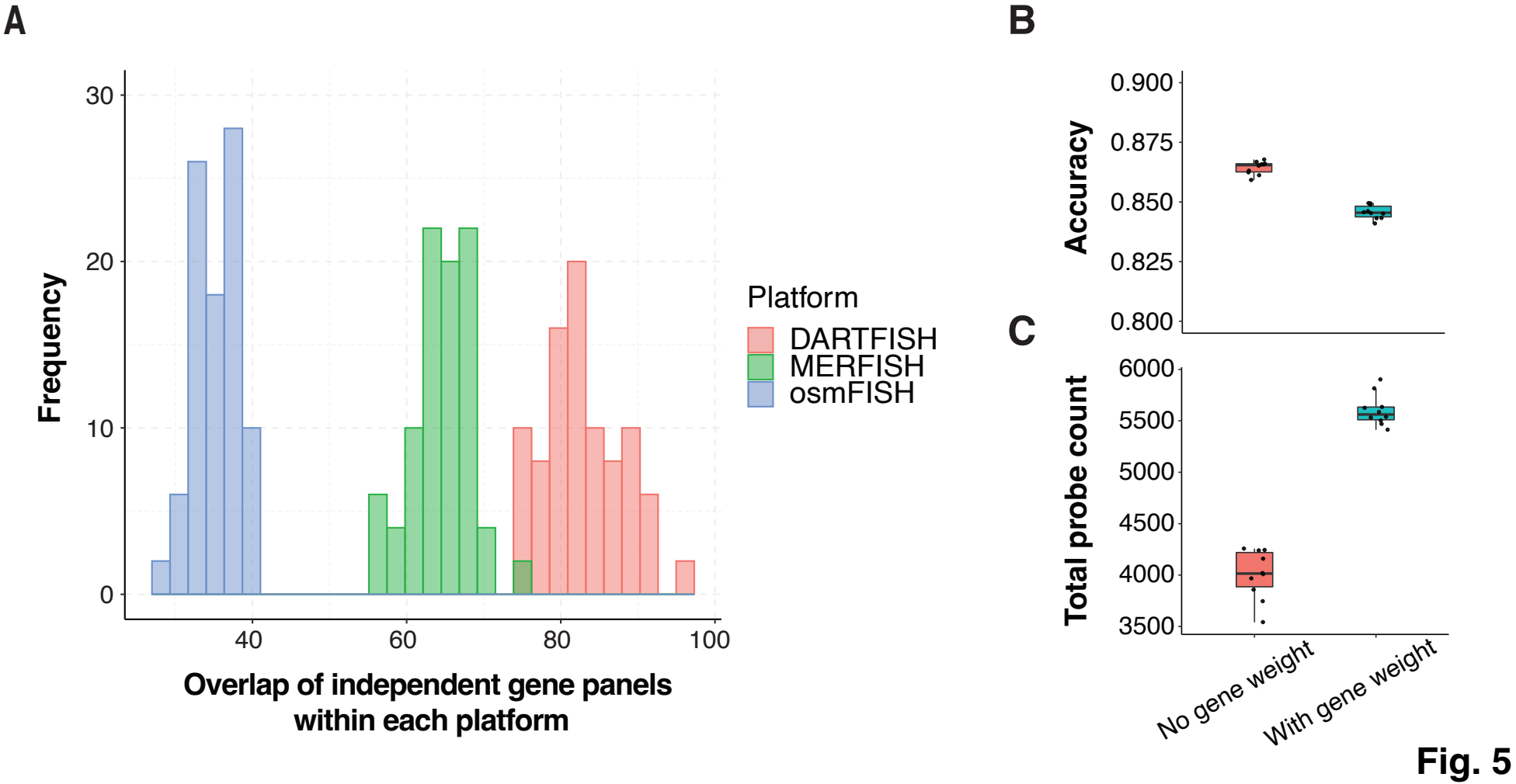

A
Frequency
30
20
10
0
40
60
80
100
Overlap of independent gene panels within each platform
Platform
DARTFISH
MERFISH
osmFISH
B
Accuracy
0.900
0.875
0.850
0.825
0.800
C
Total probe count
6000
5500
5000
4500
4000
3500
No gene weight
With gene weight

Fig. 5

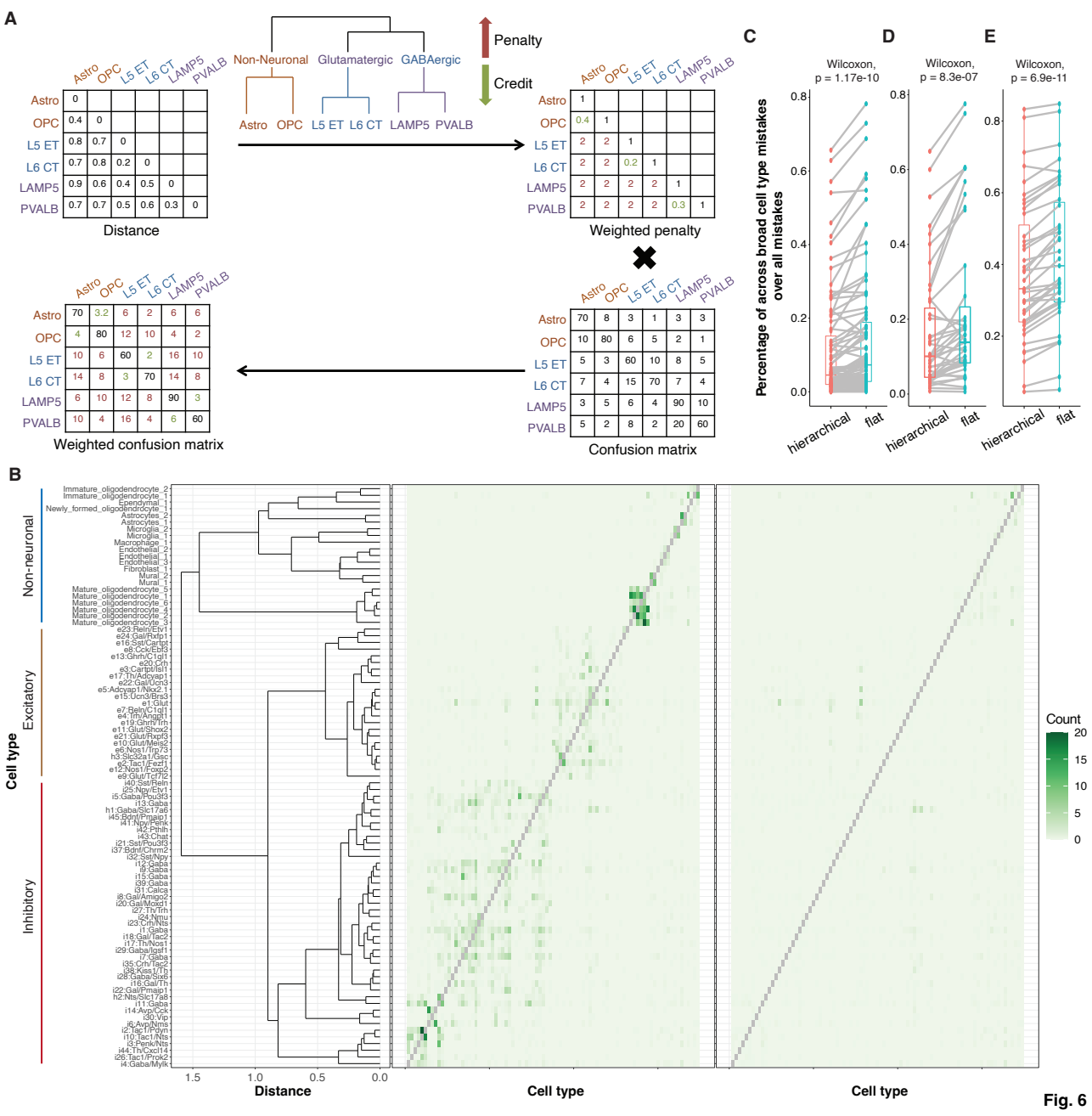
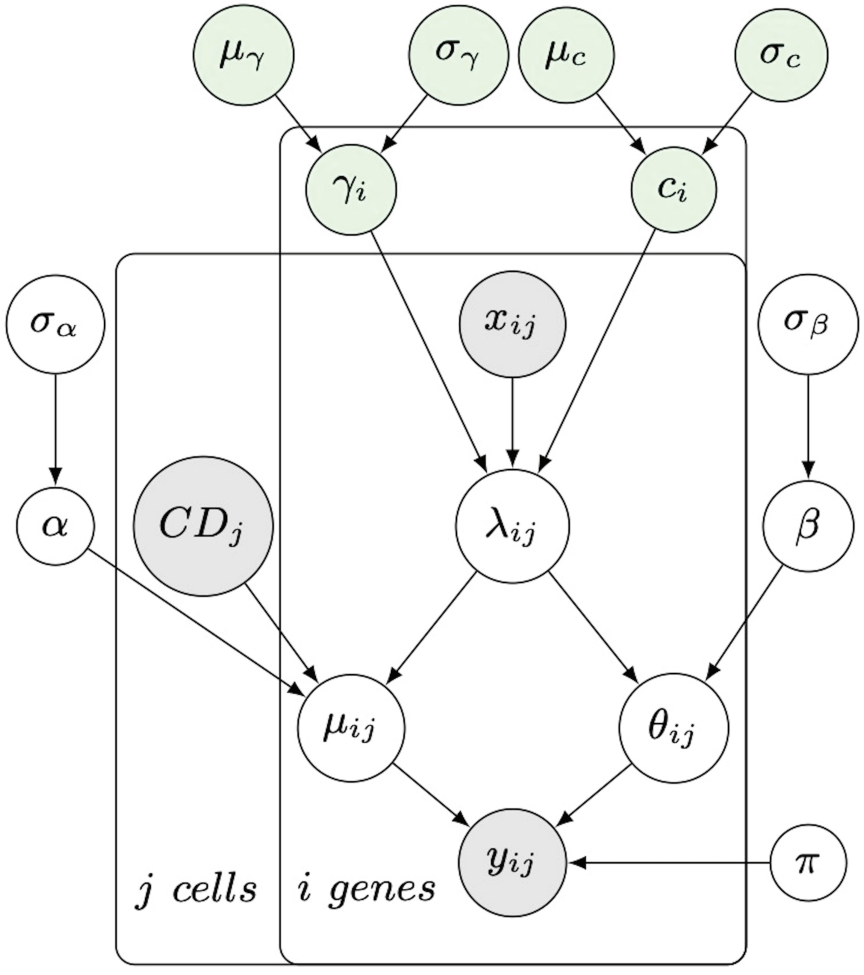

A
Non-Neuronal
Glutamatergic
GABAergic
Astro
OPC
L5 ET
L6 CT
LAMP5
PVALB
Penalty
Credit
Distance
Weighted penalty
Confusion matrix
Weighted confusion matrix
C
D
E
Wilcoxon, p = 1.17e-10
Wilcoxon, p = 8.3e-07
Wilcoxon, p = 6.9e-11
Percentage of across broad cell type mistakes over all mistakes
hierarchical
flat
B
Non-neuronal
Excitatory
Inhibitory
Cell type
Distance
Count

Fig. 6



$$y_{ij} \sim \mathrm{ZINB}(\mu_{ij}, \theta_{ij}, \pi)$$
$$\ln(\mu_{ij}) = \alpha + \ln(\lambda_{ij}) + \ln(CD_j)$$
$$\ln(\theta_{ij}) = \beta + \lambda_{ij}$$
$$\mathrm{logit}(\lambda_{ij}) = \gamma_i \times \sqrt{x_{ij}} + c_i$$

$$\alpha \sim \mathcal{N}(0, \sigma_\alpha)$$
$$\beta \sim \mathcal{N}(0, \sigma_\beta)$$
$$\gamma_i \sim \mathcal{LN}(\mu_\gamma, \sigma_\gamma)$$
$$c_i \sim \mathcal{N}(\mu_c, \sigma_c)$$
$$\sigma_\alpha, \sigma_\beta, \sigma_\gamma, \sigma_c \sim \mathrm{HalfCauchy}(0, 5)$$
$$\mu_\gamma, \mu_c \sim \mathrm{Cauchy}(0, 5)$$
$$\pi \sim \mathcal{B}(1, 1)$$

y_{ij}: number of molecules in gene i and cell j from spatial data
x_{ij}: relative expression of gene i in cell j from scRNA-seq data
CD_j: total number of molecules in cell j from spatial data
λ_{ij}: corrected relative expression of gene i in cell j
γ_i and c_i: platform effect magnitude
$\mu_\gamma, \sigma_\gamma, \mu_c, \sigma_c$: platform effect hyperparameters
μ_{ij}: mean expression of gene i in cell j
θ_{ij}: dispersion of gene i in cell j
π: zero inflation parameter

Fig. S1

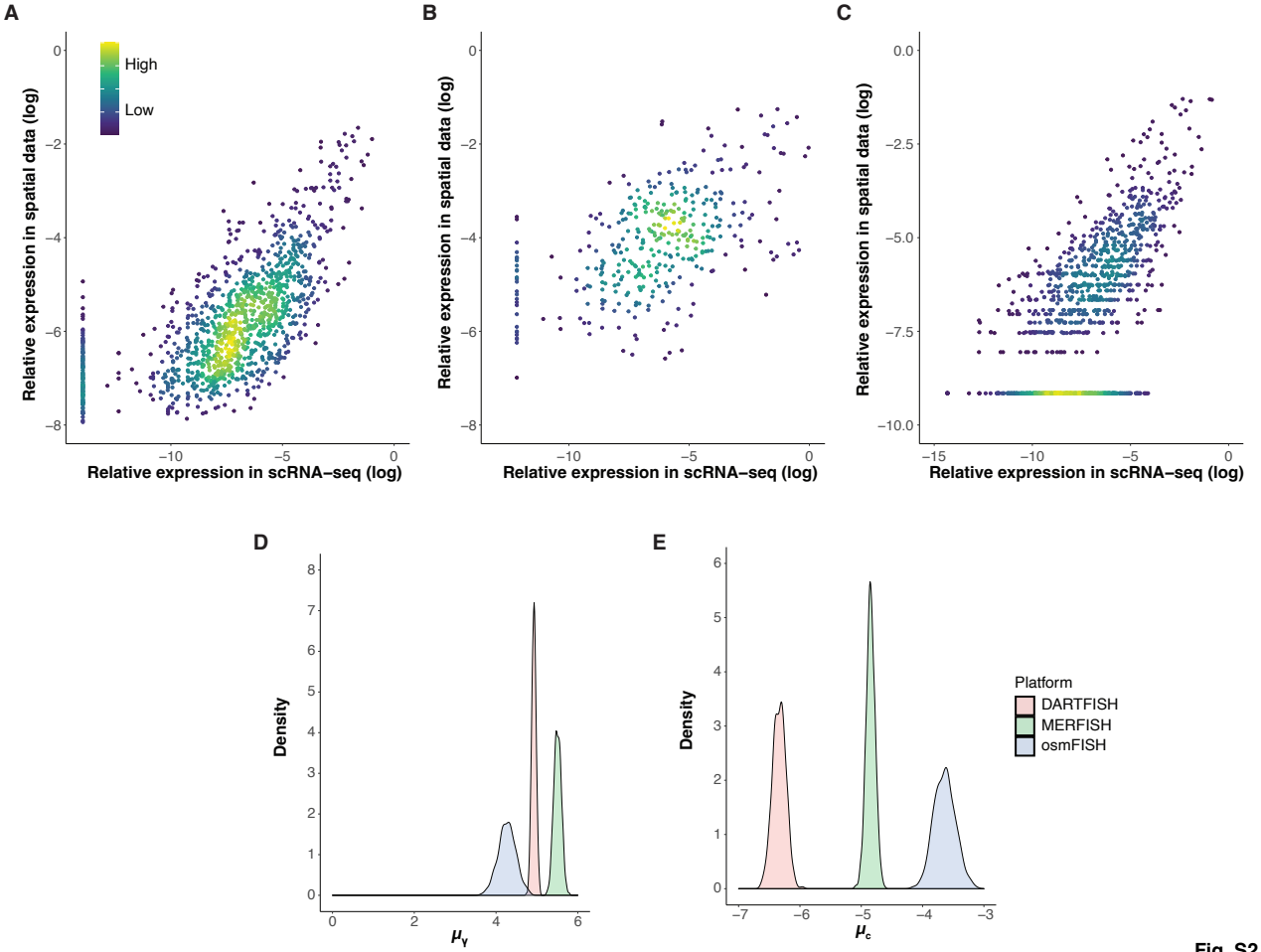

A
High
Low
Relative expression in spatial data (log)
Relative expression in scRNA−seq (log)
B
Relative expression in spatial data (log)
Relative expression in scRNA−seq (log)
C
Relative expression in spatial data (log)
Relative expression in scRNA−seq (log)
D
Density
μ_Y
E
Density
μ_c
Platform
DARTFISH
MERFISH
osmFISH

Fig. S2

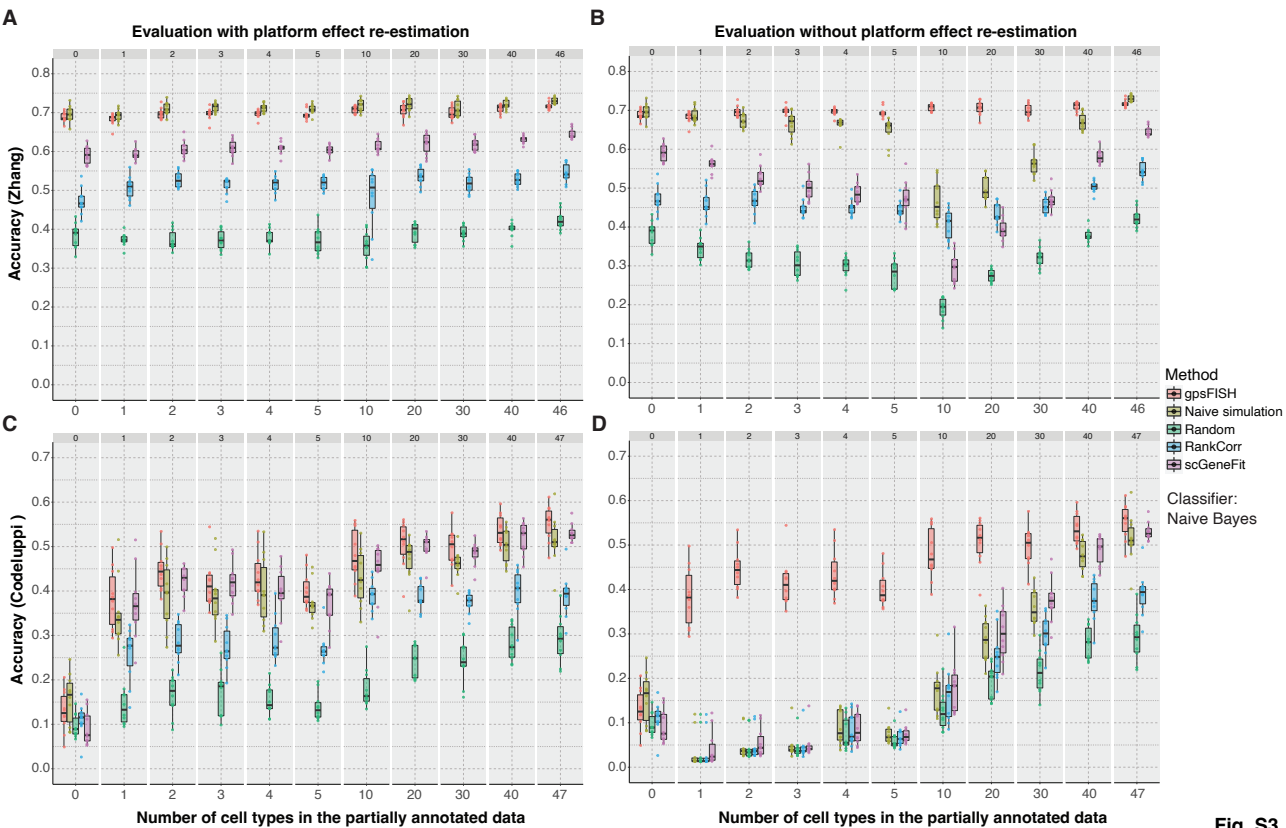

A
Evaluation with platform effect re-estimation
B
Evaluation without platform effect re-estimation
C
D
Accuracy (Zhang)
Accuracy (Codeluppi)
Number of cell types in the partially annotated data
Number of cell types in the partially annotated data
Method
gpsFISH
Naive simulation
Random
RankCorr
scGeneFit
Classifier:
Naive Bayes

Fig. S3

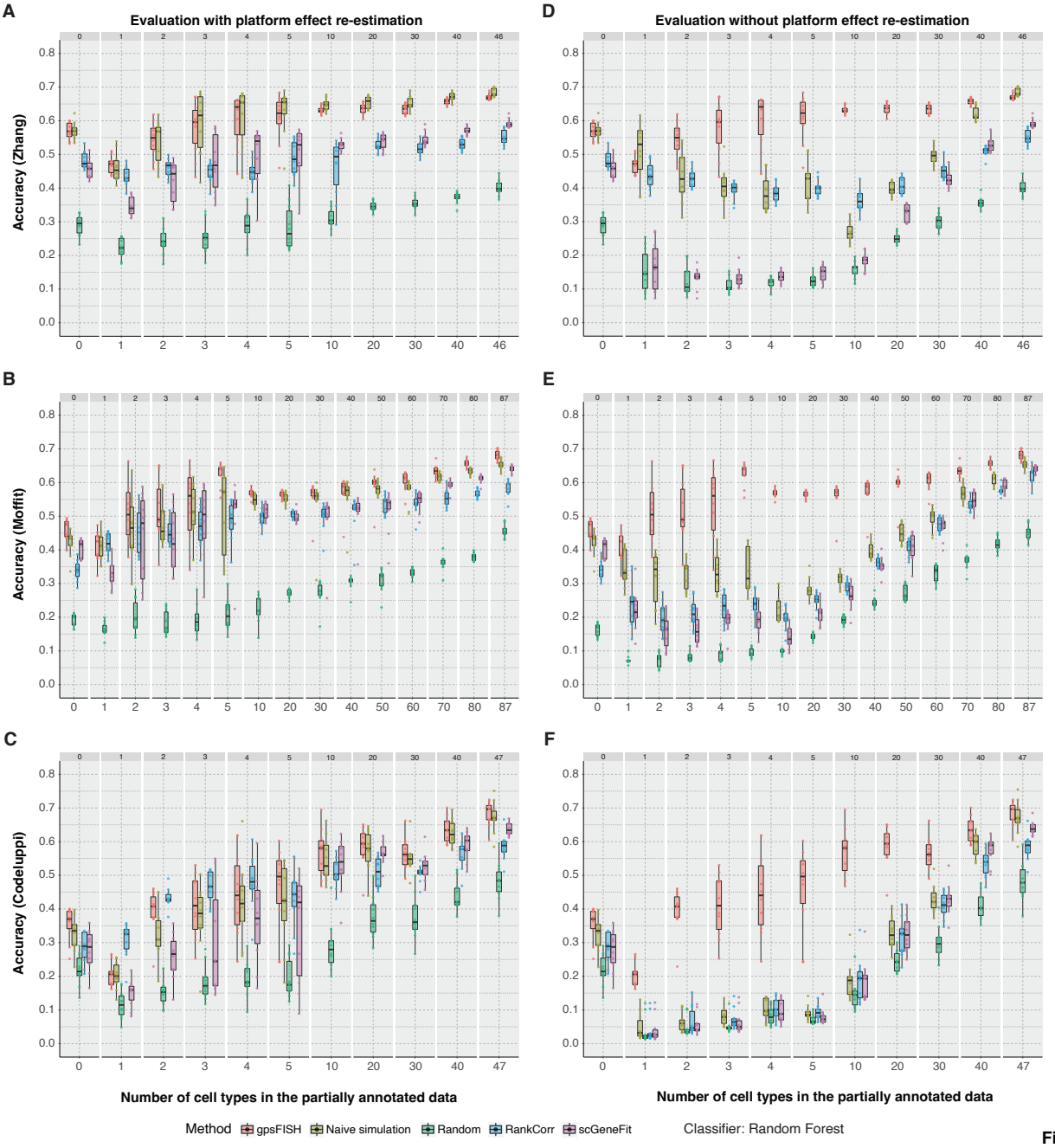

A
Evaluation with platform effect re-estimation
D
Evaluation without platform effect re-estimation
B
E
C
F
Accuracy (Zhang)
Accuracy (Moffit)
Accuracy (Codeluppi)
Number of cell types in the partially annotated data
Number of cell types in the partially annotated data
Method
gpsFISH
Naive simulation
Random
RankCorr
scGeneFit
Classifier: Random Forest

Fig. S4

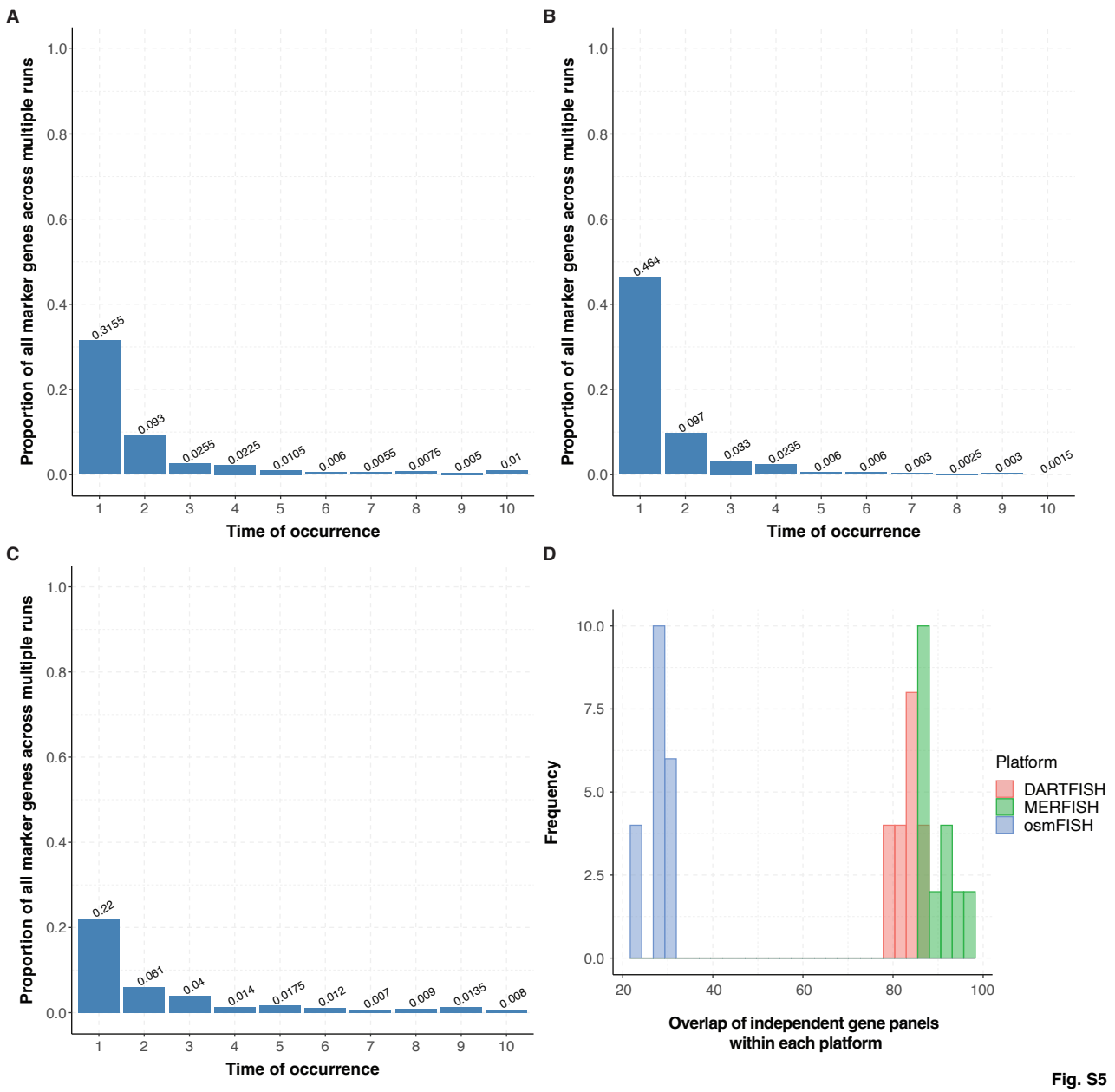

A
Proportion of all marker genes across multiple runs
1.0
0.8
0.6
0.4
0.2
0.0
0.3155
0.093
0.0255
0.0225
0.0105
0.006
0.0055
0.0075
0.005
0.01
1 2 3 4 5 6 7 8 9 10
Time of occurrence
B
Proportion of all marker genes across multiple runs
0.464
0.097
0.033
0.0235
0.006
0.006
0.003
0.0025
0.003
0.0015
Time of occurrence
C
Proportion of all marker genes across multiple runs
0.22
0.061
0.04
0.014
0.0175
0.012
0.007
0.009
0.0135
0.008
Time of occurrence
D
Frequency
10.0
7.5
5.0
2.5
0.0
20 40 60 80 100
Overlap of independent gene panels
within each platform
Platform
DARTFISH
MERFISH
osmFISH

Fig. S5

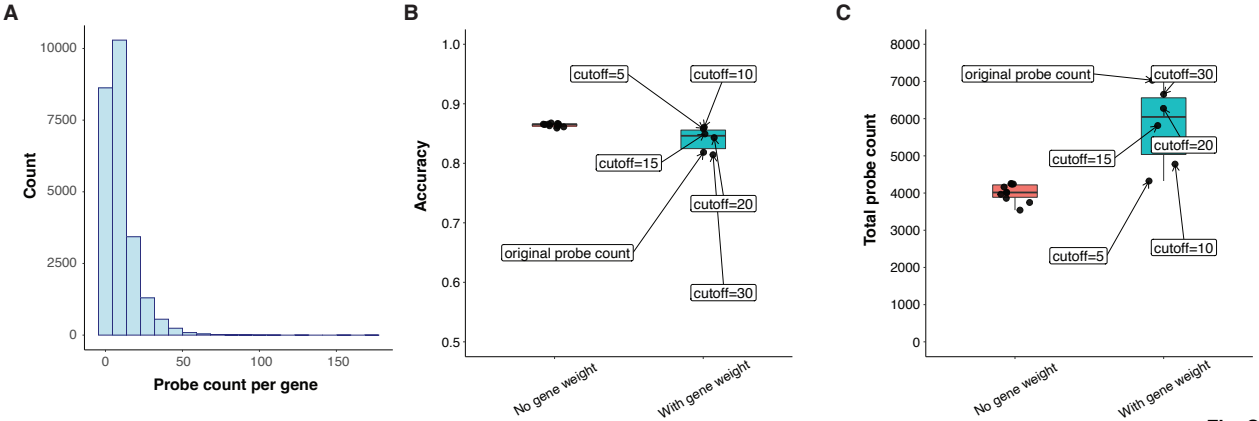

A
Count
10000
7500
5000
2500
0
0
50
100
150
Probe count per gene
B
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
cutoff=5
cutoff=10
cutoff=15
cutoff=20
original probe count
cutoff=30
No gene weight
With gene weight
C
Total probe count
8000
7000
6000
5000
4000
3000
2000
1000
0
original probe count
cutoff=30
cutoff=20
cutoff=15
cutoff=5
cutoff=10
No gene weight
With gene weight

Fig. S6

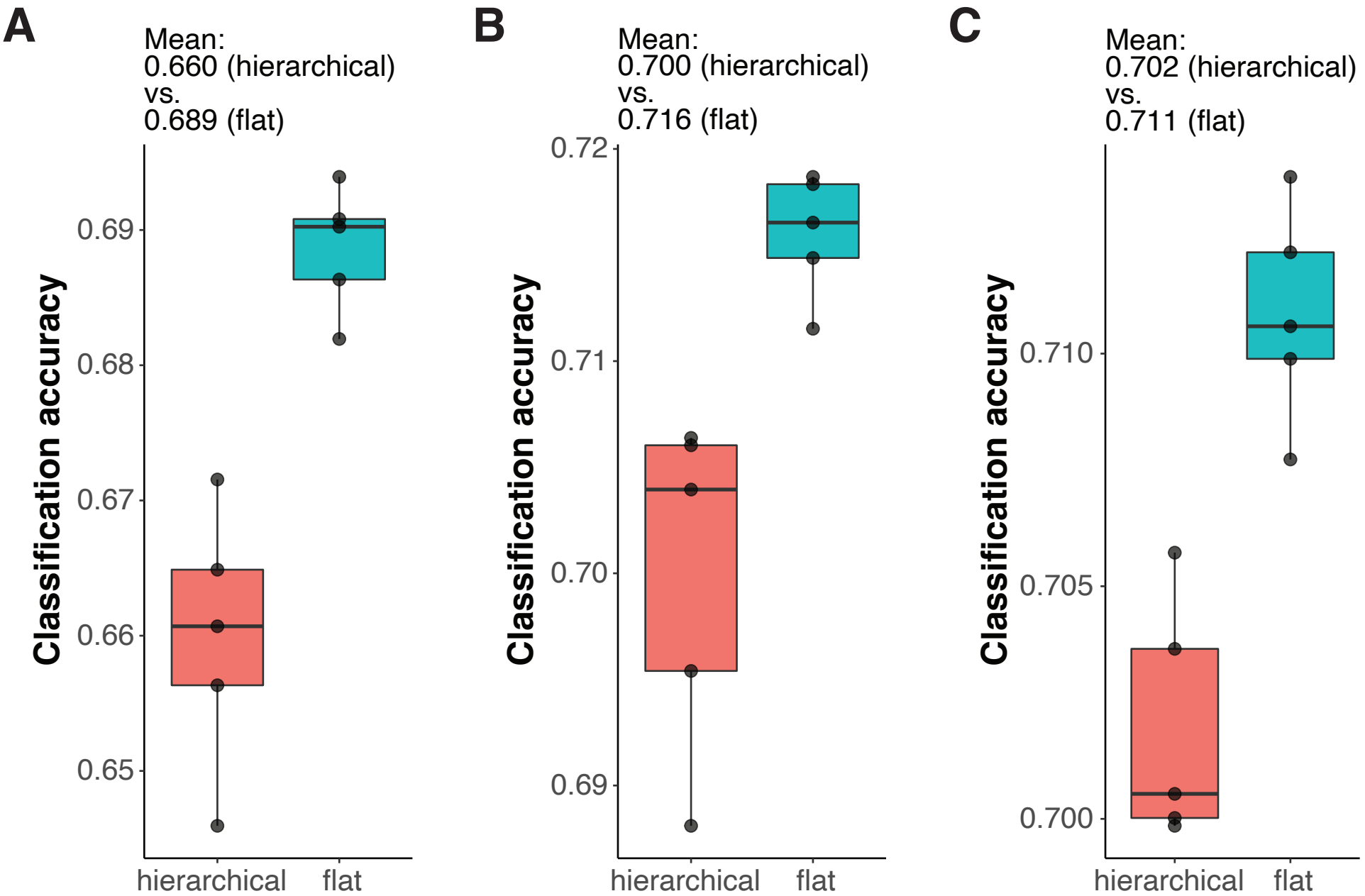

A
Mean:
0.660 (hierarchical)
vs.
0.689 (flat)
Classification accuracy
0.69
0.68
0.67
0.66
0.65
hierarchical
flat
B
Mean:
0.700 (hierarchical)
vs.
0.716 (flat)
Classification accuracy
0.72
0.71
0.70
0.69
hierarchical
flat
C
Mean:
0.702 (hierarchical)
vs.
0.711 (flat)
Classification accuracy
0.710
0.705
0.700
hierarchical
flat

Fig. S7

Figure 1 Platform effect between scRNA-seq and targeted spatial transcriptomics technologies.
A-C:
Scatter plot showing the log transformed relative expression of genes measured by both scRNA-seq and targeted spatial transcriptomics across three datasets, Moffit (**A**), Codeluppi (**B**), and Zhang (**C**), respectively. A small value is added to avoid negative infinity after log transformation. Each dot represents the relative expression of one gene in one cell type. Denominator for relative expression calculation is from all genes measured by both technologies. Color indicates density of dots. Dots should fall on the diagonal when there is no platform effect.
D:
Density plot of Deming regression coefficient for each dataset. Deming regression is fitted for each gene using relative expression measured by scRNA-seq and spatial transcriptomics data with intercept fixed to 0.
E:
Posterior predictive check of the Bayesian models fitted using each of the three datasets. QQ plot showing the distribution of simulated vs. observed spatial transcriptomics measurements.
F-G:
Density plot showing the estimated posterior distribution of σ_γ (**F**) and σ_c (**G**).

Figure 2: Schematic overview of gpsFISH.
Upper left, an scRNA-seq dataset with cell type annotation is used as input. Bottom, a genetic algorithm framework is used for gene panel selection. Platform effects are accounted for using a Bayesian model. Cell type hierarchy can also be incorporated. Upper right, output includes optimized gene panel with classification statistics.

Figure 3: Gene panel selection using gpsFISH.
A:
UMAP of cells based on the mouse hypothalamic scRNA-seq data from Moffit dataset at level 1 cell type annotation.
B:
Normalized confusion matrix of the optimized gene panel for Moffit dataset at level 1 cell type annotation.
C:
AUC for each cell type of the same gene panel.
D-E:
UMAP of cells based on simulated spatial transcriptomics measurements with platform effect of the optimized gene panel selected with (**D**) and without (**E**)considering platform effect at level 1 cell type annotation.

Figure 4: Comparison between gpsFISH and other gene selection methods.
A-B:
Box plot showing classification accuracy distribution of gene panels selected by 5 gene panel selection methods at different levels of partial annotation. The result is

based on the Moffit dataset using evaluation with (**A**) and without (**B**) platform effect re-estimation. Naïve Bayes is used as classifier.
C:
Normalized confusion matrix of the optimized gene panel for Moffit dataset at level 2 cell type annotation with dendrogram showing the cell type hierarchy. Diagonal values of the confusion matrix are removed for better visualization of misclassifications.

Figure 5: Redundancy in gene space across independent gene panel optimizations enables incorporation of customized preferences.
A:
Distribution of overlap of independent gene panels across 10 optimizations within each platform at level 1 cell type annotation.
B:
Accuracy of optimized gene panels without vs. with gene weight across 10 optimizations.
C:
Total number of probes of optimized gene panels without vs. with gene weight across 10 optimizations.

Figure 6: Gene panel selection with cell type hierarchy.
A:
Schematic of hierarchical gene selection using cell type hierarchy. A weighted penalty matrix is constructed using cell type hierarchy information quantified by pairwise distance between cell types. Additional penalty can be specified according to the cell type hierarchy. The weighted penalty matrix is then multiplied element-wise with the original confusion matrix to get the weighted confusion matrix for fitness evaluation.
B:
Original (left) vs. weighted (right) confusion matrix of the same optimized gene panel from Moffit dataset at level 2 cell type annotation with dendrogram showing the cell type hierarchy. Diagonal values of the confusion matrix are removed for better visualization of misclassifications.
C-E:
Percentage of across broad cell type (level 1) misclassifications over all misclassifications for flat vs. hierarchical classification on the Moffit (**C**), Codeluppi (**D**), and Zhang (**E**) dataset. Each dot represents one cell type with dots representing the same cell type connected. Wilcoxon paired test is performed between the percentages from flat vs. hierarchical classification and the p value is shown.

Figure S1: Schematic of the Bayesian model for platform effect estimation.
Circles in gray represent observed variables. Circles in green correspond to platform effect related variables to be estimated.

Figure S2: Bayesian model captures platform effect between scRNA-seq and targeted spatial transcriptomics technologies.

A-C:
Scatter plot showing the log transformed relative expression of genes measured by scRNA-seq vs. simulated spatial transcriptomics data using fitted Bayesian model for Moffit (**A**), Codeluppi (**B**), and Zhang (**C**), respectively.
D-E:
Density plot showing the estimated posterior distribution of μ_γ (**D**) and μ_c. (**E**).

Figure S3: Comparison between gpsFISH and other gene selection methods on the Zhang and Codeluppi dataset.
Box plot showing classification accuracy distribution of gene panels selected by 5 gene panel selection methods at different levels of partial annotation. (**A**) Zhang dataset using evaluation with platform effect re-estimation. (**B**) Zhang dataset using evaluation without platform effect re-estimation. (**C**) Codeluppi dataset using evaluation with platform effect re-estimation. (**D**) Codeluppi dataset using evaluation without platform effect re-estimation. Naïve Bayes is used as classifier.

Figure S4: Comparison between gpsFISH and other gene selection methods using random forest as classifier.
Box plot showing classification accuracy distribution of gene panels selected by 5 gene panel selection methods at different levels of partial annotation for the three datasets using evaluation with (**A-C**) and without (**D-E**) platform effect re-estimation. Random forest is used as classifier.

Figure S5: High redundancy across optimizations using gpsFISH.
A-C:
Bar plot showing among all the genes selected in 10 optimizations, the percentage of them that are included in 1 to 10 optimized panels for Moffit (**A**), Codeluppi (**B**), and Zhang (**C**) dataset at level 1 cell type annotation.
D:
Distribution of overlap of independent gene panels across 10 optimizations within each platform at level 2 cell type annotation.

Figure S6: Weighted gene panel selection based on probe count per gene.
A:
Distribution of probe count per gene for the Zhang dataset.
B-C:
Distribution of accuracy (**B**) and total number of probes (**C**) of optimized gene panels from optimization without and with gene weight. Optimization without gene weight is performed 10 times. Optimization with gene weight is performed 6 times, each time with a different probe count cutoff (no cutoff, 5, 10, 15, 20, 30).

Figure S7: Accuracy of optimized gene panels using flat vs. hierarchical gene selection.
A-C:

Distribution of accuracy of optimized gene panels using flat vs. hierarchical gene selection for Moffit (**A**), Codeluppi (**B**), and Zhang (**C**), respectively. Both flat and hierarchical gene selection are performed 5 times.

Information of the Moffit, Codeluppi, and Zhang dataset

Dataset	Moffit	Codeluppi	Zhang
Single-cell RNA sequencing platform	10X Genomics Chromium v2 & Illumina NextSeq500	Illumina HiSeq 2000	10X Genomics Chromium v3 & Illumina NovaSeq
Spatially resolved transcriptomics platform	MERFISH	osmFISH	DARTFISH
Number of cell types in scRNA-seq at level 1 for gene panel selection	12	12	16
Number of cells in scRNA-seq at level 1 for gene panel selection	30370	2816	64693
Number of cell types in scRNA-seq at level 2 for gene panel selection	87	47	46
Number of cells in scRNA-seq at level 2 for gene panel selection	30370	2816	64693
Number of overlapping cell types at level 1 for platform effect estimation	9	11	7
Number of cells in scRNA-seq at level 1 for platform effect estimation	29760	2139	43261
Number of cells in spatial transcriptomics data at level 1 for platform effect estimation	417026	3127	1341

Curated marker genes for the Zhang dataset

Subclass (Full Name)	Subclass Level 3	Subclass Level 1	Class	Substructure	Curated: Positive Markers	Curated: Negative Markers	Degenerative State Upregulated	Degenerative Downregulated	Adaptive State	Cycling State
Podocyte	POD	POD	epithelial cells	glomerulus	NPHS1, NPHS2, PTPRQ, CLIC5, NTNG1		CDKN1C, SPOCK2	PTPRQ		
Parietal Epithelial Cell	PEC	PEC	epithelial cells	glomerulus	CLDN1, VCAM1, CFH					
Proximal Tubule Cell		PT	epithelial cells	proximal tubules	LRP2, CUBN, AQP1		CST3, HAVCR1, CLU, APOE, S100A6, B2M		ITGB8, CDH6, DCDC2, VCAM1, DLGAP1, HAVCR1, PLSCR1	MKI67, TOP2A
Proximal Tubule Epithelial Cell Segment 1	PT-S1	PT	epithelial cells	proximal tubules	SLC5A12, SLC22A6,SLC22A8, SLC5A2					
Proximal Tubule Epithelial Cell Segment 2	PT-S2	PT	epithelial cells	proximal tubules	SLC34A1, SLC22A7					
Proximal Tubule Epithelial Cell Segment 3	PT-S3	PT	epithelial cells	proximal tubules	SLC5A11, MOGAT1, SLC22A7, SLC22A24, SLC7A13					
Thin Limb Cell		TL	epithelial cells	intermediate tubules	CRYAB, TACSTD2, AKR1B1					
Descending Thin Limb Cell Type 2	DTL2	DTL	epithelial cells	intermediate tubules	AQP1, UNC5D	CLDN10				
Descending Thin Limb Cell Type 1	DTL1	DTL	epithelial cells	intermediate tubules	ADGRL3, ID1	CLDN10, AQP1				
Descending Thin Limb Cell Type 3	DTL3	DTL	epithelial cells	intermediate tubules	CLDN1, SH3GL3, SLC14A2, SMOC2	CLDN10, AQP2				
Ascending Thin Limb Cell	ATL	ATL	epithelial cells	intermediate tubules	CLDN1, SH3GL3, CLDN10, PROX1					
Thick Ascending Limb Cell		TAL	epithelial cells	Distal tubules	SLC12A1, CASR, UMOD, EGF			UMOD, EGF	ITGB6, PROM1, CCL2, PLSCR1, DCDC2	
Medullary Thick Ascending Limb Cell	M-TAL	TAL	epithelial cells	Distal tubules	PROX1					
Cortical Thick Ascending Limb Cell	C-TAL	TAL	epithelial cells	Distal tubules						
Macula Densa Cell	MD	TAL	epithelial cells	Distal tubules	NOS1, ROBO2	UMOD				
Distal Convoluted Tubule Cell		DCT	epithelial cells	Distal tubules	Distal tubules					
Distal Convoluted Tubule Cell Type 1	DCT1	DCT	epithelial cells	Distal tubules						
Distal Convoluted Tubule Cell Type 2	DCT2	DCT	epithelial cells	Distal tubules	SLC8A1					
Connecting Tubule		CNT	epithelial cells	Collecting tubules	SLC8A1, HSD11B2, CALB1					
Connecting Tubule Cell	CNT	CNT	epithelial cells	Collecting tubules						
Connecting Tubule Principal Cell	CNT-PC	CNT	epithelial cells	Collecting tubules	SCNN1G, SCNN1B					
Principal Cell		PC	epithelial cells	Collecting tubules	AQP2, AQP3					
Cortical Collecting Duct Principal Cell	CCD-PC	PC	epithelial cells	Collecting tubules	SCNN1G, SCNN1B					
Outer Medullary Collecting Duct Principal Cell	OMCD-PC	PC	epithelial cells	Collecting tubules	SCNN1G, SCNN1B					
Inner Medullary Collecting Duct Cell	IMCD	PC	epithelial cells	Collecting tubules	SLC14A2, HS3ST5					
Papillary Epithelial Cells	PapE	PC	epithelial cells	Collecting tubules	TP63, KRT5					
Intercalated Cell		IC	epithelial cells	Collecting tubules	ATP6V0D2					
Cortical Collecting Duct Intercalated Cell Type A	CCD-IC-A	IC	epithelial cells	Collecting tubules	SLC26A7, SLC4A1					
Connecting Tubule Intercalated Cell Type A	CNT-IC-A	IC	epithelial cells	Collecting tubules	SLC26A7, SLC4A1, SLC8A1					
Outer Medullary Collecting Duct Intercalated Cell Type A	OMCD-IC-A	IC	epithelial cells	Collecting tubules	SLC26A7, SLC4A1, KIT					
Intercalated Beta Cell	IC-B	IC	epithelial cells	Collecting tubules	SLC26A4, SLC4A9					
Endothelial Cell		EC	endothelial cells	vessels	PECAM1, CD34					
Glomerular Capillary Endothelial Cell	EC-GC	EC	endothelial cells	glomerulus	EMCN, HECW2, PLAT, EHD3					
Afferent / Efferent Arteriole Endothelial Cell	EC-AEA	EC	endothelial cells	vessels	BTNL9, PALMD, TM4SF1, SERPINE2, AQP1					
Descending Vasa Recta Endothelial Cell	EC-DVR	EC	endothelial cells	vessels	BTNL9, PALMD, TM4SF1, SERPINE2, AQP1, SLC14A1					
Peritubular Capilary Endothelial Cell	EC-PTC	EC	endothelial cells	vessels	DNASE1L3, PLVAP					
Ascending Vasa Recta Endothelial Cell	EC-AVR	EC	endothelial cells	vessels	DNASE1L3, PLVAP, TLL1	PALMD, BTNL9, SLC14A1				
Lymphatic Cell	EC-LYM	EC	endothelial cells	vessels	PROX1, MMRN1					
Vascular Smooth Muscle Cell and Pericyte		VSM/P	stroma cells	interstitium	PDGFRB, NOTCH3		TAGLN, ACTA2			
Mesangial Cell	MC	VSM/P	stroma cells	glomerulus	POSTN, PIEZO2, ITGA8					
Renin-positive Juxtaglomerular Granular Cell	REN	VSM/P	stroma cells	interstitium	REN					
Vascular Smooth Muscle Cell	VSMC	VSM/P	stroma cells	interstitium	MYH11, MCAM					
Vascular Smooth Muscle Cell / Pericyte	VSMC/P	VSM/P	stroma cells	interstitium						
Fibroblast		FIB	stroma cells	interstitium	C7, DCN, COL1A1, PDGFRA				FLRT2, FGF14, IGF1	
Fibroblast	FIB	FIB	stroma cells	stroma cells	MEG3, LAMA2					
Medullary Fibroblast	M-FIB	FIB	stroma cells	interstitium	SYT1, TNC					
Myofibroblast	MyoF	FIB	stroma cells	interstitium	FAP, ACTA2, TAGLN, POSTN, GLI2, COL5A1					
Immune Cells		IMM	immune cells	interstitium	PTPRC					
B Cell	B	IMM	immune cells	interstitium	MS4A1, BANK1					
Plasma Cell	PL	IMM	immune cells	interstitium	IGKC, MZB1					
T Cell	T	IMM	immune cells	interstitium	CD3E, CD4					
Natural Killer T Cell	NKT	IMM	immune cells	intrrstitium	NKG7, GNLY, CD96, RUNX3					
Mast Cell	MAST	IMM	immune cells	interstitium	MS4A2, CPA3, KIT	IL3RA				
M2 Macrophage	MAC-M2	IMM	immune cells	interstitium	CD163, F13A1, MRC1, CD14					
Classical Dendritic Cell	cDC	IMM	immune cells	interstitium	ITGAX, FLT3	CD14				
Plasmacytoid Dendritic Cell	pDC	IMM	immune cells	interstitium	IL3RA, FLT3	CD14				
Non-Classical Monocyte	ncMON	IMM	immune cells	interstitium	FCN1, HLA-DRA, FCGR3A					
Neutrophil	NC	IMM	immune cells	interstitium	S100A8, S100A9, IFITM2, FCGR3B					
Schwann Cell / Neural	SC/NEU	NEU	neural like cells	interstitium	CDH19, NRXN1, PLP1, S100B					