Erasure of Biologically Meaningful Signal by Unsupervised scRNAseq Batch-correction Methods

Scott R Tyler[1,*], Ernesto Guccione[1,2,3], and Eric E Schadt[4,*]

[1] Department of Oncological Sciences, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2] Center for OncoGenomics and Innovative Therapeutics (COGIT); Department of Pharmacological Sciences and Mount Sinai Center for Therapeutics Discovery, Icahn School of Medicine at Mount Sinai, New York, NY, USA; Black Family Stem Cell Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

[3] Bioinformatics for Next Generation Sequencing (BiNGS) Shared Resource Facility, Icahn School of Medicine at Mount Sinai, New York, NY, USA[4] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

*Corresponding authors

Abstract

Single cell RNAseq (scRNAseq) batches range from technical-replicates to multi-tissue atlases, thus requiring robust batch-correction methods that operate effectively across this spectrum of between-batch similarity. Commonly employed benchmarks quantify *removal* of batch effects and preservation of *within-batch* variation, the preservation of biologically meaningful differences *between* batches has been under-researched. Here, we address these gaps, quantifying batch effects at the level of cluster composition and along overlapping topologies through the introduction of two new measures. We discovered that standard approaches of scRNAseq batch-correction erase cell-type and cell-state variation in real-world biological datasets, single cell gene expression atlases, and *in silico* experiments. We highlight through examples showing that these issues may create the artefactual appearance of external validation/replication of findings. Our results demonstrate that either biological effects, if known, must be balanced between batches (like bulk-techniques), or technical effects that vary between batches must be explicitly modeled to prevent erasure of biological variation by unsupervised batch correction approaches.

32 An ideal batch-correction method removes batch confounded differences in technical
33 effects, while retaining biological signals of interest. With single cell RNA sequencing
34 (scRNAseq) advancing at a rapid pace, a wave of normalization, batch-correction, and
35 benchmarks to assess the accuracy of such methods have stormed onto the scene. Such
36 methods attempt to tackle difficult issues in scRNAseq experiments, such as individual samples
37 often being processed as their own batch, rather than in multiplex. However, our knowledge of
38 unwanted technical variation that can occur between batches is presently limited to count depth
39 per cell[1], nuclear-long non-coding RNAs (lncRNAs)[2, 3], and cellular damage[4, 5]. While some
40 unsupervised approaches attempt to use our (limited) understanding of the technical effects
41 (referred to here as Class-1 algorithms)[6-8], many methods in common use do not explicitly
42 model technical or biological effects (referred to here as Class-2 algorithms), aligning topologies
43 agnostically to the variation source. Class-2 algorithms include CombatSeq-Negative-Binomial,
44 which rotates datasets within count-space, and others that perform a latent dimension mapping
45 (LDM) with a loss function decreasing batch differences: Multiple Canonical Correlation Analysis
46 (MCCA),[9] Harmony,[10] Liger,[11] Scanorama, and scVI[12, 13]. There are also semi-supervised Class-
47 3 algorithms, such as scANVI and scGEN[14, 15]; these attempt to model the biology of interest
48 assigned to a given sample by the experimenter in the form of prior-expectation informed by
49 domain-knowledge. These use a joint learning approach with respect to the human-assigned
50 class labels and the corresponding functional-omic measurements. Here we focus on the
51 efficacy of unsupervised methods (Class-1/2), which cannot be fairly compared to semi-
52 supervised given that semi-supervised methods include prior expectation of “important biology”
53 into the results.
54 The variations among samples we observe in scRNAseq experiments are a mixture of
55 technical and biological effects. We reasoned that batch correction algorithms that do not
56 explicitly model or discriminate between either type of effect (biological or technical), are at risk
57 of removing both sources of variation (**Supplemental Note-1, Supp. Fig. 1; proof-1**). Because

58 Class-2 algorithms do not explicitly discriminate technical or biological effects, we hypothesized
59 that they may erase biologically meaningful sources of variation in practice. However, even
60 testing this is a challenge, because the ground-truth biology in complex, biologically derived
61 datasets by definition are not known *a priori* suggesting that to discriminate biological and
62 technical effects, a deeper understanding of technical effects impacting batch-to-batch variation
63 is critical.

64 While others have previously noted the risks of over-mergers[16] by Class-2 algorithms for
65 the reasons described above, the potential for this property has not been quantified.
66 Furthermore, in practice, a literature mining meta-analysis of Class-2 algorithms indicates their
67 broad use for batch-correction (**Extended Data** (**ED**) **Table 1**); Chi-square test of independence
68 "batch-correction" and <algorithm>; all comparisons: $P \approx 0$), suggesting that they are none-the-
69 less used with the intention of removing batch confounded technical effects (i.e. batch-
70 correction).

71 An increasing number of benchmarks attempt to assess batch-correction algorithm
72 efficacy by judging the ability to thoroughly mix batches from different biological samples and
73 technologies (despite these components being intrinsically confounded),[9, 17]; these benchmarks
74 will often assess the ability to re-discover the cluster/cell-type labels defined by those who
75 originally performed the analysis, thus setting their approach to the ground truth[17, 18]. While
76 these approaches are important for cross-technology/species meta-analyses of highly similar
77 datasets, a typical bench biologist performs a treatment/control or disease/non-diseased
78 experiment on a single platform, without *a priori* knowledge of the similarity *between* batches.
79 Furthermore, in practice, the typical bench-biologist will generally follow guidance provided by
80 algorithm authors through their publications/websites when applying the algorithms. In this
81 common use-case, we sought to understand and benchmark the most widely used methods of
82 batch-correction, characterizing how well they can remove technical effects while preserving
83 biological effects of interest.

84 Here we show that many existing benchmark-metrics either assume biological-equality
85 between separate biological-individuals or assume that published labels accurately identified all
86 meaningful sources of variation. We introduce a new functional framework to aid in parsing
87 technical and biological sources of variation. We then introduce three new methods/tools to
88 extend existing benchmarks for scRNAseq experiments by characterizing local topological
89 differences within-clusters/across batches (WCAB) and within topologically overlapped regions
90 within clusters/across batches (TOWCAB), and by characterizing cluster composition between
91 batches in a way that is invariant to the number of clusters [percent-maximum-difference
92 (PMD)]. We used these metrics to assess the ability of Class 1 and 2 unsupervised algorithms[9-
93 11, 19, 20] to remove technical differences, and preserve batch-confounded, and non-confounded
94 cell-types/-states both from a cluster-centric perspective (PMD) and along topologically
95 overlapping regions (TOWCAB). We reprocessed several published datasets using different
96 batch correction approaches and found that these issues can dramatically alter interpretation of
97 scRNAseq experiments. First, we examine single- and multi-tissue atlases and demonstrate
98 biological heterogeneity represented in contiguous topologies using the published coordinates,
99 identifying cell-type erasure by two class-2 algorithms. Second, we demonstrate with *in silico*
100 and real-world datasets that transcriptome mixtures (representing cell-states or impure droplets)
101 can result in aberrant co-clustering across batches. Finally, we reanalyzed published human
102 brain scRNAseq datasets, showing that a Class-2 algorithm created the impression of between-
103 dataset concordance, in a manner that has already impacted cancer treatment research, thus
104 raising the stakes of batch-correction, as these approaches move towards clinical care.

105 **Results**

106 Despite the broad acceptance of cell-identity/-state being fluid and dynamic, many scRNAseq
107 batch-correction benchmarks use manually labeled cluster annotations, assuming that cells
108 from different donors, labeled by different bioinformaticians, are biologically equivalent. As a
109 result of this assumption, such benchmarks penalize a batch correction method for not
110 thoroughly mixing cells of the same label. However, these labels are generated by
111 experimenters doing their best to identify the biological sources of variation that are expected to
112 be important in their experiment. In practice, manually generated labels may not capture *all*
113 aspects of cell-identity, including differences in cell-state (i.e.: circadian state) that may differ
114 between biological samples assayed as separate batches. Common metrics such as ASW,
115 iLISI, cLISI, and kBET-cell-type penalize for maintaining this unlabeled source of variation
116 (**Supplemental Note-1; Supplemental Fig. 2**, **ED Table 2**). The rationale behind these
117 penalties is the presumption that user-defined labels annotate *all* biologically meaningful
118 variation, and within-label topologies should thus be thoroughly mixed, discounting the
119 possibility that the experimenters providing the reference labels may not have provided labels
120 reflecting all biologically derived attributes of cell identity.

121 We therefore sought to characterize unsupervised batch-correction algorithms in a
122 manner that did not have these limitations or assumptions. Ideally, a batch-correction algorithm
123 should remove batch confounded differences in technical effects while preserving biological
124 effects. Given that we seek to study biological effects and cannot presume to know them *a*
125 *priori*, we began from first principles, asking what types of functions are involved in defining the
126 signals detected from various high-dimensional molecular measures (e.g., functional genomic
127 data) and how might they be confounded or balanced between batches.

128 We began from the premise that all observations in a scRNAseq dataset are manifest
129 from either biological (B) or technical (T) effects (**Fig. 1**; **Supplemental Note-2**). Prior to
130 experimenter intervention to perform the measurements, only biological effects exist (**Fig. 1a**).

131 Once sample collection has begun technical effects begin to accumulate, the technical effects
132 will be generated through both the biological response to sample processing (T_r), which is not of
133 interest and can be reduced by inhibition of transcription/translation[21, 22], technical effects of
134 sample processing/measurement (T_m) (different lysis buffer composition, sequencing depth, etc)
135 (**Fig. 1b**). In the context of technical-replicates, we know that the biology of interest (B) is the
136 same between the two batches, leaving only the differences in technical effects of measurement
137 and processing (T_m), and the cell's response to processing (T_r) (which has been shown to be
138 cell-type specific[22]). When treated as a 'pattern matching' problem however, we are at risk of
139 removing biology of interest, when these sets of functions are not explicitly distinguished, if
140 there is any batch confounded biology, including in biological replicates (**Fig. 1c**, **Supplemental**
141 **Note-3**).
142 This type of framework captures the fact that with true technical replicates, local
143 differential expression analysis could measure the impact of technical effects directly (**Fig 2a**,
144 **Supplemental Note-2**; **Supplemental Fig. 3**). We thus developed an open-source tool to
145 assess differential gene detection (rather than expression, given that true technical replicates
146 have no true differences in transcription prior to processing) within-clusters/across-batches
147 (WCAB)[23], and then used protein-protein interaction and pathway analysis to identify the
148 biological coherence of the differential transcriptomic signatures, using local backgrounds of
149 only what was detected within each cluster to prevent generic tissue enrichment, high
150 expression, or annotation bias[24, 25]. Surprisingly, in technical-replicates derived from separate
151 tissues (mouse-brain[26] and mouse-heart[27]), we observed consistent differential detection
152 signatures regardless of correction strategy (**Fig. 2b**), that were enriched for large structural
153 components of the cell, such as mitochondria, the endoplasmic reticulum (ER), nuclear targeted
154 genes, and separately hemoglobin (**Fig. 2c**). These gene-complexes (T_m; **ED Table 3**) were
155 differentially detected within-clusters/across-batches in a large proportion of clusters, and
156 conserved across tissue types (**Fig. 2d-g**). However, the cell-stress and iron response (T_r) was

157 highly divergent across tissue-types in comparison to T_m genes/pathways (T=10.69, P=1.24e-4,
158 1-sample T-test) (**Fig. 2d**), fitting with prior reports of cell-type specific responses to
159 processing[21, 22]. We observed these same complexes were differentially detected in a different
160 species and tissue-type (human PBMCs) (**Fig. 2e-g**). The more conservative intersection of
161 mouse technical replicates (or their human orthologues) for this classification system accounted
162 for 77.3%, 69.2%, and 57.3% of differentially detected genes between technical replicates for
163 mouse-brain/-heart, and human-PMBCs, respectively.
164
165 ***Efficacy of batch-correction algorithms with identical cells and technical-replicates***
166 While technical replicates capture the combined impact of batch effects, such effects may still
167 contain several sources of technical variation. We therefore began our benchmark specifically
168 quantifying the ability of algorithms to remove the effect of count depth, quantifying the number
169 of sub-sampled cells that are co-clustered with their original cells, an approach recently dubbed
170 "molecular-cross-validation"[28]. Class-2 methods performed marginally better than no correction
171 or Class-1 algorithms when only analyzing mouse brain, however, this advantage was no longer
172 present when analyzed in the presence of a third dataset for which we expected little overlap in
173 cluster composition (homeostatic intestine) (**Box 1b**; **ED Fig. 1**).
174 Many scRNAseq analyses build on clustering results; it is therefore important to quantify
175 the effect of batch-correction on clustering results. However, batch-correction algorithms may
176 leave differing levels of variation in the final dataset, resulting in differing numbers of clusters;
177 yet there is generally no ground-truth number of clusters in biological datasets. We therefore
178 created the "percent-maximum-difference" (PMD) (**Fig. 3a**) metric/test to quantify the overall
179 similarity in cluster composition across batches. Unlike χ^2, χ^2 -log10(p-value), Cramer's V,
180 inverse Simpson's Index, or count normalized Shannon Entropy across clusters, PMD is
181 provably invariant to the number of clusters found when relative overlap in cluster composition is
182 preserved, operates linearly across the spectrum of batch similarity, is unaffected by batch size

183 differences or overall number of cells, and does not require that batches be similar (**ED Fig. 2**;
184 see **Methods** and **Supplemental Note-4** for proof, benchmark, and detailed characterization of
185 PMD; **Supplemental Figs. 7,8**).[29] PMD yields a single quantitative representation on a scale of
186 0-1 ranging from highly similar batches to highly dissimilar batches. A PMD=1 indicates batches
187 with no overlapping clusters (highly dissimilar), while PMD≈0 indicates highly similar batches,
188 say, from the same single cell suspension (see **ED Fig. 2b,c** for a visual representation and
189 calculation details).
190 Using PMD, we quantified batch similarity after correction, when cells came from the
191 same single cell suspension (technical-replicates), where the PMD is expected to be low or
192 near-zero (**Box 1c:** using technical replicates from neurons, heart, and PBMCs as model
193 systems)[30, 31]. Most algorithms (and no-correction) performed quite well at the cluster level by
194 PMD (**ED Fig. 3a-c**). The results were similar across a range of tissues, including mouse brain,
195 mouse heart, and human peripheral blood mononuclear cells (PBMCs) (**ED Fig. 3d,e**), with
196 none of the batch correction methods giving rise to a PMD greater than 0.04.
197
198 ***Efficacy with dissimilar biologic datasets***
199 We next performed negative controls (no expected cluster overlap) by comparing brain[31] and
200 intestinal epithelium[32], given that the only immune cells recovered after epithelial and crypt
201 enrichment were T-cells, and the brain dataset contained no detected CD3E or CD3G; however,
202 this is still an expectation of dissimilarity based on biological knowledge (**Box 1d**; **Fig. 3b**;
203 expected PMD=1). While Class-1 algorithms indeed indicated no overlap in cluster composition
204 (PMD=1), Class-2 algorithms with latent-dimension-mapping (LDM) had PMD scores up to 0.4,
205 which corresponds to 60% relative similarity in cell type composition between the brain and
206 intestine (**Fig. 3c**). Class-2+LDM methods indeed harbored mixed neuron/intestinal epithelial
207 clusters (>30% of clusters mixed in full datasets), while Class-1 approaches did not (**Fig. 3d,e**).
208

209 ***Results with biologic datasets of uncertain overlap***

210 In contrast to the above positive- and negative-controls, batches will typically give rise to
211 clusters of cells that have unknown overlap, and even cells of the same type may differ by cell-
212 state across batches. We therefore analyzed homeostatic intestine and irradiated intestine (**Box**
213 **1e**; **Fig. 4a**)[32], expecting some overlap given both tissue types began with the same base cell-
214 types, but then given cellular damage caused by radiation exposure, we would also expect
215 dramatically different cell-states, replication, and repair (**Fig. 4b**). Results ranged from PMD
216 values of 0.99 with no-correction, to PMD values of 0.20 (Harmony) (**Fig. 4c**). These results
217 demonstrate that the choice of correction method can drastically alter the biologic conclusions of
218 an experiment with batches of divergent cell-states.

219 To quantify the biological pathways that were subject to erasure by some batch-
220 correction approaches, we refined our WCAB approach to filter each cluster to include *only*
221 topologically overlapping regions of the pairwise batch-batch comparisons (**Fig. 4d**). We
222 performed this procedure because topologically overlapping regions are often interpreted as
223 containing biologically similar cells. This is in contrast to another possibility in which two
224 topological locations are "zippered" together or placed adjacent by a batch-correction algorithm
225 in a manner that creates an apparently contiguous object in similarity space, but where the
226 objects are better described as two conjoined, but non-overlapping objects. To identify and
227 distinguish these types of situations, we developed the open-source Topologically Overlapping
228 Within-Cluster/Across-Batch (TOWCAB) method (**ED Fig. 4**).

229 We applied the TOWCAB method to the homeostatic/irradiated intestine data to directly
230 quantify which attributes of cell identity were removed by the different batch correction
231 algorithms. We identified 9 clusters of pathways that were differentially expressed within
232 topologically overlapping areas produced by Class-2+LDM algorithms (**Fig. 4e**; **ED Fig. 5a**).
233 These included immunologic pathways related to antigen presentation, response to pathogens,
234 reactive oxygen species, and other processes that would be expected to be biologically

important during irradiation mediated disruption of the intestinal barrier (**Fig. 4e**). Importantly, these pathways were pre-filtered to exclude any pathway that was observed as significantly different in any of the technical replicates (**Fig. 2**), thus quantifying only pathways that had no prior evidence of being impacted by technical sources of variation.

Discrete sample types vs continuously divergent sample types

A remaining question is the impact of integrating samples with many replicates that originate from discrete sample-types or those that exist across a continuum of changing cell-states and cell-type abundances. To address the former, we integrated cortical and substantia nigra samples, each with biological replicates[33]. Pairwise sample-sample PMDs were calculated and displayed for within vs between region similarity (**Fig. 5a**); this show a clear separation in which samples from within a region (cortical from substantia nigra) were more similar relative to across region comparisons for all methods (including no-correction), with the exception of Seurat's MCCA (Class-2).

For clusters to 'split', however, a large effect must be present, meaning that the topologies may be slightly askew across batches, but not sufficiently distinct to induce cluster splitting. We assessed topological heterogeneity, via TOWCAB, revealing that Class-2 algorithms topologies were more heterogeneous than Class-1 (**ED. Fig. 5b**). These altered pathways included those related to neurons, synapses, dendrites, and ion transport, indicating that it was variation in these pathways that were erased in order to produce a "corrected" topology.

In contrast to the above (differential cell type abundance between two regions of the same tissue), another realistic scenario is the presence of dynamic continuous processes across batches. We therefore analyzed the endometrium across the menstruation cycle which has previously described changes in cell-states and cell-type abundances, particularly at the window of implantation[34]. Class-1 approaches were consistent with each other identifying these

261 samples as largely dissimilar, with changing cell states and cell-type abundances. On the other
262 hand, Class-2 algorithms increased sample-sample similarity (lower PMD) (**Fig. 5b**). TOWCAB
263 analyses found that Class-2 algorithms erased pathways that appeared biologically pertinent,
264 including angiogenesis, microvillus related structures, immune cells, and endometrial cells (all
265 well described aspects of the estrous cycle[35, 36]; **Fig. 5c**). Notably, variation of the Progestagen
266 Associated Endometrial Protein gene (*PAEP*; or Glycodelin-A), (which is dynamically regulated
267 throughout estrous[37]) is topologically erased by Class-2 algorithms along the axis of this cell-
268 state (**Fig. 5d**).
269 It is again important to note that these pathway analyses each used custom
270 backgrounds of locally expressed genes within clusters ensuring that this was not generic
271 enrichment for brain or endometrium related pathways relative to the whole genome, but were
272 rather locally enriched among the genes expressed in topologically overlapping regions of the
273 datasets, taking into account what *could have been* detected. Further, any pathways that were
274 previously shown to be differentially detected in our technical replicate studies were excluded
275 for this analysis, as removal of variance along these dimensions should not be penalized, as
276 this is the goal of batch-correction.
277
278 ***Erasure of cell-type signal within an atlas***
279 To differentiate the effects of our specific implementations of these algorithms from what is seen
280 in practice, we performed TOWCAB analyses on a recent release of human urinary bladder
281 cells (30,942 cells; 10 donors, aligned using Seurat's MCCA), based on their released UMAP
282 coordinates[38]. Indeed, we found that different cell-types were merged across batches including
283 B- and urothelial-cells (**ED Fig. 6**). This further highlights both the efficacy of TOWCAB analysis
284 in the discovery of biologically heterogeneous regions of the topology, and also highlights that
285 we cannot always rely upon cell-type labels from atlases as "ground-truth" labels for
286 benchmarking. We similarly observed what appeared to be B-cell-related droplets, that were

287 labeled as satellite stem cells, from a muscle sample, that were topologically placed within an
288 otherwise fibroblast related cluster within the *tabula sapiens* resource (based on stromal UMAP
289 coordinates following scVI correction; see **Supplemental Note-1, Supplemental Fig. 1**)[39].
290

291 ***Synthetic dataset benchmarking with simulated cell-types/-states***

292 Given these examples in which Class-2 algorithms appear to erase biological signal of interest,
293 we turned to simulations to further demonstrate this issue, where through the simulations
294 complete control of the ground-truth cell-lineage -type and/or -state can be achieved. To this
295 end, we simulated cell-states *de novo*, *in silico*, with a linear-mixed model of negative-binomial
296 distributions, coarsely representing a possible cell-type/-state mixture under various batch
297 confounded and non-confounded circumstances, comparing known ground-truth results to
298 results downstream of the different batch correction algorithms. Using this approach, with a
299 known ground truth (rather than reliance upon consistency with previously published labels), we
300 found that Class-2 algorithms largely erased batch confounded mixed expression programs
301 (**Supplemental Note-5**; **Fig. 6a-c**). These results mirror our biological findings that Class-2
302 approaches may over-merge discrete cellular populations, which should be either split by cell-
303 type (**Fig. 3,5**, **ED Fig. 5**) or cell-state (**Fig. 4**, **ED Fig. 5**).

304 By combining the results of our molecular cross-validation (higher percent-same),
305 technical replicates (lower PMD), negative control integrations (higher PMD), more complex
306 integrations (lower TOWCAB significant pathways), and simulations (closer to ground-truth), we
307 found that Scater's relative-log-expression (RLE) ranked the highest of methods assessed in
308 this work (**Fig. 6d**).
309

310 ***Real world implications of mixed-model based over-mergers by Class-2 algorithms***

311 Given that we had found *in silico* linear mixture models create shared sources of variation that
312 enable over-mergers by Class-2 algorithms, we hypothesized that *biophysical* linear-mixtures

313 (doublets or cell-fragment mixtures in a droplet[40, 41]) may act as “seeds” for over-integration, with
314 contaminated droplets masquerading as a “novel cell-type” with unexpected marker gene
315 positivity. Given our above findings, application of Class-2 integrations may result in apparent
316 topologic-replication of these results (biologically heterogeneous cells intermingled topologies
317 containing mixed droplets).

318 For example, CD19 (a canonical B-cell marker) has recently been described as
319 expressed in a subset of human brain mural cells taking this as evidence that an anti-CD19
320 CAR-T cell therapy would also kill these brain cells[42]. This finding appeared to be confirmed
321 following a Class-2 batch-correction (Seurat’s MCCA) with other datasets, after which they
322 observed co-clustering which was interpreted as a repeated discovery of this population[42].

323 We performed in-depth analysis of these datasets including examination of 1) total
324 counts, 2) number of observed genes, 3) creation of linear mixed models of “pure” B-cell and
325 “pure” mural cells comparing these models to the global transcriptome of the candidate
326 CD19+/mural-droplets, 4) transcriptome-wide correlations with CD19, and 5) expression of
327 marker genes (See **Supplemental Note-6**; **Supplemental Fig. 11**). All of these analyses were
328 consistent with a model that the originally observed CD19+/mural-marker co-positive droplets
329 were simply a mixture of transcripts from B- and mural-cell lineages. Indeed, in the original
330 dataset nearly all droplets were contaminated with transcripts from multiple lineages
331 (**Supplemental Note-6**). Lastly, we analyzed a recently released atlas of fresh human brain
332 vasculature and associated cells; we observed only 1 out of 16,681 droplets that were positive
333 for both CD19 and mural cell markers, while also being negative for other B-cell markers
334 (**Supplemental Fig. 11**)[43]. We hypothesized that the appearance of independent replication
335 through co-clustering following MCCA (**Fig. 6e**) may have been an over-merger that may not
336 replicate at the bench.

337 While peroxidase based anti-CD19 immuno-histochemistry in the original report showed
338 peroxidase signal in some cells near or in vasculature, there was no co-stain with B- or mural-

339 cell markers. To test whether CD19 is in fact expressed in mural cells, we co-stained for CD19,
340 CD79a (pan-B-cells), and PDGFRβ (mural cells) in four human brains using no-primary negative
341 controls to set thresholds for CD19 positivity. We did not observe cells that appeared positive for
342 CD19 and PDGFRβ that were also negative for CD79a, despite CD19 staining positive controls
343 working well (human tonsil) (**Fig. 6f-h**). Notably however, a negative finding can never be
344 definitive; given the exceptional clinical importance of CD19 expression in off-targets in the brain
345 for anti-CD19 CAR-T-cell therapies, these findings should be further interrogated.
346 Overall, our findings confirm the necessity of checking the assumption that topologically
347 mixed cells are biologically similar after Class-2 alignments, through methods such as
348 TOWCAB, or Class-1 algoirthms my be applied, with the best performing method by our
349 benchmark being Scater's implementation of relative-log-expression (RLE)[44, 45].
350

351 **Discussion**

352 The field of scRNAseq has been evolving rapidly, both at the bench and computationally.
353 Here we have shown that the most commonly used metrics for benchmarking methods of batch-
354 correction often rely on either 1) the assumption of biological equality across different samples,
355 or 2) the assumption that humans who performed the analysis to generate the reference labels,
356 were successful in optimally finding and labeling all meaningful attributes of biological variation.
357 The methods that rely on the second assumption reward for *similarity* to the methods used by
358 the original experimenters rather than accuracy. While all methods are subject to biases and
359 losses of accuracy given deviations from underlying assumptions realized in practice,
360 understanding the sensitivity of a method to such deviations is critical in achieving accurate
361 interpretations of method outputs. Here we have attempted to highlight a range of issues that
362 can potentially arise in seeking accurate interpretations from single cell RNAseq data when
363 assumptions for different batch correction algorithms are unknowingly violated.

364 Given the premise that scientific measurements of biologic systems are manifest from
365 the biological- and technical-functions, where, importantly, the biological-functions are what we
366 aim to *measure*, a more complete understanding of the technical functions at play in complex
367 experiments such as scRNAseq, is needed to disentangle them (either computationally or
368 experimentally) when batch and biological variation are intrinsically confounded. To this end, we
369 described a functional framework to help formalize the issues, which ultimately led to our
370 identification of differentially detected transcripts encoding proteins that appear to localize to
371 large structural components of the cell, (and hemoglobin when RBCs have not been removed),
372 as strong contributors to the technical effects.
373 We also fill a gap in our ability to accurately quantify batch overlap based on cluster
374 composition, which required our development of the PMD measure: a measure that is provably
375 invariant to the number of clusters (of course under its own set of assumptions), is linear with
376 respect to similarity, and robust to dataset size or differences in batch sizes (**Supplementary**
377 **Fig. 7,8**). These properties, for the first time, enabled a quantitative assessment of
378 normalization/batch-correction algorithms across a full spectrum of similarities based on cluster
379 composition without requiring the assumption of biological equality between biological
380 replicates, or that near perfect annotation of the underlying biology was achieved by the
381 experimenters.
382 From the perspective of topologies, we introduced (TO)WCAB analyses that quantify
383 differentially expressed/detected genes and pathways, specifically within topologically
384 intermingled areas across batches. Our results suggest that when using unsupervised batch-
385 correction algorithms, TOWCAB-like analyses should be performed and reported so that
386 experimenters, reviewers, and readers can understand which attributes of gene expression
387 were removed by the algorithm. Such understanding can protect against misinterpretation that
388 adjacent cells within a topologic embedding are biologically similar, when they are in fact
389 different. Our results also motivate exercising further caution against interpreting topological

390 concordance as an indication of similarity after Class-2 batch-correction, especially when
391 comparing treatments, genotypes, or other condition groups. Topological concordance as an
392 indication of similarity has has become increasingly common practice [42, 46]. Similarly, WCAB
393 can be employed when interpreting co-clustering after Class-2 algorithm application.
394 Importantly however, Class-2 approaches remain the only viable option when integrating
395 across sequencing technologies or across species. Despite this, our results demonstrate that
396 these algorithms cannot be blindly employed without checking for biological heterogeneity
397 across apparently contiguous topologies. We also provided new tools and approaches, enabling
398 the means to check these assumptions (PMD, WCAB, and TOWCAB). Lastly, we show that
399 these are not theoretical concerns, but may have already had direct impact on atlases and
400 research into clinical treatments for cancer. We further hope that the community will build
401 technical replicate atlases to enable explicit modeling and removal of technical effects in single
402 cell -omics.

Acknowledgements

Support for this work was provided by T32CA078207 (SRT), K99HG011270 (SRT), , U01 AG046170 (EES), and RC2 DK122532 (EES).

Author Contributions

SRT conceived of and performed all analysis, bench-work, and wrote manuscript. EG provided feedback and edited figures. EES wrote Proofs 1 and 2, and guided research and edited manuscript.

Competing Interests Statement

Authors declare no conflicts of interest.

415 **Methods**

416

417 ***Biologic datasets: 10x Genomics datasets***

418 All datasets, baring those of intestinal origin were obtained from Chromium's publicly available

419 datasets.

420 PBMCs:

421 (https://cf.10xgenomics.com/samples/cell-

422 vdj/3.0.0/vdj_v1_mm_c57bl6_pbmc_5gex/vdj_v1_mm_c57bl6_pbmc_5gex_filtered_feature_bc_

423 matrix.h5)[47]

424 1k neurons:

425 (https://cf.10xgenomics.com/samples/cell-

426 exp/3.0.0/neuron_1k_v3/neuron_1k_v3_filtered_feature_bc_matrix.h5)[31]

427 10k neurons:

428 (https://cf.10xgenomics.com/samples/cell-

429 exp/3.0.0/neuron_10k_v3/neuron_10k_v3_filtered_feature_bc_matrix.h5)[30]

430

431 ***Other biologic datasets***

432 The intestinal datasets, which were aligned to the Ensembl reference, as with those from

433 Chromium, were previously published.[32] Several datasets from the human cortical and

434 substantia nigra[33] and the endometrium were also previously released[34]. However, all raw input

435 datasets used here are re-distributed in the benchmarking repository for ease of replication.

436 Human brain datasets are distributed as an Rdata file within the data/cd19_mural_cell_data

437 directory of the repository. Any other public datasets are downloaded in-line within the code in

438 the repository.

439

440

441 ***Class-2 algorithm literature mining for co-mentions with the term "batch-correction"***

442 Searches on google scholar were performed to quantify the relative frequency at which
443 Class-2 algorithms were co-mentioned with the term batch-correction. Searches were
444 performed on July 24th, 2022. The number of results containing "single cell RNAseq" were used
445 as the baseline "total possible" results, then search terms were added to this base including with
446 and without the term "batch-correction " and with each Class-2 method included in this
447 manuscript ultimately to tabulate a contingency table for each method within the subset of single
448 cell results that either mention or don't mention the method at hand and along the other
449 contingency table axis either mention or don't mention the term "batch-correction." Results are
450 contained in **ED Table 1**.

451

452 ***Global dataset downsampling***

453 For the benchmarks that included sub-sampled "child cells," this 50% transcript sub-
454 sampling of the dataset was created using the pyminer_norm package (available by pip:
455 python3 -m pip install bio-pyminer-norm) using the command line call:

456 python3 -m pyminer_norm.random_downsample_whole_dataset -i <dataset.tsv> -o
457 <dataset_ds50.tsv> -percent 0.5

458

459 ***Implementation of normalization approaches***

460 *No normalization*

461 For no normalization, the datasets are simply concatenated by genes with no other
462 modifications.

463

464 *UMI downsampling*

465 We created a new python package that is pip installable called bio_pyminer_norm to
466 provide a parallelized and scalable implementation of UMI downsampling. This approach

467 requires the user to determine lower (and optionally upper) bounds of the sum of the UMI
468 counts and total number of observed genes for which cells will be included or discarded. As
469 implemented here, we used 3 cutoffs to show the impact of changing sensitivity with this
470 parameter. Cutoffs were as follows for all biological datasets: low:$10^{3.55}$, mid: $10^{3.6}$, high: $10^{3.65}$
471 all with a minimum number of observed features of $10^{3.1}$; for simulated datasets (which were
472 deeper), cutoffs were: low:$10^{3.65}$, mid: $10^{3.7}$, high: $10^{3.75}$ all with a minimum number of observed
473 features of $10^{3.15}$. Next, each cell is bootstrap-sampled, having its observed transcriptome
474 simulated, shuffled, and an equal number of transcripts randomly selected. This is
475 downsampled to the lower cutoff value for total sum UMI.
476
477 *Count Normalization*
478 Cells/droplets were normalized as counts per x in which x is an arbitrary scalar. In our
479 implementation, we use three different cutoffs, which were the same as implemented for
480 downsampling.
481
482 *Scater-RLE*
483 Following guidance from the Scater repository, we implemented this normalization
484 approach using the following call: scater::normalize(exprs, exprs_values = "counts", return_log
485 = TRUE, log_exprs_offset = NULL, centre_size_factors = TRUE, preserve_zeroes = FALSE).[19]
486
487 *Combat-seq Negative Binomial*
488 Combat-seq with Negative binomial distribution is distributed in the sva-devel package.
489 Following author suggested tutorial, we used the following implementation: ComBat_seq(exprs,
490 as.character(batch), group=NULL, full_mod=FALSE)[20].
491
492 *Seurat-v3 Multi-CCA*

493 Following the SeuratV3 website tutorial on integration, we first normalized and
494 processed each dataset using: NormalizeData then FindVariableFeatures. Next, we used the
495 functions: FindIntegrationAnchors and IntegrateData to complete the MCCA data integration in
496 SeuratV3; as recommended, clustering was performed on the first 30 PCs[9].
497
498 *Harmony*
499 To use Harmony, we followed the author guidelines available on the repository website,
500 first normalizing counts to counts per thousand, then log2 transforming the matrix, next genes
501 were min-max scaled to range between 0 and 1. We used the following call:
502 HarmonyMatrix(exprs, batch_table, "batch", do_pca = TRUE)[10].
503
504 *Liger*
505 Following the Liger repository tutorial, we used the createLiger function on the input list
506 of datasets to be normalized; next we utilized the liger::normalize function, followed by the
507 selectGenes, and scaleNotCenter functions.
508 As per author guidance on their repository tutorial, we used the Liger suggestK function
509 to generate a K vs median KL divergence curve. They authors suggest using the 'elbow-rule' of
510 this curve, however, given that there is no actual relationship between the units of number of
511 clusters and KL-divergence, we implement a K selection method that sums the area under that
512 curve (AUC), and select the K that first reaches 95% of this AUC. This K was used for the
513 optimizeALS function. Lastly, we used the quantileAlignSNF function to complete Liger based
514 batch-correction.[11]
515 Given that Liger only yields a low-dimensional representation of the data, no additional
516 feature selection was performed and we used this representation for clustering directly.
517
518 ***WCAB differential expression***

519 To determine differential expression within clusters, we used the recently released Libra
520 package with pseudo-bulking and pseudo-replicates[23]. We used three pseudo-replicates for
521 each batch, randomly partitioning them to contain at least 3 cells per pseudo-replicate, giving a
522 minimum requirement of 9 cells per batch to be able to perform DE analysis for each pairwise
523 batch comparison. The default EdgeR flavor of DE analysis was used. For comparisons with
524 more than two batches, a gene was considered differentially expressed if it was globally
525 significant by Benjamini-Hochberg FDR $q<0.05$ for any pairwise comparison.
526
527 ***Topological filtering for TOWCAB analysis***
528 The topologies used for filtering were the same graphs used for Louvain clustering in all
529 cases, except for when topologies were built from published coordinates; this is described in its
530 own section.
531 For each batch-batch comparison, each cluster was subset. Within each cluster, each
532 possible batch-batch pair is compared; within this subset, cells were quantified to assess the
533 relative proportion of edges shared with droplets of their "host" batch, or the other batch it is
534 being compared to. For each cell, 100 bootstrap shuffled edges were used to generate a cell-
535 wise background to empirically measure the host-batch proportion with fully entropic mixing
536 would yield for that given cell. Cells were kept that were within 1.5 standard deviations from the
537 mean of the empirical background (indicating entropic mixing). Cells were also excluded that
538 had >90% or <10% host-batch edges within the sub-network as these distantly connected cells
539 could be spurious and not represent a large scale topological pattern.
540
541
542 ***Illustration of assumptions made by standard benchmark metrics***
543 The human 10k PBMC dataset from 10x genomics (and included within the data zip from
544 the repository) was first analyzed to identify "primary" clusters which served as the "cell-state

545 not annotated" references. Next, this dataset was split into random thirds, with each third getting
546 its own cell state signal added.
547
548 *Primary cluster identification*
549 Using Seurat V3, we calculated the normalized dispersion values, using the top 2000
550 over-dispersed features which were used for clustering via PyMINEr, in which the cell-cell
551 Spearman correlation matrix was calculated, which was then fed into the cell-cell Euclidean
552 distance calculation. These distances were then used to create the locally weighted kNN
553 followed by Louvain modularity as described in detail later in this methods section.
554
555 *Cell-state spiking using underdispersed template genes*
556 Some have questioned the utility of use of simulation based on our assumed
557 parameterized distributions. To circumvent this concern, we used actual gene expression
558 observed from within this dataset as templates. Given that we used a split of three, we desired
559 three brackets of low, medium, and high expression in which we randomly assigned a bracket
560 (low/mid/high) to each gene from the gene-set specific to each data split such that for gene-1 all
561 of the cells from split-1 received low-template, split-2 received high, split-3 received mid; and
562 gene-2 cells of split-1 received high, split-2 received mid, split-3 received low, etc. Each bracket
563 to split assignment was performed randomly so that total counts were not wildly different
564 between splits as would happen by assigning all of split-1's spiked-in signal to the low bracket,
565 split-2 to mid, split-3 to high, etc.
566 The number of genes that were needed as templates (n) from each bracket was equal to
567 the number of target genes (defined in the next section). Summary statistics at the gene level
568 for expression and variability were obtained through Seurat using the HVFInfo function. Non-
569 expressed genes were removed, then genes were ranked by their mean expression and
570 converted into percentiles. The low bracket was defined as beginning at the 15th percentile, mid:

571 65th, high: 92.5th. Next, 2-times the number of required genes (n) were selected moving up from
572 these lower bounds. Each bracket's genes were then annotated as the n genes within the 2n
573 selection that had the lowest dispersion values. Each genes' bracket and dispersion status are
574 visually displayed in **Supplemental Note-1**; **Supplemental Fig. 1**.
575
576 *Identification of genes to target for spike-ins*
577 We aimed for these spike-in genes to be of biologically cohesive and meaningful signal.
578 To this end, we begin with a list of all genes from several go pathways relevant to MEK/ERK/
579 JAK-STAT signaling (GO:0000165, GO:0070371, GO:0007259), T-cell Activation
580 (GO:0042110), T-cell differentiation (GO:0045580), and immune system process
581 (GO:0002376). Genes of these pathways were collated via gProfiler, then filtered for only those
582 genes that were not previously annotated as over-dispersed and used for identification of the
583 primary clusters. This prevents any data leakage because the original labels were generated
584 agnostically of the expression of these non-over-dispersed genes from these biologically
585 meaningful pathways. Similarly, the above described template genes did not include any of
586 these target genes to prevent self-sampling, again towards the goal of preventing any data-
587 leakage.
588
589 *Spiking signal into only a single population*
590 In the experiments where we spiked signal into only a single cluster, we identified this
591 "primary cluster" as described above that showed the highest expression of *CD3E*, a pan-T-cell
592 marker. All other attributes of this analysis were identical to those described above, except for
593 the fact that only this cluster was annotated as being split into thirds, and correspondingly, only
594 this cluster received the spiked in signals to each of the three splits.
595

596 *Assessment of status quo benchmarking metrics for the apparent performance of methods with*
597 *cell-state annotated vs non-annotated labels*
598 The three splits of the dataset were integrated using the same pipeline as described in
599 the above section. We assessed: Adjusted Rand Index (ARI) by batch, ARI by celltype,
600 normalized ARI by batch, normalized ARI by cell type, fscore ARI, average within distance,
601 "average distance to nnkth nearest neighbor within cluster"[48] (mnnd), "coefficient of variation of
602 dissimilarities to the nearest within-cluster neighbor measuring, uniformity of within-cluster
603 densities, weighted over all clusters"[48] (cvnnd), maximum cluster diameter, widest gap between
604 clusters, separation index, minimum separation between cluster, average silhouette window
605 (ASW), dindex ("this index measures to what extent the density decreases from the cluster
606 mode to the outskirts"[48]), denscut ("this index measures whether cluster boundaries run through
607 density valleys"[48]), highdgap ("this measures whether there is a large within-cluster gap with
608 high density on both sides"[48]), "correlation between distances and a 0-1-vector where 0 means
609 same cluster, 1 means different clusters"[48] (Pearson gamma), within cluster sum of the squared
610 distances, entropy, average distance to the centroid of a cluster (pamc), average between
611 cluster distance, minimum of average distance to other cluster, minimum separation between
612 clusters, minimum of the average cluster silhouette width, lowest iLISI, average iLISI, highest
613 cLISI using cluster labels with either the spiked cell state observed or not, average cLISI with
614 spiked cell state observed or not, highest cLISI using "observed state" labels, average cLISI, the
615 kBET average rejection rate, kBET-cell-type rejection rate, kBET-cell type with observed cell-
616 states, mean and max kBET within primary clusters, mean and max kBET within cell-state
617 annotated labels. R packages used for these metrics included: kBET[29], lisi[10], and fpc[48]. All
618 quoted descriptions come from the fcp package documentation.
619 These metrics were compared for when cell state was "observed" vs "not-observed"
620 based on primary clusters, or those split into the thirds as described above. The Spearman and
621 Pearson correlations between metrics were quantified within these different cluster labels being

622 used as noted in **Supplemental Note-1; Supplemental Fig. 2**. Given the large number of
623 metrics quantified, we display only several of interest that showed discordant results when the
624 cell-state was annotated by the experimenter or unknown to the experimenter.
625
626 *Assessment of the impact of PCA reduction in generation of the reference primary clusters*
627 The entire above procedure was repeated a second time, but first using reference labels
628 that were generated by taking the first 30 principal components, then using them for input to the
629 symmetric Euclidean distance matrix calculation prior to locally weighted kNN creation and
630 Louvain modularity. This is in contrast to using the cell-cell Spearman correlation matrix as input
631 for the calculation of the Euclidean distance matrix, thus preventing the dataset from ever being
632 compressed into a low dimensional space.
633 Similar to comparing cell-state observed vs not-observed, we also quantified the
634 correlations between PCA reduced vs non-reduced when generating reference labels as shown
635 in **Supplemental Note-1**; **Supplemental Fig. 2**.
636
637
638 ***Identification of T_m complexes through StringDB***
639 Once the complete list of DEGs was collated for a given technical replicate comparison
640 via WCAB analysis, within cluster Spearman correlations were performed within those DEGs,
641 using a bootstrap shuffled null background to account for any non-normal distributions inherent
642 to scRNAseq data, using a false positive rate cutoff of 0.05. These correlations were
643 represented as an undirected graph network using iGraph[49]. This graph was then partitioned
644 into connected components. The genes from each component were then pulled from the
645 StringDB network of protein-protein interactions (PPI) graph networks[25]. These PPI networks
646 were then used to create the final co-regulated (based on correlation) physically interacting
647 graphs (based on the StringDB annotations). These graphs were then subjected to Louvain

648 modularity to identify highly connected sub-networks which were then annotated by performing
649 gProfiler analyses[24], using the globally expressed genes in these datasets, that were *also*
650 present within the StringDB database as the custom background, therefore preventing any
651 generic enrichment for database annotation or high expression. When merging Tm lists
652 between mouse brain and heart datasets, in some cases a gene appeared in different T_m
653 complexes (due to slight differences in the boundaries found between clusters as called by
654 Louvain modularity); to prevent double-counting these genes, it was assigned to the smaller
655 complex. Consensus T_m annotations were arrived upon by taking the intersection of the T_m
656 annotated genes from the mouse brain and heart technical-replicates.
657
658 ***Quantification of non-technical DEG pathways***
659 We aimed to be very conservative in identifying putative “biologically important”
660 pathways that were differentially expressed across non-technical replicate comparisons. To this
661 end, we performed within-cluster/across-batch DEG analysis as described above followed by
662 gProfiler pathway analysis using cluster-level custom backgrounds based on what was
663 expressed within this cluster. These significant pathways were then filtered to remove any
664 pathway that appeared at all in the union of both brain and heart technical-replicates. Given the
665 premise that all observations are the union of biological and technical, removing the technically
666 differentially expressed pathways leaves the biologically differentially expressed pathways.
667
668
669 ***scRNAseq simulations***
670 *Simulating cell-types*
671 Using Splatter,[50] we simulated 1500 cells with 10,000 genes using the parameters:
672 "de.downProb"=.25, "de.prob"=.75, "de.facScale"=.75, "bcv.common"=.75. All other parameters
673 were left at their default levels. Each cell was uniformly randomly assigned to one of 6 clusters,

674 and to either batch 1 or batch 2, depending on the details of the simulation as noted in
675 **Supplementary Fig. 9**.
676
677 *Simulating cell-states*
678 To simulate a cell-state module, a single 'cluster' was simulated with the same
679 dimensions as the original dataset. All cells were then paired with a cell-state transcriptome.
680 25% of genes were then considered to be members of the cell-state. A gaussian weighting
681 vector was generated using rnorm * 1/50 to shrink variance, and shifted to center around 0.25.
682 Each gene belonging to the cell-state was assigned to a weight from this distribution. A linear
683 mixture of the cell transcriptome and states is then created to form the final cell, using weights
684 for the mixture similar to above, mixing cell-type and state with random weights either centered
685 at 33.3% state or a half-Gaussian whose lower bound was zero for cells that were being
686 injected with the cell-state or not respectively.
687
688 ***Statistical analyses***
689 Statistics comparing percent-same, PMD, relative mutual information, purity, and reverse
690 purity across methods were performed using an ANOVA via the aov function in R on rank
691 transformed versions of the metrics, given that variance was different across groups:
692 aov(rank(<metric>) ~ method)
693 For analysis of simulations, the number of groups shared across batches were also modeled as
694 a covariate:
695 aov(rank(<metric>) ~ method + num_groups_shared)
696 TukeyHSD was used for post-hoc comparisons. All statistics were analyzed as two-sided
697 comparisons.
698
699 ***Clustering***

700 *Feature selection prior to clustering*

701 Because the goal of this work was to examine the real-world realistic implementations of
702 a standard bench-biologist following tutorials as advertised, feature selection algorithms were
703 implemented as described in each methods' tutorials. When not available, and the corrected
704 version of the transcriptome as a 'corrected expression matrix,' automated highly variable gene
705 selection, and clustering was applied equivalently to all methods that yield corrected
706 transcriptomes, as previously reported by PyMINEr[51]. The mean variance relationship was fit by
707 a lowess locally weighted regression curve. Residuals were calculated, and those genes whose
708 residual was ≥2 standard deviations above the loess fit curve were selected. For methods
709 yielding latent dimensions, these dimensions were used for clustering rather than applying
710 additional feature selection as this is typically performed *before* latent dimension mapping.

711

712 *Clustering*

713 Clustering is performed similar to standard practice with Louvain modularity, with one
714 modification; instead of approximating similarity space through compression of data into a latent
715 space, we measure actual similarity/distance space, staying in high dimensions without latent
716 dimension compression. The locally weighted kNN used for Louvain modularity was constructed
717 as follows:

718 - Calculate the symmetric pairwise Spearman correlation matrix of each cell against all
719 other cells
720 - Calculate the squared Euclidean distance of all cells to all other cells using the
721 Spearman correlation matrix as input.
722 - For each cell, mask the 95% most dissimilar cells (or all but the 200 closest cells,
723 whichever gives the lower number of connections), and the diagonal, self-distance=0,
724 thus creating a local distance matrix.

725 ● For all remaining cells that were not masked, the inverse Squared Euclidean distances
726 were linearly normalized between 0 and 1 within each cell. This creates a local weighted
727 distance matrix, which is then converted to a weighted graph, treating the local distance
728 matrix as a weighted adjacency matrix, which is then subjected to Louvain modularity
729 based clustering.
730
731 This process is fully automated in PyMINEr,[51] hyperparameter free, and implemented
732 equally to all batch-correction methods, using the -louvain_clust argument. Louvain modularity
733 was used via the python-louvain package.
734
735 ***TOWCAB assessment of published atlas based on published coordinates***
736 These analyses were performed with the script called bin/from_coords.R within the main
737 repository. Coordinates were converted to a kNN with k=ceil(sqrt(n)), where n is the number of
738 cells in the dataset using the FastKNN package's k.nearest.neighbors function based on the
739 cell-cell Euclidean distance matrix calculated by the pdist function from the rdist package. This
740 kNN and the corresponding expression matrix were fed into the run_towcab_analysis function
741 within the towcab package. Figures were made using the UCSC Cell Browser
742 (https://cells.ucsc.edu) browser that hosts the data[52].
743
744 ***Clustering performance metrics***
745 The "percent-same" metric that was implemented in some biologic benchmarks is
746 calculated by taking the pairs of cells from the original dataset, and its downsampled
747 counterpart, and quantifying the number of pairs that were placed into the same cluster
748 (numerator), relative to the total number of pairs (denominator). This metric only quantifies cells
749 that are present in both datasets; if a parent or child cell was removed by a pipeline for quality

control purposes, they were not counted given that it would not have a matching cell-pair to assess if the cluster was the same.

Mutual information was calculated using the mi.empirical function from the entropy R package. Relative mutual information is the ratio of the observed mutual information to the theoretic maximum mutual information, if the results were 100% accurate relative to the ground-truth clusters.

Purity was calculated using the purity function from the NMF package in R.[53] The purity function takes in two arguments, first: the cluster results, and second: the ground-truth clusters. Purity is a metric that takes the method's clusters, identifies its plurality ground-truth cluster, then returns the percent of all cells that fall into a method's cluster whose plurality is the same ground-truth cluster. This particularly penalizes for merging ground-truth clusters together.

Reversing the arguments of the purity function does the opposite however, we therefore use reverse purity to quantify splitting ground-truth clusters (reverse purity).

The "true percent max difference" is reported in **ED Table 1**; this metric calculates the PMD based on the ground-truth known clusters in the simulated batches. This metric is used in combination with the percent maximum difference calculated using the observed clustering results for each method. The log ratio percent maximum difference is $\ln(PMD_{observed}/PMD_{ground\text{-}truth})$. Difference in percent maximum difference is $PMD_{observed}-PMD_{ground\text{-}truth}$. Absolute difference in percent maximum difference is $|PMD_{observed}-PMD_{ground\text{-}truth}|$.

The "state merged with base purity across batch" is the typical purity calculation, but first filtered to include only cells from clusters for which the base cluster appears both with and without the state added, thus penalizing specifically for merging clusters that belong to the same base cluster with and without a state. This quantifies an algorithm's propensity for erasing batch confounded 'states' or expression programs.

774 The “state merged with different base purity across batch” quantification is the purity
775 function applied specifically to the filtered subset of ground-truth clusters with the cell-state
776 added that did not belong to the same base “cell-type.”
777 The “split by batch reverse purity” metric is the reverse purity metric but selectively
778 applied to the subset of cells that belonged to ground-truth clusters that appeared in both
779 batches, whereas cell clusters that appeared only in a single batch were removed.
780
781 **Reanalysis of brain datasets for CD19+/mural-marker+ populations**
782 Because no code or datasets were released with the original manuscript[42], we could not
783 re-perform the exact procedures used there. Where did however perform our own independent
784 analysis of the same datasets which were release elsewhere in their original papers. All analysis
785 and code is contained in the “bin/cd19_mural_cell_analysis.R” script in the primary benchmark
786 repository (https://bitbucket.org/scottyler892/sc_norm_bench_v2). Reprocessed data is included
787 in the data distribution Rds or is downloaded on the fly within the script.
788
789 *Supervised annotation of B-cell-related and mural-cell related droplets and candidate*
790 *CD19+/mural-marker+ copositive candidate droplets*
791 To provide a strict conservative definition of droplets that were the most “pure” in
792 identity, we defined conservative “pure” B- and mural-droplets as those containing ≥5 UMI for
793 any of the indicated markers, with 0 sum expression of across-lineage markers: B-cell markers
794 were: (*Vpreb1, Vpreb3, Ighm, Ighg3, Ighe, Ighg1, Ighg3, Ighd, Igll1, Sdc1, Tnfrsf8, Ms4a1,*
795 *Cd79a, Cd79b*); mural-cell markers were: (*Cd248, Rgs5, Foxf2, Pdgfrβ, Cspg4).* The ‘candidate’
796 populations were defined as droplets with any observed expression of CD19 and simultaneous
797 detection of a mural cell marker. The “clearly mixed” populations were defined as cells with ≥5
798 UMI for at least one mural marker, but also showed ≥10 UMI total expression of B-cell related
799 genes (listed above), lymphocyte/leukocyte related genes: (*Ptprc, Cd38, Cd34, Cd3e, Cd3d,*

800 *Cd3g, Cd3a, Il2rA, Cd28, Ctla4, Cd4, Cd8a, Cd8b; Cd14, Cd80, Cd84, Cd86, Tnfrsf4*), or MHC-
801 II genes [mouse:(*H2-K1, H2-Ke6, H2-Oa, H2-DMa, H2-DMb2, H2-DMb1, H2-Ob, H2-Ab1, H2-*
802 *Aa, H2-Eb1, H2-Eb2, H2-D1, H2-Q1, H2-Q2, H2-Q4, H2-Q6, H2-Q7, H2-Q10, H2-T24, H2-T23,*
803 *H2-T22, H2-T3, H2-M10.2, H2-M10.1, H2-M10.3, H2-M10.4, H2-M11, H2-M9, H2-M1, H2-*
804 *M10.5, H2-M10.6, H2-M5, H2-M3, H2-M2*), human:(*HLA-DMA, HLA-DMB, HLA-DOA, HLA-*
805 *DOB, HLA-DPA1, HLA-DPB1, HLA-DPB2, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2,*
806 *HLA-DRA, HLA-DRB1, HLA-DRB5, HLA-DRB6*)].
807 Marker genes of other lineages were used only for heatmap displays rather than
808 analysis or identification of populations. These included the following markers: [astrocyte:(*Gfap*);
809 erythroid:(*Hbb, Hbg1, Hba1*); oligodendrocyte precursor cells(OPCs):(*Sox10*); microglia:(*Csf1r*);
810 endothelial:(*Cdh5, Pecam1, Vcam1*); excitatory neurons:(*Neurod2*, *Neurod1*); inhibitory
811 neurons:(*Lhx6, Gad1, Dlx5, Dlx2*).
812
813 *Number of observed genes and counts*
814 The number of observed genes and total counts per droplet have been previously
815 described as a characteristic of doublets/multiplets[54], and was therefore analyzed and found to
816 be higher in candidate CD19+/mural-marker+ copositive droplets as noted in the manuscript.
817
818 *Transcriptome correlation and MSE with 50:50 linear model of "pure" B-cell and "pure" mural-*
819 *cell transcriptomes*
820 After "pure" B- and mural-related droplets were identified, the average B- and mural-
821 related transcriptomes were calculated. Then the average log2(counts+1) of these two
822 transcriptomes was taken as the putative 50:50 mixture. Note that to account for differing
823 abundance of these populations, we took random samples of droplets to equal depth (minimum
824 count of either population of cells) over 20 iterations to generate a distribution of fits from
825 random subsets of each reference controlling for an equal number of cells in each reference.

826 This accounts for any excess “smoothing” and lower error that would result simply from
827 sampling a greater number of cells. This transcriptome vector was used as a reference for both
828 Spearman correlations and mean squared error (MSE) calculations with a ‘test’ population. This
829 ‘test’ population was either the candidate CD19+/mural-marker copositive droplets, or a random
830 subset of droplets from the “pure” mural population. Each ‘test’ population was quantified for its
831 fit with sampling from a single cell through to all candidate cell populations, and a matching
832 population of control droplets from the “pure” mural population (held out from references) (x-axis
833 on **Supplemental Note-6; Supplemental Fig. 11c**); assaying this range of sample depth also
834 controlling for the any “smoothing” differences seen by simply sampling more cells. This
835 comparison allowed us to directly compare whether the CD19+ candidates are globally more B-
836 like than the pure mural population.
837 Note that in the scenarios where we fit the “pure” mural subsets to the 50:50 reference,
838 the test subset was excluded from the “pure” mural population used to generate the reference
839 average, thus preventing data leakage. All marker genes used for identification of these
840 populations were excluded from these analyses to both 1) ensure no data leakage, and 2) that
841 we were quantifying the broad transcriptome patterns independently of the expression of CD19,
842 further indicating that these cells were not simply mural cells that expressed CD19.
843
844 *Fitting a linear mixed model of “pure” B- and mural-references*
845 A linear model of the average log2(counts+1) candidate and reference pure-mural and
846 pure-b transcriptomes was calculated as described above, but instead of assuming a 50:50
847 mixture, we quantified the relative proportion of these references that explain the candidate
848 transcriptome by fitting a linear mixed model: candidates = b_1*pure_mural_reference +
849 b_2*pure_b_reference, with an intercept of 0. The b_1 and b_2 coefficients were calculated by the
850 FLXglm function from the flexmix package[55] with the offset argument set to zero, fixing the
851 intercept, thus preventing the attribute of overfitting that result from different sequence depth.

852 Every gene not included in the marker-sets used to identify these populations was quantified
853 independently, yielding a measure at the per-gene level quantifying the contribution from each
854 lineage. Relative contributions of each "pure" population was calculated as $b_1/sum(b_1+b_2)$ and
855 $b_2/sum(b_1+b_2)$, removing the net impact of differential 'count-depth' for specific cell subsets. For
856 example, a random Poisson sub-sampling of a single transcriptome vector by 50% may
857 symmetrically decrease the beta-coefficients by half, yet their relative contribution would remain
858 the same ± random Poisson sampling error.
859
860 **Protein staining on human brain and tonsil**
861 *Samples*
862 Samples were obtained from BioMax USA (Human tonsil: HuFPT161, Brain: HuFPT017,
863 HuFPT018, HuFPT019, HuFPT020).
864
865 *Sample preparation*
866 Slides were baked for 1hr at 50C, then deparaffinized as follows: Xylenes 10 min, 2x
867 100% EtOH, 10 min, 2x; 95% EtOH, 5 min; 70% EtOH, 5 min; 50% EtOH, 5 min; rinse in diH2O;
868 PBS, 10 min.
869
870 *Antigen retrieval*
871 Antigen retrieval was performed using citrate buffer in an instapot pressure-cooker.
872 Pressure cooker was filled with fresh diH2O, with a vegetable steamer insert placed inside to
873 hold plastic staining jar containing slides submerged in 220mL of 1x citrate buffer made from
874 10x concentrate (Invitrogen #00-4955-58). Slides were subjected to full temperature/pressure
875 cooking conditions for 3 minutes excluding ramp time. Pressure was released and samples
876 were allowed to cool to on a bench-top for 30 minutes prior to transfer into PBS.
877

878 *Pre-bleaching*

879 To decrease autofluorescence, slides were pre-bleached as previously described[56].
880 Place slides tissue side up in glass dish filled with 1L pre-bleach buffer (PBS+0.05% sodium
881 azide. Glass dish was placed on reflective aluminum surface inside of a 4C walk-in fridge. A
882 cardboard box lined with polyvinyl shrink-wrap (to protect cardboard from evaporating water),
883 and further lined with aluminum foil to reflect light back onto sample was created and placed
884 overtop of dish containing samples. Holes were cut out of cardboard container to allow for
885 insertion of autofluorescence photo-bleaching lamps. Samples were then exposed to high
886 intensity full spectrum LED lights (3x Relassy: Sunlike Full Spectrum Grow Light 45 Watt light
887 incandescent equivalents and 1x SANSI full spectrum 400 Watt incandescent equivalents) for
888 48 hours. Note that some residual autofluorescence existed, highlighting the importance of the
889 no primary negative controls that we employed. Sodium azide containing PBS was then
890 disposed of in hazardous waste. Samples were then washed 3x in PBS.

891

892 *Protein staining*

893 Samples were blocked for 2hrs at room temperature in blocking buffer (PBS with 1mM
894 $CaCl_2$, 1mM $MgCl_2$, 20% Donkey serum (Jackson Immuno Research: 017-000-121), 0.3%
895 Triton X-100). Samples were stained in staining buffer (PBS with 1mM $CaCl_2$, 1mM $MgCl_2$, 1%
896 Donkey serum, 0.3% Triton X-100). Antibody dilutions were as follows: goat anti-PDGFRβ 1:100
897 (Novus Biologicals: AF385), rabbit anti-CD79a 1:100 (Fisher Scientific PIMA516344; clone
898 SP18), mouse anti-CD19 1:50 (GeneTex: GTX42325, clone LE-CD19).

899 Note that CD19 primary (1:50) and secondary (1:50) were used at higher concentrations
900 than other channels to address any concerns over lack of sensitivity. Our positive control in
901 human tonsil showed abundant signal as expected indicating the efficacy of our staining
902 procedure.

903 Slides were washed 3x with PBS (5 min each), and stained in staining buffer with
904 secondary antibodies, each from Jackson Immuno Research: 1:50 Alexa Fluor 488 AffiniPure
905 Donkey Anti-Mouse IgG (H+L) (715-545-151); 1:100 Rhodamine Red-X (RRX) AffiniPure
906 Donkey Anti-Rabbit IgG (H+L) (711-295-152); 1:100 Alexa Fluor 647 AffiniPure Donkey Anti-
907 Goat IgG (H+L) (705-605-147). Slides were then washed 3x with PBS (5 min each) and
908 mounted in Prolong Diamond Antifade with DAPI (Fisher Scientific: P36962) and sealed with
909 nail polish. Images were all collected using the same laser settings on a Leica SP8 confocal
910 microscope.
911
912 *Image processing*
913 Images were processed in ImageJ, using the median fluorescence value for each
914 channel in the secondaries only negative control, as the minimum inclusion value, to provide an
915 objective negative control for thresholding. This was performed for all donors and all channels.
916 Donor HuFPT018, showed unquenchably high global background in all cells however, (equally
917 between the test and no CD19 primary antibody conditions), and was therefore excluded.
918
919 **Software versions**
920 All R analyses were performed in R version 3.6.0 (2019-04-26) -- "Planting of a Tree". R
921 packages require for analysis were: splatter: 1.8.0, scater: 1.12.2, Seurat: 3.1.2, DESeq2:
922 1.24.0, sva: 3.33.2, irlba: 2.3.3, liger: 0.4.2, harmony: 1.0, BiocParallel: 1.18.1, ggplot2: 3.3.0,
923 reshape2: 1.4.3, pals: 1.6, tidyverse: 1.3.0, stringr: 1.4.0, scales: 1.1.0, entropy: 1.2.1, NMF:
924 0.22.0. Python was version 3.7.1.
925

Data Availability

All dataset inputs used in this manuscript are distributed in the data folder/zip of the benchmark repository, or are public data that are downloaded by the scripts on the fly during analysis: https://bitbucket.org/scottyler892/sc_norm_bench_v2

Code Availability

All code for this benchmark is freely available at the following repository:

https://bitbucket.org/scottyler892/sc_norm_bench_v2

The UMI downsampling package we created is available by pip installation:

python3 -m pip install bio-pyminer-norm

The R version of the downsampling package is located here:

https://github.com/scottyler89/downsample

and can be installed via the devtools package:

devtools::install_github('scottyler89/downsample')

The UMI downsampling package repository and tutorials are located here:

https://bitbucket.org/scottyler892/pyminer_norm/

The R implementation of PMD and associated functions/statistical tests are available here:

https://github.com/scottyler89/PercentMaxDiff

and can be installed via the devtools package:

devtools::install_github('scottyler89/PercentMaxDiff')

The R implementation of towcab and pathway analyses is available here:

https://github.com/scottyler89/towcab

and can be installed via the devtools package:

devtools::install_github('scottyler89/towcab')

References

1. Bacher, R. et al. Enhancing biological signals and detection rates in single-cell RNA-seq experiments with cDNA library equalization. *Nucleic Acids Research* **50**, e12-e12 (2021).
2. Gehring, J., Hwee Park, J., Chen, S., Thomson, M. & Pachter, L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nature Biotechnology* **38**, 35-38 (2020).
3. Suryawanshi, H. et al. A single-cell survey of the human first-trimester placenta and decidua. *Science Advances* **4**, eaau4788 (2018).
4. Muskovic, W. & Powell, J.E. DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. *Genome Biology* **22**, 329 (2021).
5. Chen, J. et al. PBMC fixation and processing for Chromium single-cell RNA sequencing. *Journal of Translational Medicine* **16**, 198 (2018).
6. Charles, A.H. et al. Human prefrontal cortex gene regulatory dynamics from gestation to adulthood at single-cell resolution. *Cell* **185**, 4428-4447.e4428 (2022).
7. Booeshaghi, A.S., Hallgrímsdóttir, I.B., Gálvez-Merchán, Á. & Pachter, L. Depth normalization for single-cell genomics count data. *bioRxiv*, 2022.2005.2006.490859 (2022).
8. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nature Methods* **20**, 665-672 (2023).
9. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821 (2019).
10. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289-1296 (2019).
11. Welch, J.D. et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e1817 (2019).
12. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053-1058 (2018).
13. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964-965 (2019).
14. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology* **17**, e9620 (2021).
15. Lotfollahi, M., Wolf, F.A. & Theis, F.J. scGen predicts single-cell perturbation responses. *Nature Methods* **16**, 715-721 (2019).
16. Luecken, M.D. & Theis, F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15**, e8746 (2019).
17. Luecken, M. et al. Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*, 2020.2005.2022.111161 (2020).
18. Tran, H.T.N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology* **21**, 12 (2020).
19. McCarthy, D.J., Campbell, K.R., Lun, A.T.L. & Wills, Q.F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics (Oxford, England)* **33**, 1179-1186 (2017).
20. Zhang, Y., Parmigiani, G. & Johnson, W.E. ComBat-Seq: batch effect adjustment for RNA-Seq count data. *bioRxiv*, 2020.2001.2013.904730 (2020).
21. Wu, Y.E., Pan, L., Zuo, Y., Li, X. & Hong, W. Detecting Activated Cell Populations Using Single-Cell RNA-Seq. *Neuron* **96**, 313-329.e316 (2017).
22. Marsh, S.E. et al. Dissection of artifactual and confounding glial signatures by single-cell sequencing of mouse and human brain. *Nature Neuroscience* **25**, 306-316 (2022).
23. Squair, J.W. et al. Confronting false discoveries in single-cell differential expression. *Nature Communications* **12**, 5692 (2021).

1006 24. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package
1007 for gene list functional enrichment analysis and namespace conversion toolset g:Profiler.
1008 *F1000Research* **9** (2020).
1009 25. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased
1010 coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic*
1011 *Acids Res* **47**, D607-d613 (2019).
1012 26. Genomics, x. 1k Brain Cells from an E18 Mouse (v3 Chemistry). (2018).
1013 27. Genomics, x. (2018).
1014 28. Batson, J., Royer, L. & Webber, J. Molecular cross-validation for single-cell RNA-seq.
1015 *BioRxiv*, 786269 (2019).
1016 29. Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A. & Theis, F.J. A test metric for assessing
1017 single-cell RNA-seq batch correction. *Nature methods* **16**, 43-49 (2019).
1018 30. Genomics, x., Vol. 2020 (https://support.10xgenomics.com/single-cell-gene-
1019 expression/datasets/3.0.0/neuron_10k_v3?; 2019).
1020 31. Genomics, x., Vol. 2020 (https://support.10xgenomics.com/single-cell-gene-
1021 expression/datasets/3.0.0/neuron_1k_v3; 2019).
1022 32. Ayyaz, A. et al. Single-cell transcriptomes of the regenerating intestine reveal a revival
1023 stem cell. *Nature* **569**, 121-125 (2019).
1024 33. Agarwal, D. et al. A single-cell atlas of the human substantia nigra reveals cell-specific
1025 pathways associated with neurological disorders. *Nature Communications* **11**, 4183
1026 (2020).
1027 34. Almanzar, N. et al. A single-cell transcriptomic atlas characterizes ageing tissues in the
1028 mouse. *Nature* **583**, 590-595 (2020).
1029 35. Demir, R., Yaba, A. & Huppertz, B. Vasculogenesis and angiogenesis in the endometrium
1030 during menstrual cycle and implantation. *Acta Histochemica* **112**, 203-214 (2010).
1031 36. Zhou, J.Z., Way, S.S. & Chen, K. Immunology of the Uterine and Vaginal Mucosae. *Trends*
1032 *in Immunology* **39**, 302-314 (2018).
1033 37. Stavreus-Evers, A. et al. Glycodelin is present in pinopodes of receptive-phase human
1034 endometrium and is associated with down-regulation of progesterone receptor B. *Fertility*
1035 *and Sterility* **85**, 1803-1811 (2006).
1036 38. Fink, E.E. et al. Ureter single-cell and spatial mapping reveal cell types, architecture, and
1037 signaling networks. *bioRxiv*, 2021.2012.2022.473889 (2021).
1038 39. Jones, R.C. et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas
1039 of humans. *Science (New York, N.Y.)* **376**, eabl4896 (2022).
1040 40. Wolock, S.L., Lopez, R. & Klein, A.M. Scrublet: Computational Identification of Cell
1041 Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8**, 281-291.e289 (2019).
1042 41. Amrhein, L., Harsha, K. & Fuchs, C. A mechanistic model for the negative binomial
1043 distribution of single-cell mRNA counts. *bioRxiv*, 657619 (2019).
1044 42. Parker, K.R. et al. Single-Cell Analyses Identify Brain Mural Cells Expressing CD19 as
1045 Potential Off-Tumor Targets for CAR-T Immunotherapies. *Cell* **183**, 126-142.e117 (2020).
1046 43. Garcia, F.J. et al. Single-cell dissection of the human brain vasculature. *Nature* **603**, 893-
1047 899 (2022).
1048 44. McCarthy, D.J., Campbell, K.R., Lun, A.T.L. & Wills, Q.F. Scater: pre-processing, quality
1049 control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*
1050 **33**, 1179-1186 (2017).
1051 45. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome*
1052 *Biol* **11**, R106 (2010).
1053 46. Aguilera-Castrejon, A. et al. Ex utero mouse embryogenesis from pre-gastrulation to late
1054 organogenesis. *Nature* **593**, 119-124 (2021).
1055 47. Genomics, x., Vol. 2020 (https://support.10xgenomics.com/single-cell-
1056 vdj/datasets/3.0.0/vdj_v1_mm_c57bl6_pbmc_5gex; 2019).

48. Hennig, C., Vol. 2020 (CRAN, https://cran.r-project.org/web/packages/fpc/fpc.pdf; 2020).
49. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, complex systems* **1695**, 1-9 (2006).
50. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology* **18**, 174 (2017).
51. Tyler, S.R. et al. PyMINEr Finds Gene and Autocrine-Paracrine Networks from Human Islet scRNA-Seq. *Cell Reports* **26**, 1951-1964.e1958 (2019).
52. Speir, M.L. et al. UCSC Cell Browser: visualize your single-cell data. *Bioinformatics* **37**, 4578-4580 (2021).
53. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
54. Bais, A.S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150-1158 (2019).
55. Leisch, F. Flexmix: A general framework for finite mixture models and latent glass regression in R. (2004).
56. Sun, Y., Ip, P. & Chakrabartty, A. Simple Elimination of Background Fluorescence in Formalin-Fixed Human Brain Tissue for Immunofluorescence Microscopy. *Journal of visualized experiments : JoVE* (2017).
57. Georgiou, P. et al. Experimenter sex modulates mouse biobehavioural and pharmacological responses. *bioRxiv*, 2022.2001.2009.475572 (2022).
58. Perelis, M. et al. Pancreatic β cell enhancers regulate rhythmic transcription of genes controlling insulin secretion. *Science (New York, N.Y.)* **350**, aac4250 (2015).
59. Chari, T., Banerjee, J. & Pachter, L. The Specious Art of Single-Cell Genomics. *bioRxiv*, 2021.2008.2025.457696 (2021).
60. Luecken, M.D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods* **19**, 41-50 (2022).
61. Chen, R., Kang, R. & Tang, D. The mechanism of HMGB1 secretion and release. *Experimental & Molecular Medicine* **54**, 91-102 (2022).
62. Rosenberg, A.B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (New York, N.Y.)* **360**, 176-182 (2018).
63. Kim, J., He, Y. & Park, H. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization* **58**, 285-319 (2014).
64. Paul Hoffman, S.L., Vol. 2022 IntegrateData Function documentation in Seurat (2022).
65. McGinnis, C.S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* **16**, 619-626 (2019).
66. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology* **19**, 224 (2018).
67. Cramir, H. Mathematical methods of statistics. *Princeton U. Press, Princeton* **500** (1946).
68. Bergsma, W. A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society* **42**, 323-328 (2013).
69. Simpson, E.H. Measurement of Diversity. *Nature* **163**, 688-688 (1949).
70. Shannon, C.E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379-423 (1948).
71. Mangiafico, S. Package 'rcompanion'.
72. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* **38**, 147-150 (2020).
73. Zhong, S. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524-528 (2018).

1107

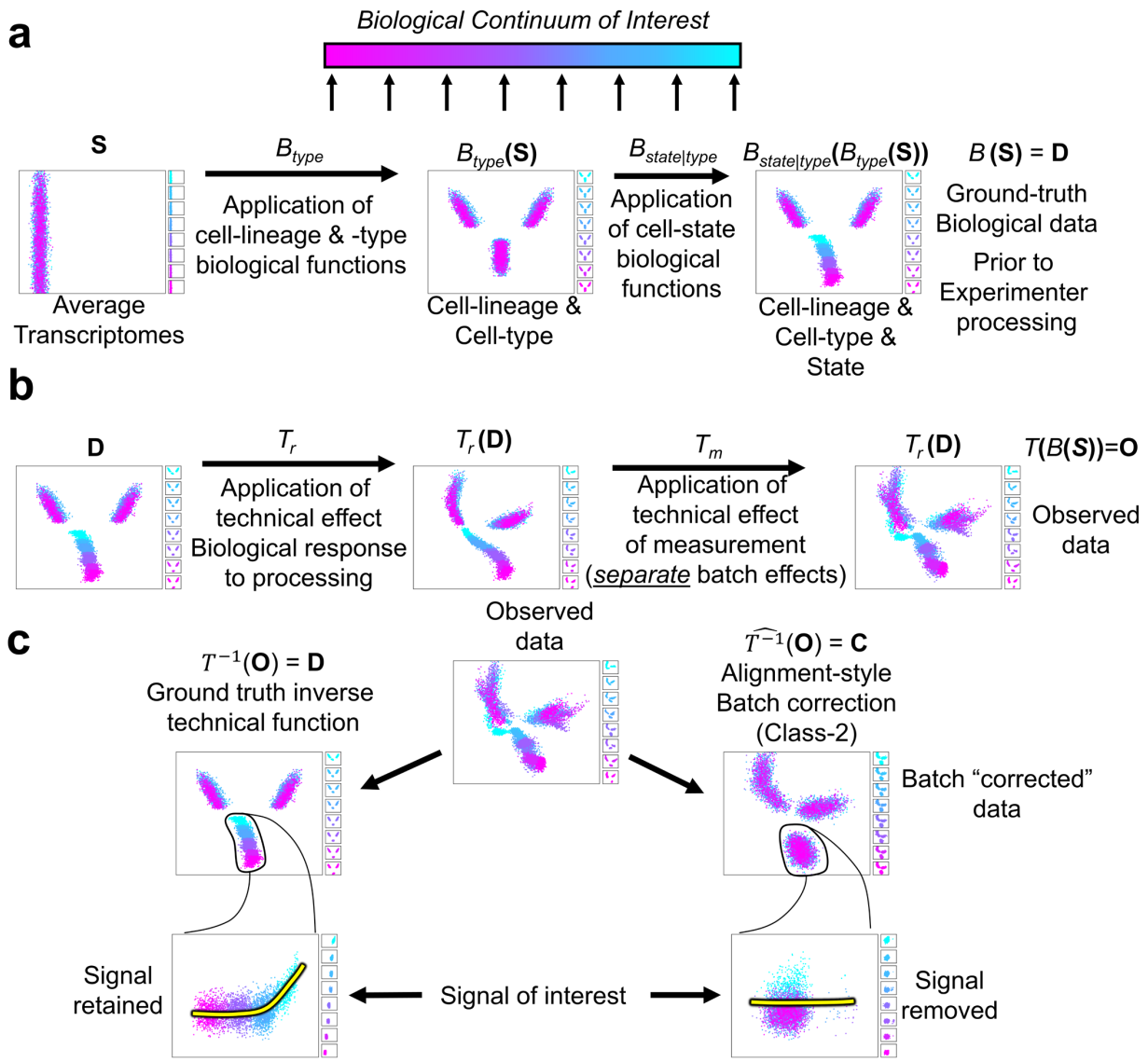


Figure 1: Functional framework for aiding experimental design in single-cell assays. A 2-dimensional representation of cells to illustrate the functional framework we devised that describe the transformations that cells (and their -omic measures) take. **a**, Biological (B) transformations are encoded in sets that are applied to the cells, beginning with the biological functions applied to an average transcriptome (**S**) that encode transformations related to cell-lineage/-type (B_{type}). Next, in an experimental design or disease process, various changes in cell-state can exist encoded by functions contained in (B_{state}). **b**, Once an experimenter begins to intervein to perform the experiment, technical effects occur, as cells react to processing through biologically derived processes (T_r), and bias/errors batch effects are introduced when samples are processed separately (T_m) (for example with different master-mixes, causing differing amplification efficiency, etc). This generates the observed data matrix (**O**). **c**, Next, batch correction algorithms are applied, in an attempt to find the ground truth inverse function of technical effect functions (T^{-1}). However, only a hypothesized approximation of this function can be made ($\widehat{T^{-1}}$); when $\widehat{T^{-1}}$ does not model whether the applied inverse functions are from the

biological set (B), or technical set (T), we cannot be certain the $\widehat{T^{-1}}$ does not contain inverse functions from the biological set.

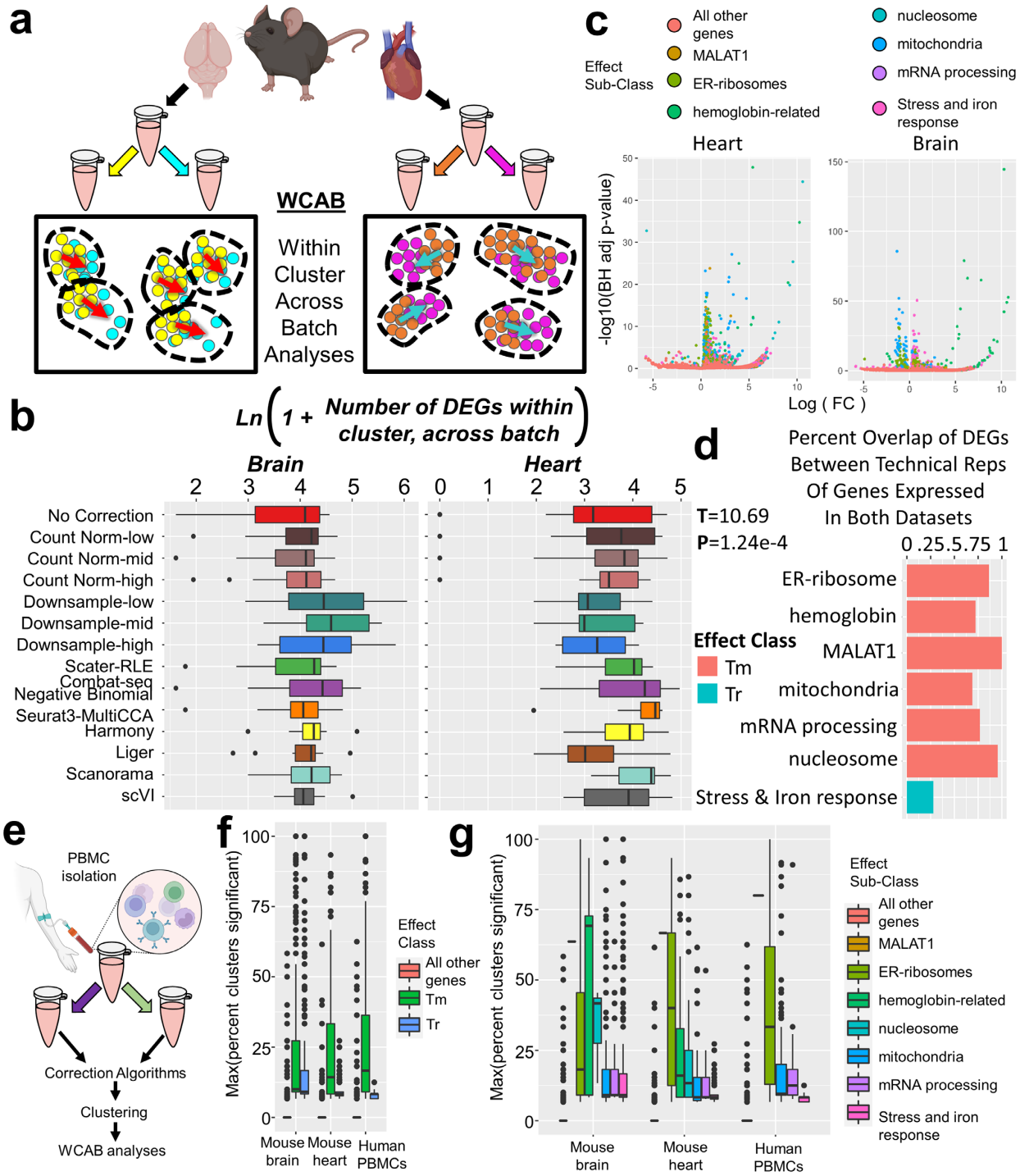


Figure 2: Locally differentially expressed genes across technical-replicates. a, Using brain and heart technical-replicates, batch-corrections were performed followed by differential expression within-clusters/across-batches (WCAB). **b**, Broad differential expression within clusters was seen regardless of correction method is shown by boxplots quantifying the log

1131 number of differentially expressed genes within clusters across batches. **c**, Differentially
1132 expressed genes were largely those localized to large structural components of the cell
1133 including mitochondria, ER, and nucleus. Another class of genes fell into the cell stress
1134 response. **d**, The conserved differentially expressed complexes are identified as the technical
1135 effects of measurement (T_m), while the technical response (T_r) to measurement is less
1136 conserved across datasets. **e**, Human PBMC technical-replicates were assayed by WCAB
1137 analyses. **f**, Boxplots showing that T_m effects uniformly impacted a higher percentage of clusters
1138 than T_r, which was more cluster specific. **g**, Boxplots showing the percentage of clusters in
1139 which a gene was differentially expressed in the noted dataset; displayed is the maximum of
1140 correction methods for each gene (See **ED Table 3** for all results).
1141

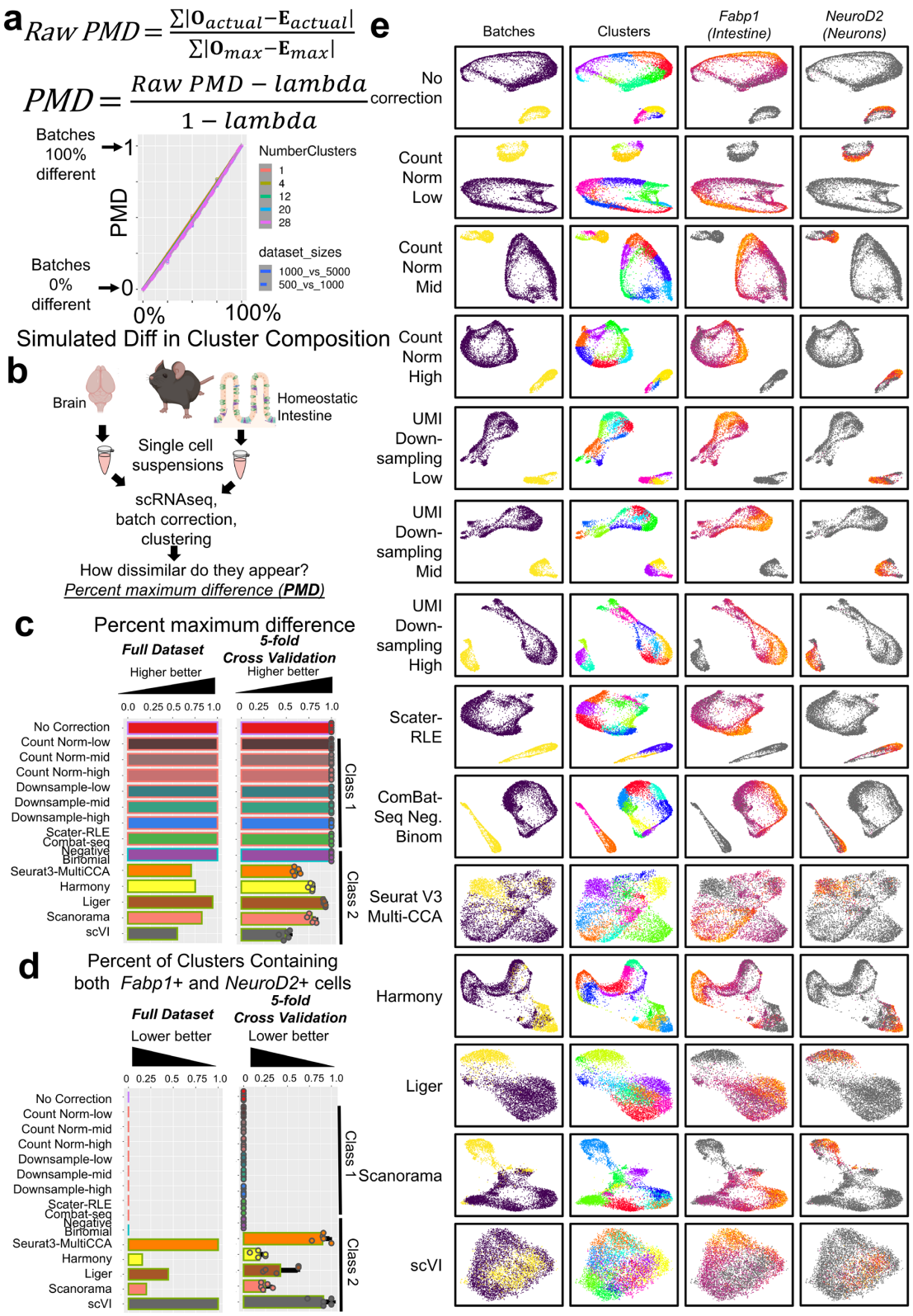


Figure 3: Batch-correction algorithm performance when samples are expected to have no overlap. a, The PMD equation, to compute relative difference between batches based on their

1145 cluster composition, and a scatterplot from simulations (Supplemental Note-4; Supplemental
1146 Figs. 7,8), showing linearity and robustness to differences in number of clusters or dataset
1147 sizes. **b**, A schematic representation of the dataset integration task in which there is little to no
1148 overlap in cell-types expected (brain and intestine samples) quantified by the percent maximum
1149 difference (PMD). **c**, A bar chart displaying the PMD between the input brain and intestine
1150 datasets. Five-fold cross-validation of this integration task show highly concordant results. **d**, A
1151 bar chart showing the percentage of clusters containing both *FABP1*+ and *NEUROD2*+ cells
1152 (markers of intestinal epithelia and neurons respectively) at ≥5 UMI in original datasets; all bar
1153 charts show means+/-SD. **e**, Dataset mixing visualized by a spring-embedding plots of the kNNs
1154 used for clustering, colorized by batch, cluster, and marker-gene positivity.

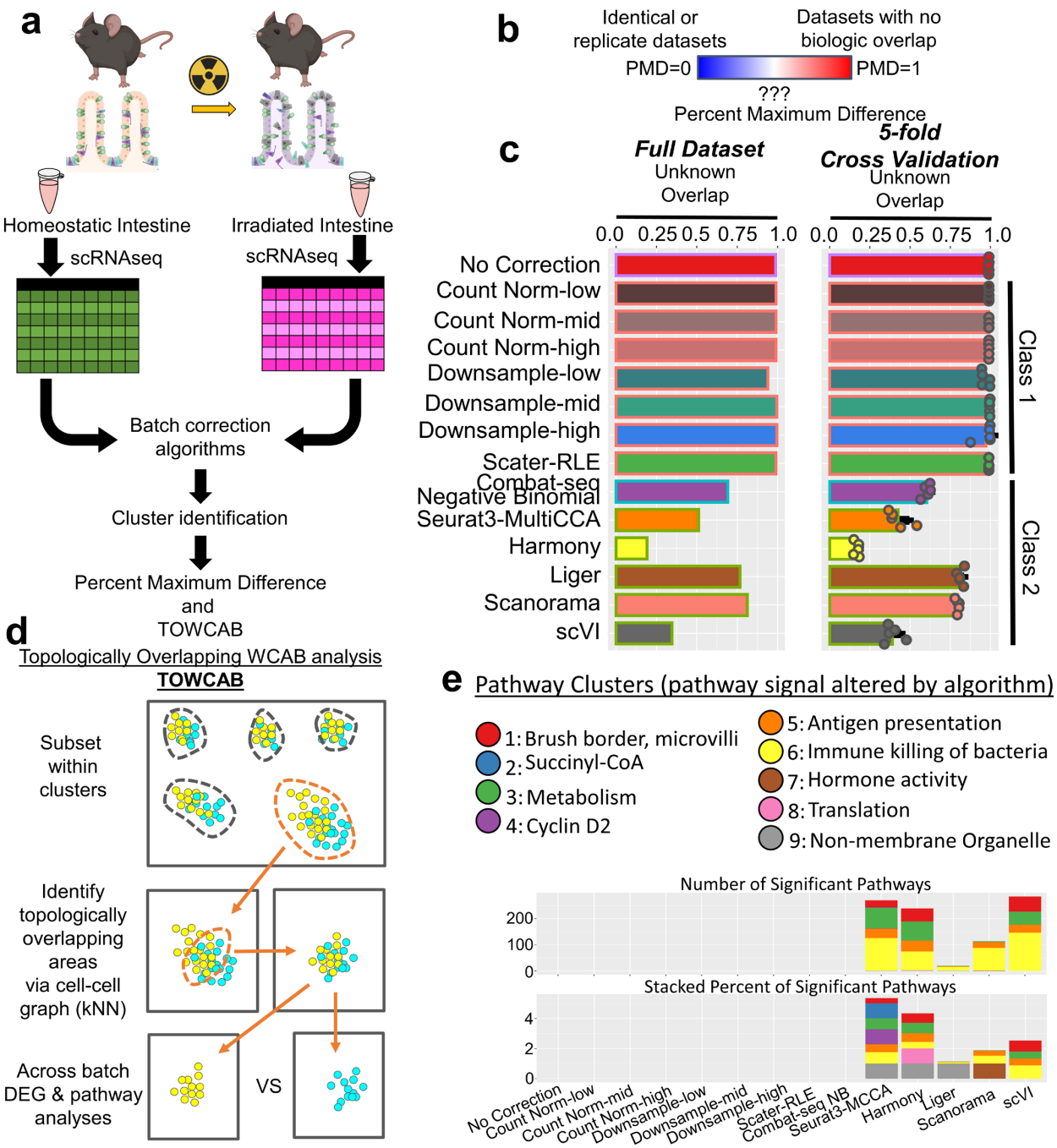


Figure 4: Class-2 algorithms with LDM can create biologically heterogenous topologies. **a**, Schematic of analysis approach: Homeostatic and irradiated mouse intestines were analyzed, and assessed both by PMD, and a refinement of WCAB, that selects only topologically overlapping areas within clusters (TOWCAB)[32]. **b**, There was no clear *a priori* expectation how similar these datasets would be by PMD. **c**, A wide range of PMD values was observed ranging

1161 from nearly completely different without batch-correction and with depth normalization

1162 approaches to mostly the same (20.2% maximum difference) with Harmony. Five-fold cross-

1163 validation of this integration task show highly concordant results (means +/- SD). **d**,

1164 Quantification of the number of, or stacked percentage of, significant pathways whose signal are

1165 altered along topologically overlapping regions.

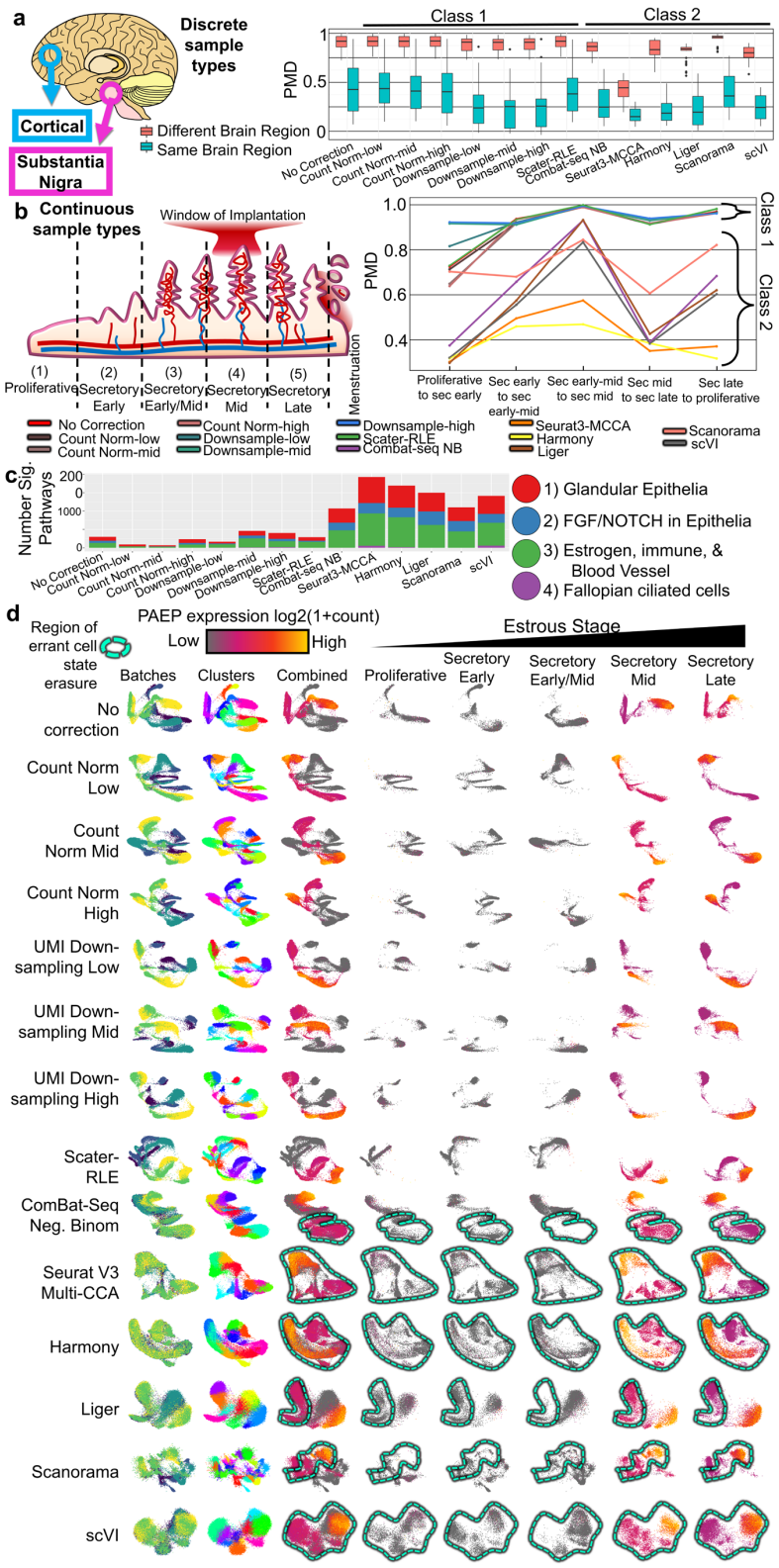


Figure 5: Batch-correction of discrete or continuous sample types. a, Boxplots showing measured PMD values for within and between tissue-types on batch-correction approaches using

1169 on two human brain sample-types (cerebral cortex (n=5) and substantia nigra (n=7)) that share
1170 all cell-types, but appear in differing abundance. **b**, Continuous variation in cell-abudnace and
1171 state within the endometrium over estrous was assayed for batch correction methods, displaying
1172 line-graphs of apparent batch-batch similarity over the estrous cycle (**g**:x-axis=time; y-axis =PMD,
1173 color=correction method). **c**, TOWCAB analyses show that Class-2 batch-correction methods
1174 erase a greater amount of biologically meaningful pathways. **d**, Plotting expression of *PAEP* (a
1175 known cell-state gene, up in mid-secretory) illustrates this cell-state erasure by Class-2 methods
1176 standard methods.
1177

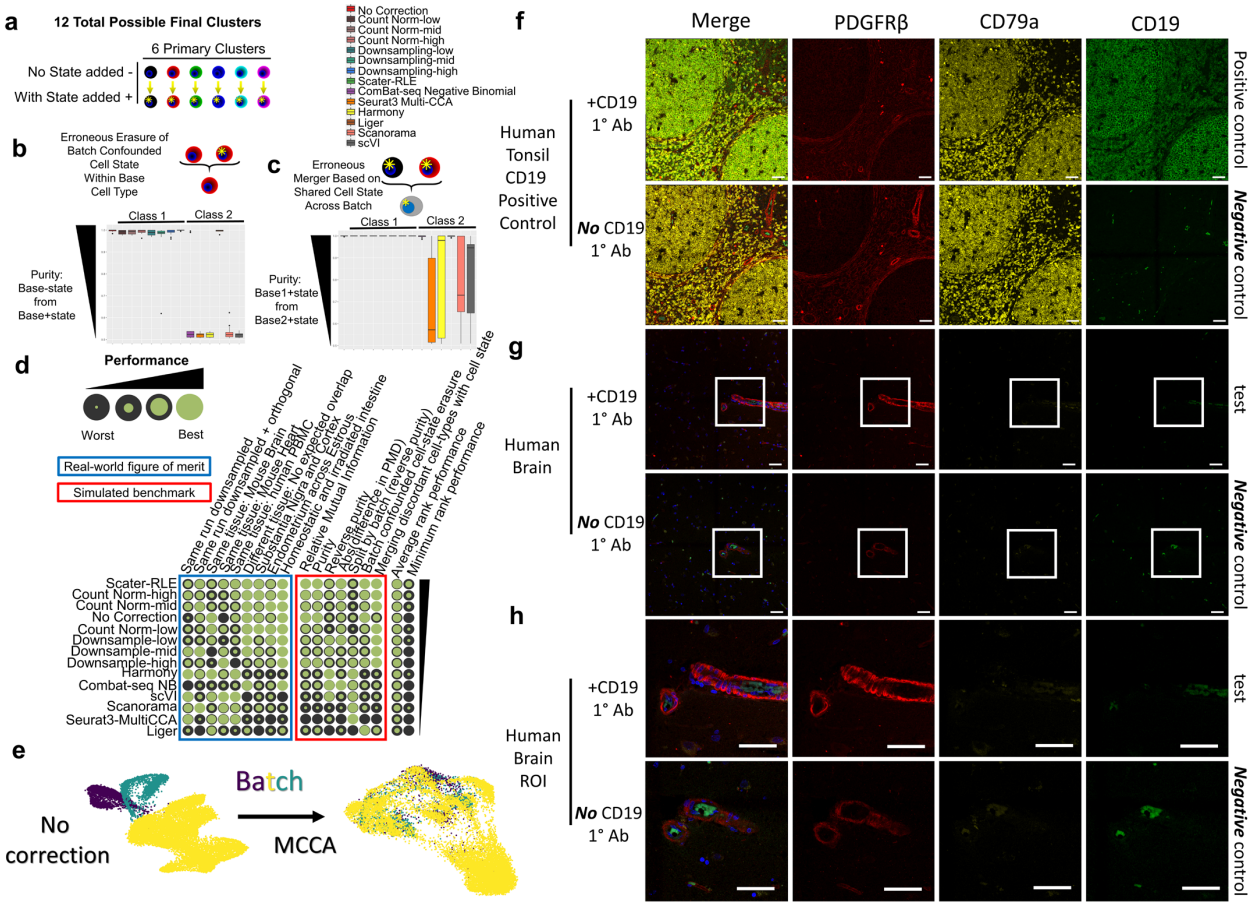
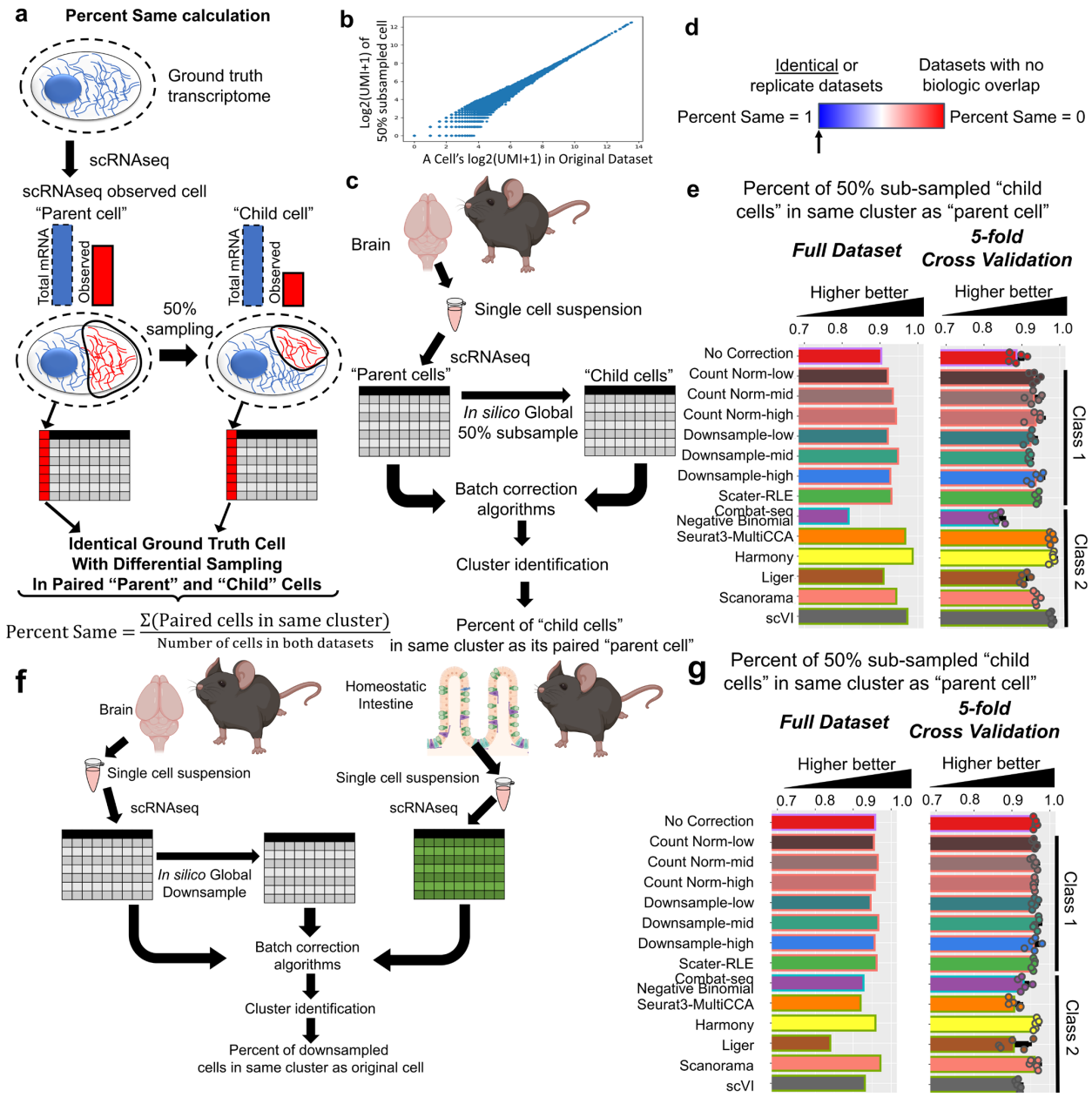


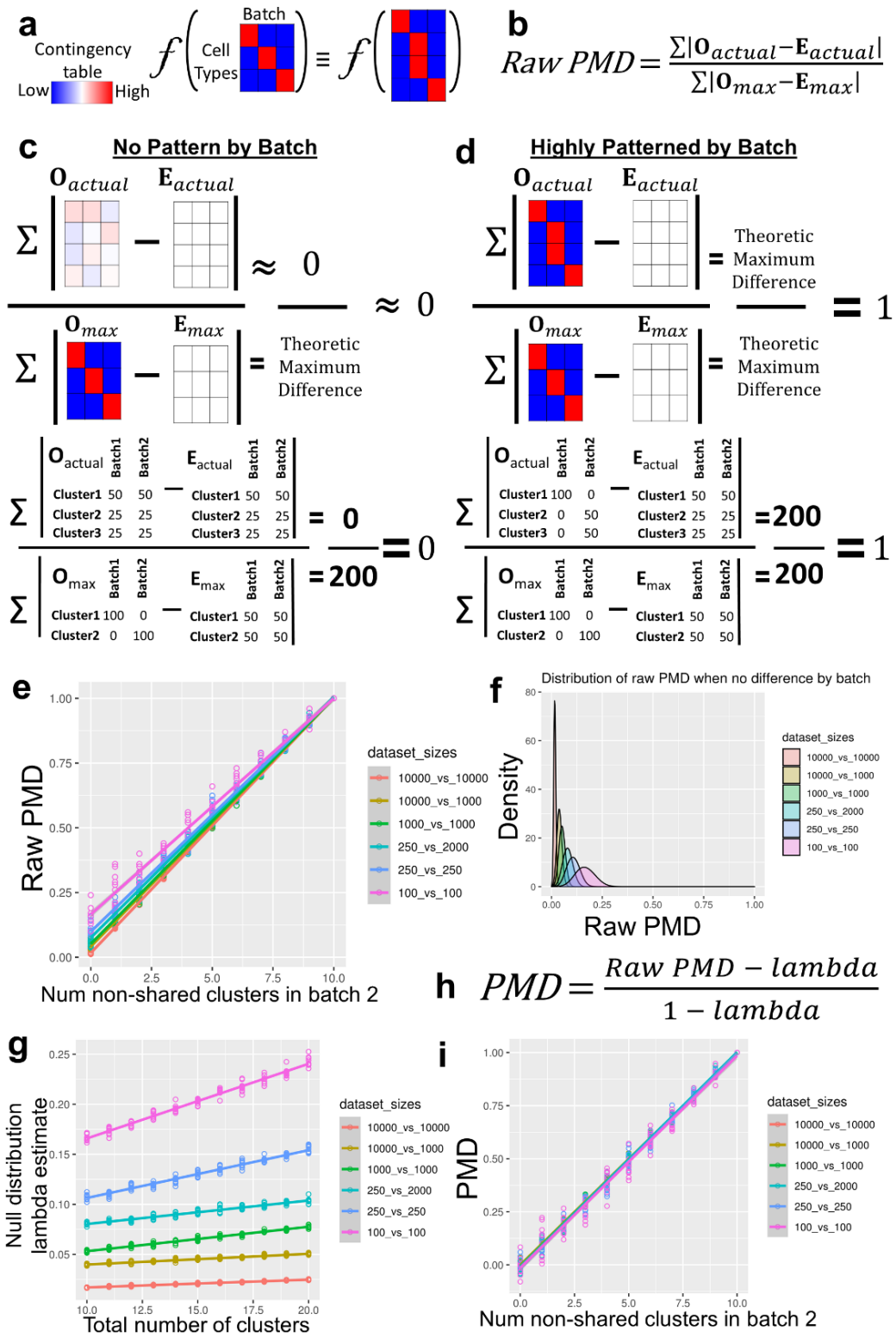
Figure 6: Class-2 algorithms over-merge when transcriptome mixtures are present. a, Twelve possible clusters were used in simulations; they were generated from six primary clusters, with a cell-state optionally added (**Supplementary Note-5**; **Supplementary Fig. 9**). **b,c,** Boxplots and schematics illustrating the possible failure-mode for erasure of "cell-state" or "cell-type" sources of variation when batch confounded. Boxplots are the typical display of median, with the box highlighting the second-to-third quartiles and whiskers extending out to the first and fourth quartiles; outliers are defined as 1.5x the inner quartile range. **d**, Each tested method (row) is sorted by their average rank in real-world (blue box) and synthetic (red box) assessments. **e**, After batch-correction of the three human brain datasets, MCCA (as was performed in the original manuscript) shows thorough mixing of batches. **f**, We performed CD19 positive control staining on human tonsil B-cells along with PDGFRβ and CD79a including a no CD19 primary antibody negative control. **g-h**, With the same exact staining and imaging

1191 conditions as our positive control, we stained 4 donor brains including the no CD19 primary
1192 negative control. The small signal in the CD19 channel was luminal rather than perivascular,
1193 was also faintly positive in the CD79a channel and present in the no-CD19 primary antibody
1194 condition, consistent with background autofluorescence. We did not observe CD19/PDGFRβ co-
1195 positive perivascular cells. Scale bars indicate 50um.
1196



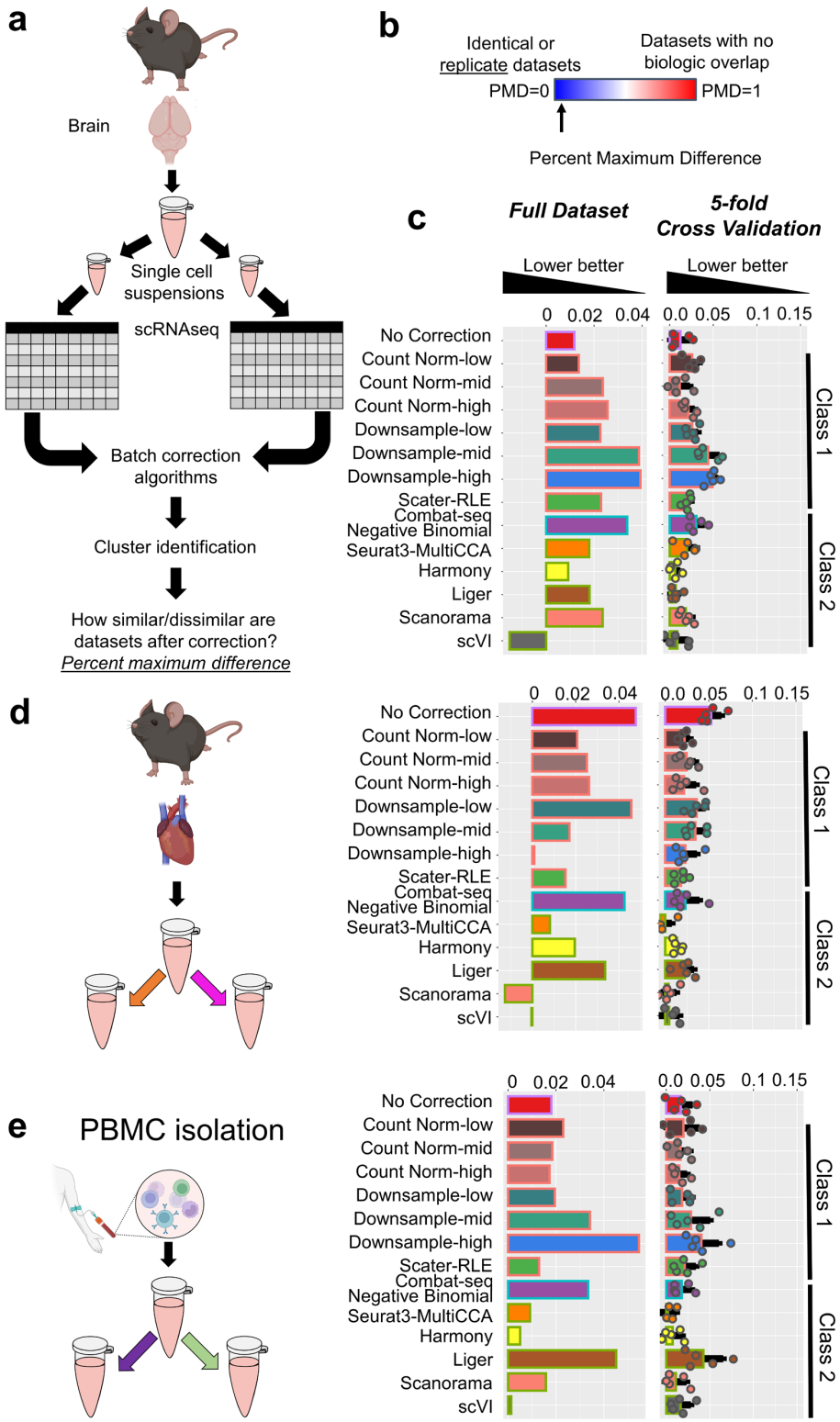
Extended Data Figure 1: Performance in integrating identical datasets with differences only in UMI depth. a, Example schematic of the same cell with different transcript capture efficiency leading to different observed total UMI. **b**, Scatter plot showing a brain scRNAseq dataset that has been globally downsampled at random. Each point is a transcript's log2(UMI+1) measure in the parental cell (X-axis) and downsampled cell (Y-axis). **c**, Pipeline overview for benchmark schematic. **d**, The percent-same should equal 1, because the datasets are from the same source droplets. **e**, A bar chart displaying the percentage of downsampled

1205 cells that were classified as belonging to the same cluster as its parental cell. A higher score
1206 indicates better performance. **f**, The overall experimental and pipeline schematic of this figure of
1207 merit for results in **g**. **g**, Bar charts of the percent-same metric in the full datasets or cross-
1208 validation splits (means +/- SD).



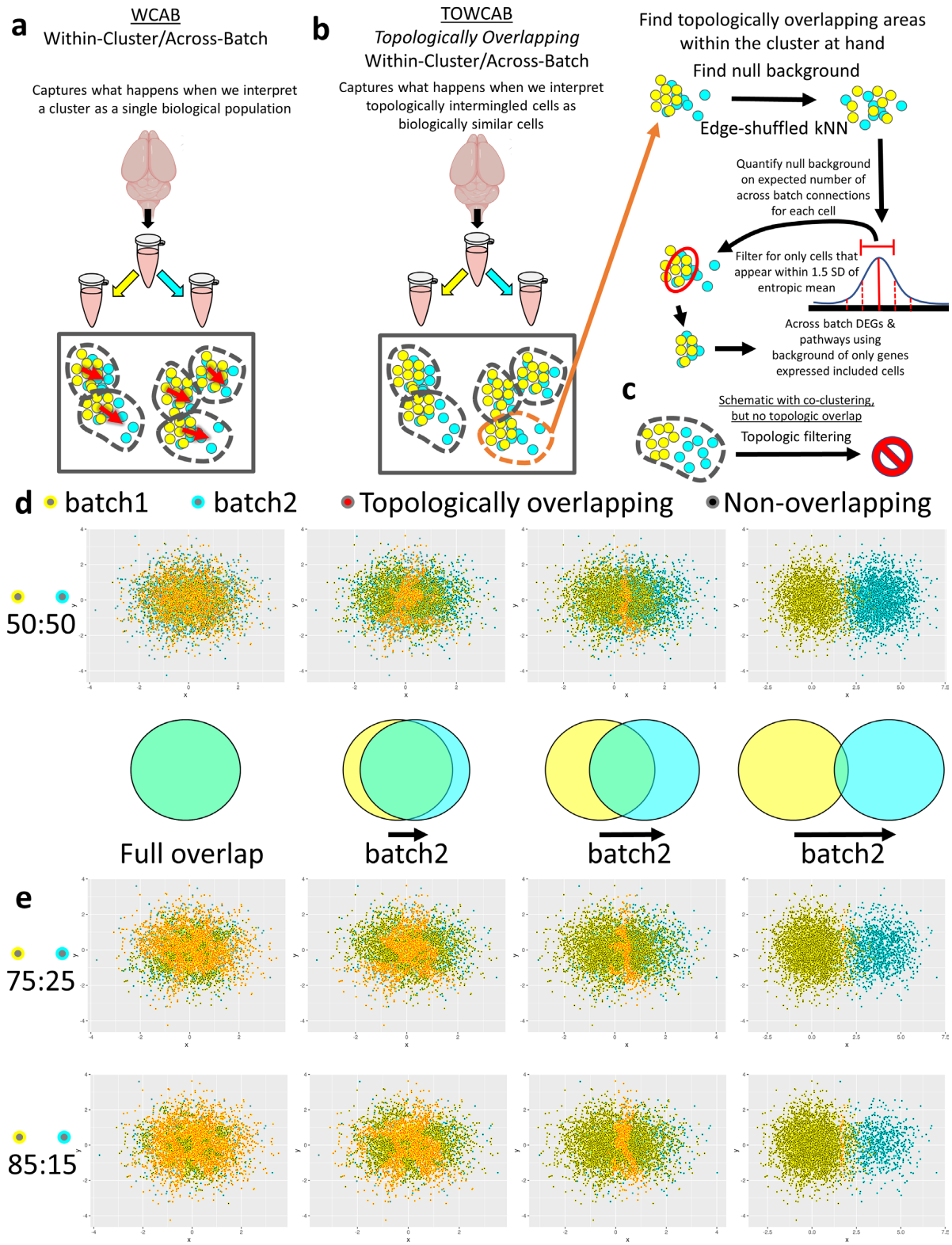
Extended Data Figure 2: Percent Maximum Difference calculation. a, A schematic showing that an ideal metric quantifying differences in batches should be invariant to the number of

clusters found within each batch if relative overlap across batches is unchanged. **b**, The equation for the Raw Percent Maximum Difference (Raw PMD) (before empiric power correction). **c**, A visual example of the Raw PMD calculation where there is very little asymmetry across batches. **d**, In contrast to the scenario in **c**, when the actual clustering results are highly asymmetric by batch, the numerator and denominator are equivalent and this ratio equals one. **e**, Raw PMD (y-axis) plotted against the number of clusters that are specific for a given batch (x-axis: higher is more batch asymmetry); simulations with different batch sizes are noted by color. **f**, Histograms of the intercepts of the lines in **e** revealed Poisson distributions, whose lambda parameter (peak of the distribution) differed based on the number of cells simulated (colors). **g**, Lambda parameter estimates (y-axis) of bootstrap shuffled null versions of the simulations shown in **e**, varied with the number of cells (color) and degree of overlap (x-axis), indicating that Raw PMD can be corrected by measuring the lambda parameter. **h**, The final PMD calculation, adjusted for power with the lambda parameter of the null distribution corrects for the intercept (numerator) and slope (denominator). **i**, After adjustment, PMD (y-axis), is linear with the degree of overlap between batches (as noted by number of batch specific clusters, x-axis), and equivalent regardless of the size of these batches (color).



Extended Data Figure 3: Batch-correction algorithm performance on merging technical replicates. a, A schematic representation of the experimental and analytic pipeline for this data

1232 integration example. **b**, Technical replicates share identical cluster composition, meaning that
1233 PMD should equal 0. **c**, A bar chart showing the percent maximum difference between full
1234 datasets and five-fold cross-validation subsets. **d-e**, Technical-replicates of (**d**) mouse heart and
1235 (**e**) human PBMCs were also assessed. Means+/-SD are shown for cross-validation splits.
1236

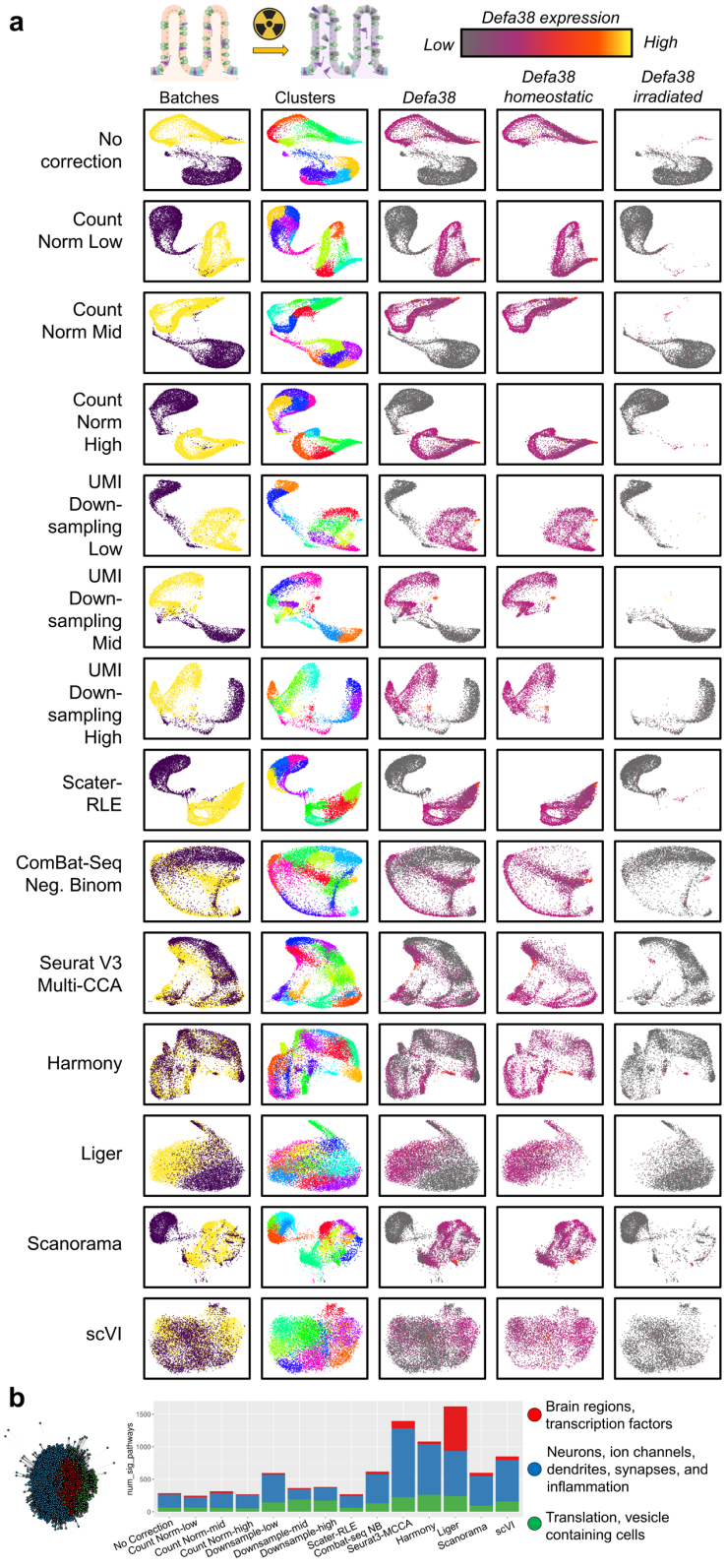


1237

1238 **Extended Data Fig. 4: WCAB and TOWCAB analysis schematic. a**, Within-Cluster/Across-

1239 Batch (WCAB) analyses take in the original count matrix, batch information, and clustering

1240 results. Within clusters, differential expression analyses are performed on a downsampled
1241 version of the datasets, providing the most conservative assessment of relative differential
1242 expression/detection with total counts per cell held constant. **b**, Topologically overlapping
1243 WCAB (TOWCAB) refines this approach, filtering for only cells that strongly match a null
1244 distribution of what entropic mixing of this topology would look like, therefore keeping only cells
1245 that are thoroughly mixed across batches within this subset of the topology. **c**, While WCAB
1246 analyses will include results in which two batches are identified as a single cluster, when in fact
1247 they were simply “zippered” along an edge, TOWCAB analyses will not find any topologically
1248 overlapped areas within the cluster because they are only adjacent rather than overlapping. **d**, A
1249 simulated example dataset shows this filtering process as implemented in our R towcab
1250 package. Displayed points are considered as a single cluster, with varying degrees of overlap
1251 by batch (colorized by border), while the center color indicates whether the point was
1252 considered thoroughly mixed enough to be considered “topologically overlapping” by batch. **e**,
1253 The use of bootstrap shuffled null distributions enables efficacy in discovering topologically
1254 overlapped regions even with class imbalance.
1255

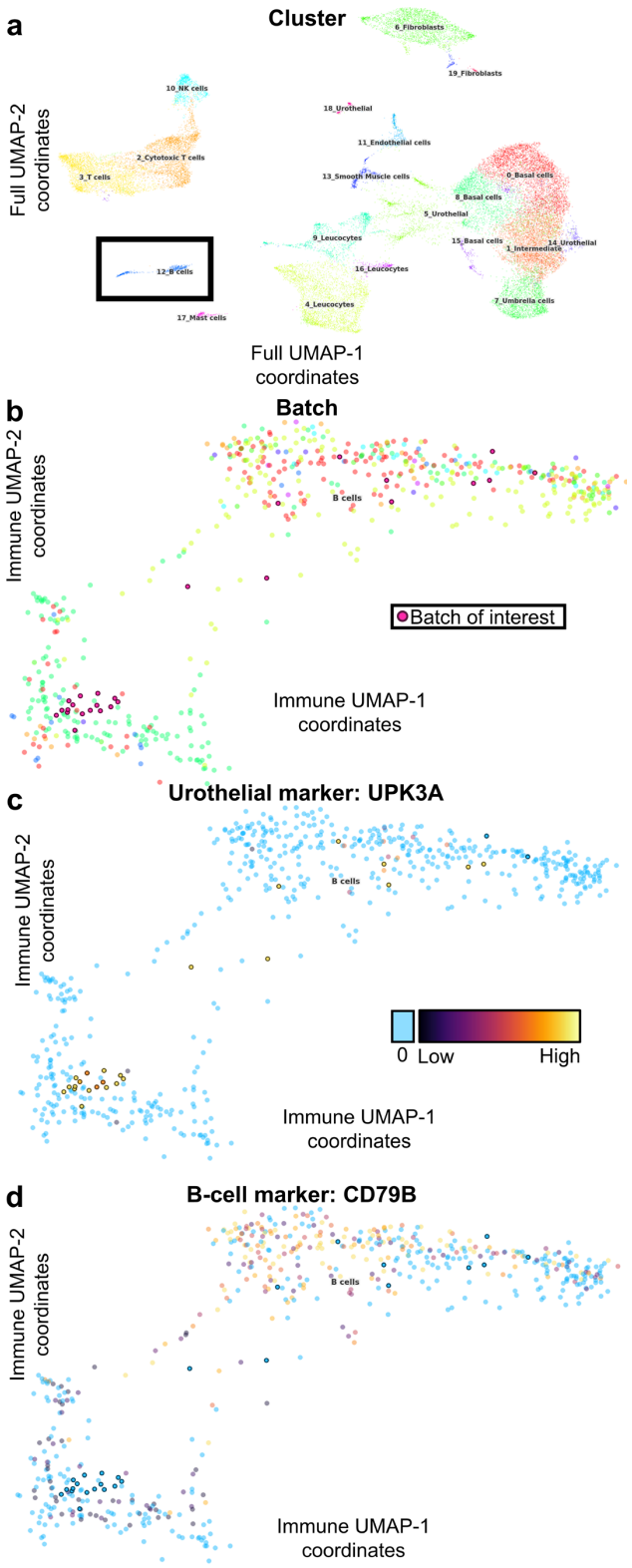


1256

1257 **Extended Data Fig. 5: Erasure of cell-state or tissue relevant transcriptional signal by**

1258 **Class-2 methods in mouse intestine and human brain. a**, Spring embedding plots of the

1259 kNNs used for clustering show that defensin alpha 38 expression is lower in irradiated intestine.
1260 Despite this, many Class-2 algorithms merge these topologies, fitting with the TOWCAB
1261 analyses shown in **Fig. 4**. **b**, TOWCAB analysis of brain regions; the graph network shows
1262 nodes representing pathways, connected by edges based on relative overlap in gene-lists,
1263 colorized by Louvain modularity to identify the pathway clusters. A bar chart showing the
1264 number of significant pathways whose signal was significantly differentially expressed within
1265 topologically overlapping regions of the cell-cell graph. All Class-2 algorithms showed a greater
1266 number of differentially detected/expressed pathways compared to Class-1 algorithms. Cluster-
1267 1 and -2 show clear biological importance to the brain.
1268



Extended Data Figure 6: Application of TOWCAB reveals erasure of cell-type signal in an atlas embedding. **a**, Published UMAP coordinates of adult human ureter colorized by cluster; the

1272 cluster labeled as B-cells is boxed. **b**, Cells annotated as B-cells along the published Immune
1273 compartment-specific UMAP coordinates colorized by batch. Note the red batch with darker
1274 circles indicates the batch of interest. **c,d**, TOWCAB analysis identified differential expression of
1275 *UPK3A* and *CD79B* across batches. **c**, Cells from the batch of interest are expressing the
1276 urothelial marker *UPK3A*. **d**, Cells from this batch are not expressing the B-cell marker *CD79B.*
1277

	Comparison Type	**Comparison Description**	**Experimental Setup**	**Figure of Merit**
a	Experimentally identical	Cells of a single tissue compared to subsamples of the transcriptomes of the same cells	Subsample from source "parent" cells to create "child" cells with all cells of the same tissue type	Cluster across all parent and child cells and compute the percentage of times child cells cluster with parent cells (**ED. Fig. 2a-e**)
b	Experimentally identical in context of highly dissimilar	As above (**Box 1a**), but in the context of unrelated cell types	Subsample from source "parent" cells to create child cells where these cells are of the same type, but mixed with unrelated cell types	Same as above **Box 1a** (**ED Fig. 1f-g**)
c	Experimentally differing only by Poisson sampling and technical	Cells from the same single cell suspension	scRNAseq data generated on multiple isolates of the same sample	Percent maximum difference where methods that minimize the PMD are preferred (**ED Fig. 3**)
d	Experimentally highly dissimilar	Cells of one type compared to cells of an unrelated type	Independent runs of scRNAseq data generated on disparate tissues expected to have little or	Percent maximum difference where methods that maximize the PMD are preferred (**Fig. 3**)

			no overlap in cellular composition	
e	Experimentally batch confounded cell-state and cell-type abundance.	Cells of a given batch are of similar cell-type, but in different cell-states, and cell-type abundance	scRNAseq data generated on cells derived from normal intestine and cell derived from irradiated intestine, differences in endometrium throughout menstruation.	Percent maximum difference; TOWCAB analyses (**Fig. 4**, **4d-e**; **ED. Fig. 5a**)
f	Batches assumed to be similar, but may have had biologically meaningful differences	Cells came from the same tissue, and may be assumed to be highly similar, but in reality may harbor biologically meaningful differences.	Within-dataset templated cell-states added, within brain region comparisons (SN vs cortex), and adult ureter atlas, and human brain mural cells.	TOWCAB analyses and protein staining (**Supplemental Fig. 2**; **Fig. 5,6e-h**; **ED Fig. 5b,6**).
g	Mixtures of similar and dissimilar cells	Cells of the same and different types compared for similar and dissimilar groupings.	Simulated batch-confounded and non-confounded cell-type/-states, the multi-tissue *tabula sapiens* atlas.	Various including batch similarity compared to ground-truth, overall clustering accuracy, extent of misclustering by batch (**Fig. 6a-c**; **Supplemental Fig.**

				9,10). Visual inspection of published coordinates (**Supplemental Fig. 1**)

Box 1: A brief overview of the figures of merit based on real biological datasets, and synthetic datasets, indicating whether they are testing how approaches perform on similar or dissimilar datasets and whether or not cell state information is included in the figure of merit or synthetic benchmark.

1283 **Extended Data Table 1: Relative co-mentions of Class-2 algorithms with the term "batch-**
1284 **correction" within the population of manuscripts that contain the phrase "single cell**
1285 **RNAseq". a**, Tabulated results of method/batch correction co-mentions and chi-square
1286 statistics. **b**, Raw data and links that were used for tabulation of co-mentions.
1287
1288 **Extended Data Table 2: Standard benchmarking metrics often rely on assumptions of**
1289 **perfect knowledge by the experimenter. a,** Raw values for each metric assessed when the
1290 added state was "observed" or "unannotated" by the experimenter. **b**, Raw values for each
1291 metric when reference labels were generated using Euclidean distance based on Spearman
1292 correlations vs Euclidean distance based on the first 30 PCs. **c**, The correlations of each metric
1293 when the state was either observed and annotated vs not observed or annotated by the
1294 experimenter. **d**, The correlations between metric values when reference labels were generated
1295 from Euclidean distance of Spearman correlations vs Euclidean distance of the first 30 PCs. **e**,
1296 Summary statistics demonstrating that many metrics yield significantly different results
1297 dependent on whether the experimenter had perfect knowledge and annotation of all cell-states
1298 compared to a situation where cell states remained unannotated in the provided reference
1299 labels. **f**, Summary statistics showing that some popular metrics yield significantly different
1300 results, rewarding for similarity to the processing methods used to generate reference labels. **g**,
1301 Summary of all results comparing correlation consistency or significant differences in means
1302 when states are observed vs non-observed.
1303
1304 **Extended Data Table 3: Study of Technical Effects on Chromium V3 with technical-**
1305 **replicates in different tissues and species**. **a**, A collated table displaying the technical class
1306 each gene belongs to, for each individual dataset. For each gene, the percentage of clusters for
1307 which it was differentially detected within a batch-correction run was quantified. **b-d**, Detailed
1308 lists of all genes and batch-correction methods and the percentage of clusters for which that

1309 gene was significantly different in the WCAB analysis is reported for (**b**) mouse brain, (**c**) mouse
1310 heart, and (**d**) human PBMCs. **e**, A collated table of the intersect of technical effects between
1311 moues brain and mouse heart can provide a draft of genes known to be impacted by technical
1312 effects. **f**, The union of pathways that were significantly different by WCAB analyses in mouse
1313 brain and heart can act as a conservative "background" of pathways that we observed as
1314 differing between technical-replicates in either of these tissues.
1315
1316 **Extended Data Table 4:** All results for simulation scenarios, and metrics used for quantifying
1317 batch-correction algorithm performance are provided.
1318
1319 **Extended Data Table 5: Pathway results of genes correlated with markers of interest in**
1320 **Zong 2018 dataset. a,** Pathways of genes significantly correlated with CD19 above a bootstrap
1321 shuffled null background. **b**, Pathways of genes significantly correlated with CD79a/b.
1322

1323 **Supplemental Notes**

1324 ***Supplemental Note-1: Status quo methods of benchmarking batch correction***

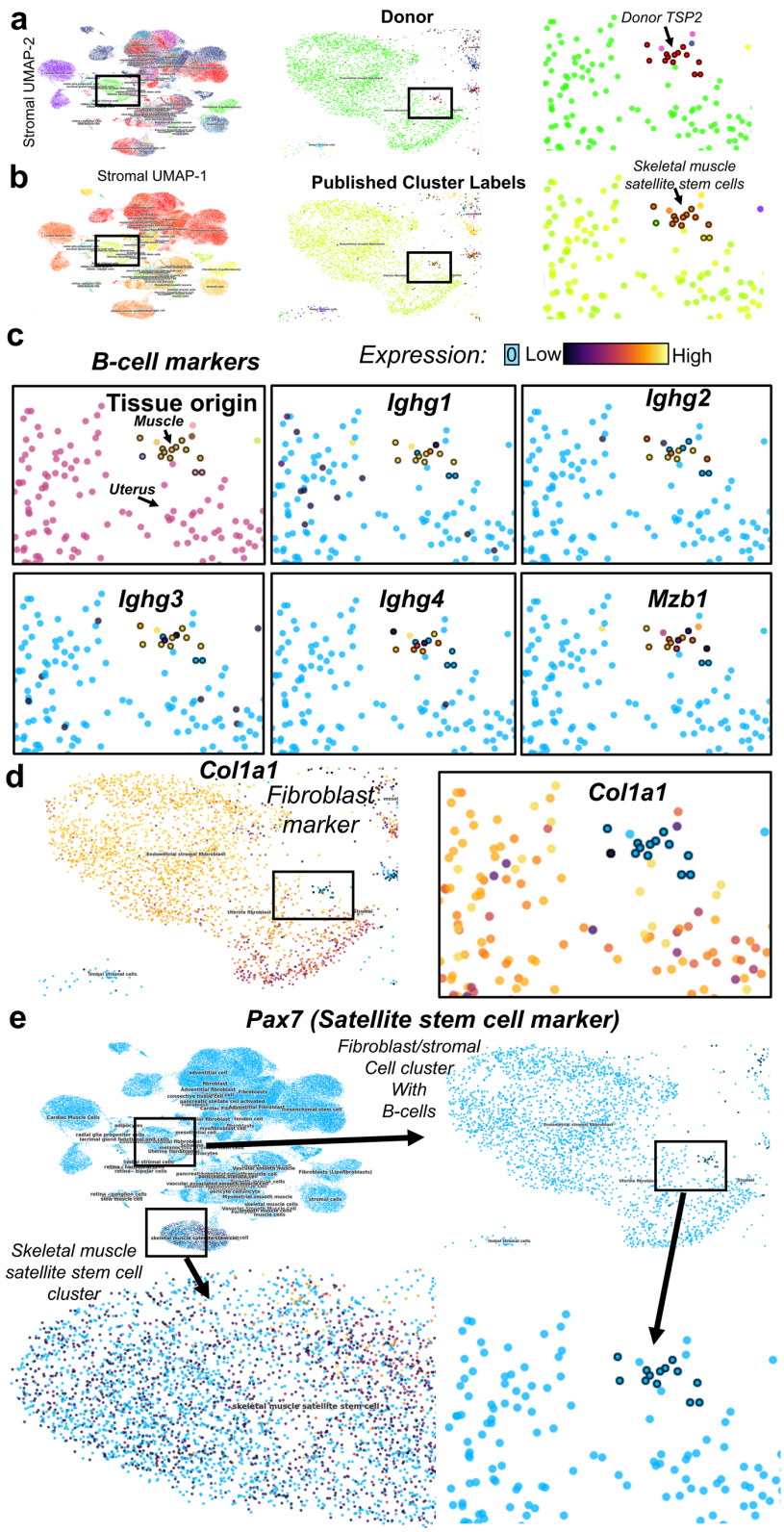
1325 *Assumptions made by status quo metrics*

1326 Most, if not all standard metrics used in benchmarking scRNAseq batch correction
1327 methods rely on one of two assumptions: either 1) they rely on the assumption that batches
1328 differ only in technical effects, and that there are no batch confounded differences in biological
1329 variation (ex: kBET and iLISI); or 2) manually provided cluster labels accurately capture all
1330 meaningful sources of biological variation, and clusters with the same label reflect biologically
1331 homogenous populations [within cluster average silhouette width (ASW), adjusted rand index
1332 (ARI), etc]. Broadly speaking, the later class tests for consistency within the manually provided
1333 labels. However, in scRNAseq benchmarks, these manually provided labels come from
1334 published annotations, which are arrived at after the application of the original manuscripts
1335 preferred analytic pipelines, doing their best to label what they think are pertinent and expected
1336 clusters.

1337 The first class of assumptions will hold as correct when there are no biological
1338 differences between batches. However, technical replicates are the only scenario under which
1339 such an assumption can be taken for granted. Even, for example, within isogenic mouse lines,
1340 subtle differences in environment, somatic variants, and epigenetic changes can all dramatically
1341 alter molecular biology pathways. One illustration of this is a recent discovery that the sex of the
1342 experimenter alters the metabolic pathway of ketamine in isogenic mice. This was ultimately
1343 traced back to the scent of human males compared the scent of human females[57]. Additionally,
1344 there are many other sources of within-subject environmental variation impacting cell states,
1345 such as time since feeding, circadian rhythm patterns, and so on – none of which can be
1346 completely controlled. True technical replicates constructed by taking two separate pipette
1347 draws from a single tube provides the only near guarantee that samples only differ in
1348 downstream technical effects and random Poisson sampling of the original biological sample.

1349 Environmental factors can confound our interpretation when benchmarking batch correction
1350 methods even when using isogenic mice; with any biological replicate, biological variation within
1351 group as well as technical variation are included.

1352 There are also several dangers in relying on previously published cluster labels. The first
1353 sub-assumption within these metrics is that the labels capture all meaningful sources of
1354 variation. For example, it has been reported that "27% of the β cell transcriptome exhibited
1355 circadian oscillation[58]." This means that if a reference label is simply "β cell," this biologically
1356 meaningful source of variation will not be captured within this label. Furthermore, there may be
1357 instances when the application of an experimenter's prior expectations of the dataset can
1358 negatively impact their ability to produce accurate labels. For example, we have found in the
1359 *tabula sapiens* multi-tissue atlas, there exists a population of droplets that came from a muscle
1360 sample, that was labeled as satellite stem cells[39]. However, these droplets were negative for the
1361 markers of this cell type, but were instead positive for B-cell marker genes, yet were inserted
1362 into a topologic region annotated as fibroblast/stromal cells following Class-2+LDM integration
1363 (scVI) (**Supplemental Fig. 1**)[39]. This highlights the dangers of 1) relying on the use of prior
1364 expectations for processing and droplet annotation (since the example just given, the B-cells
1365 were labeled as muscle satellite cells and the sample origin was muscle), and 2) this Class-
1366 2+LDM integrated dataset resulted in biologically heterogeneous populations of cells along what
1367 appears to be connected topologies.
1368



Supplemental Fig. 1: Example errant integration in the *tabula sapiens* atlas with mislabeling due to application of Class-2+LDM integration and tissue specific prior

1372 **expectations. a**, UMAP coordinate showing the *tabula sapiens* stromal compartment, colorized
1373 by donor. We focus on donor TSP2. **b**, UMAP colorized by the custom annotations provided by
1374 tissue experts; the highlighted subset is the region of interest with the donor of interest
1375 highlighted; these droplets were labeled as muscle satellite stem cells, while other cells within
1376 this topology were largely labeled as endometrial or uterine fibroblasts. **c**, The tissue of origin
1377 shows the droplets of interest were from a muscle sample, likely explaining their label of muscle
1378 satellite cells. Colorized expression for B-cell marker genes shows that these droplets are
1379 consistent with B-cells, yet are topologically integrated with endometrial/uterine fibroblasts. **d**,
1380 The droplets of interest were also negative for the fibroblast marker *Col1a1*. **e**, Droplets of
1381 interest were also negative for the satellite stem cell marker *Pax7*, despite many of the cells
1382 within the larger satellite cell cluster being positive for this marker.
1383

1384 In addition to the risks of errant labels, when relying on labels from prior publications, we
1385 are also at risk of label incompleteness. We hypothesized that if an *unknown* difference in cell-
1386 state existed across three different batches, commonly employed benchmarking metrics may
1387 reward for erasure of these non-annotated differences. To test this hypothesis, we partitioned a
1388 dataset (10k human PBMCs) randomly into thirds, and introduced high, medium, and low
1389 brackets of transcription randomly in a batch confounded manner, to a subset of genes from
1390 pertinent signal transduction and immune related pathways (**Supplemental Fig. 2a-c**). Note
1391 that these "cell-state spike-ins" were not generated by simulation, but were instead templated by
1392 three brackets of underdispersed genes, thus accounting for any concerns over simulation not
1393 matching biologically derived distributions (**Supplemental Fig. 2b**). Furthermore, because these
1394 genes were underdispersed, and therefore not included by the feature selection algorithm for
1395 the first round clustering results (state not-annotated), there was no data leakage between
1396 partitioning and cell-state spike-ins.

1397 We sought to understand the impact of labeling or not labeling batch confounded
1398 biologically meaningful signal. If an experimenter were unaware of these differences in cell-state
1399 across batches, and used only the parent population's label (as identified before *in silico* splits
1400 and signal spike-in), standard benchmarking metrics (kBET-cell-type, average silhouette width
1401 (ASW), cLISI) rewarded for erasure of biologically meaningful signal (**Supplemental Fig. 2d**),
1402 showing that perfect knowledge of cell-identity, including all biological differences between
1403 datasets, was required for accurate benchmarking with metrics that utilize user-defined cluster
1404 labels, resulting in a benchmark that rewards erasure of the unannotated batch confounded
1405 signal.

1406 Under this paradigm, we examined standard benchmarking metrics (i.e.: ARI, kBET-cell-
1407 type, cLISI, etc, 27 metrics in total) comparing apparent method performance when this batch
1408 confounded signal was or was not annotated by the experimenter with respect to the reference
1409 labels (**ED Table 2**). Indeed, 14 of 27 metrics were statistically un-correlated when comparing

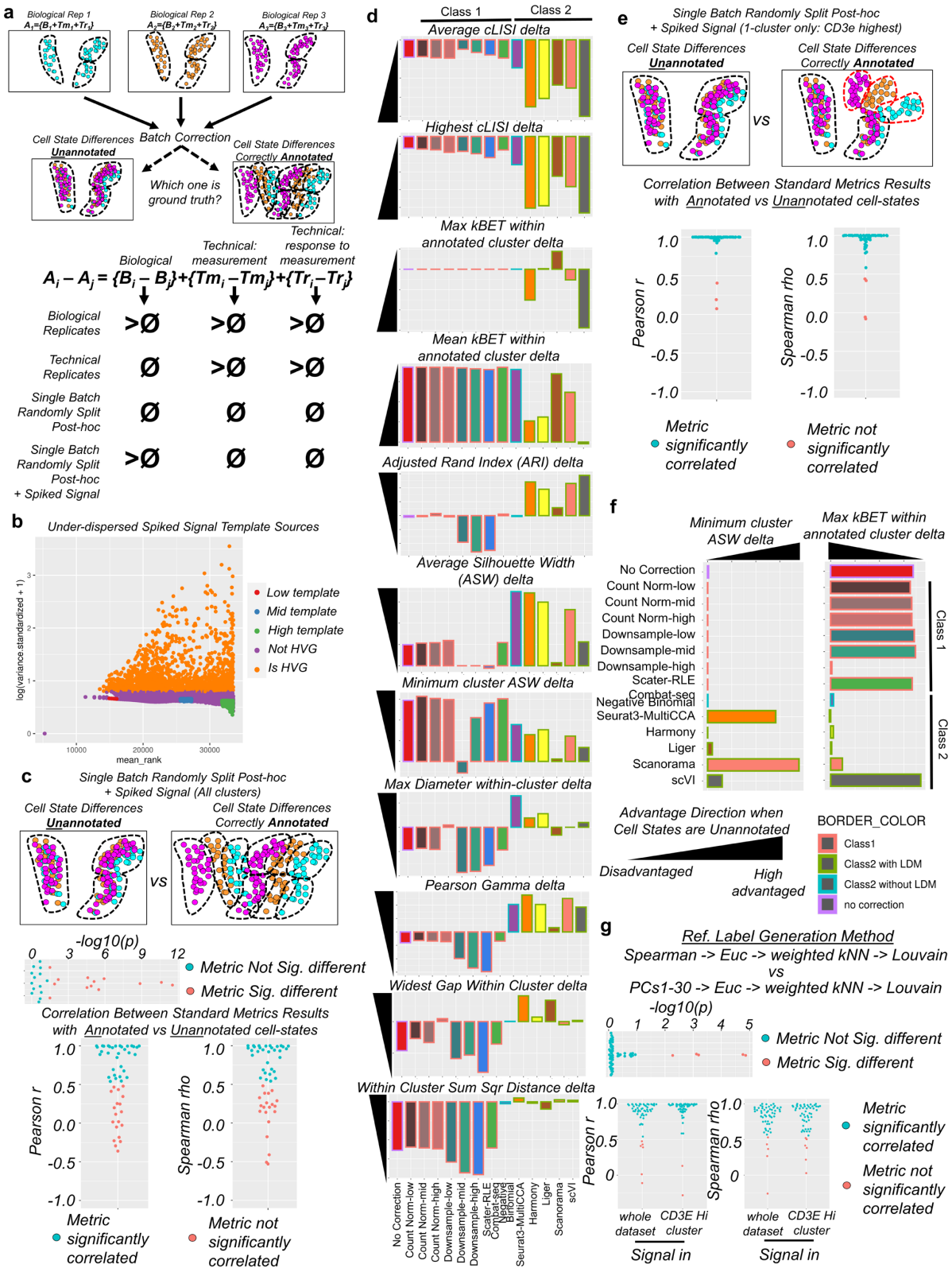
1410 results based on reference labels that did annotate state-differences vs results tabulated where
1411 state-differences were not included in reference labels (**ED Table 2a,c**). A further 13 of 27
1412 metrics showed significant and consistent changes in means dependent on perfect annotation
1413 of cell-state (**ED Table 2a,e**). This left only 13/27 metrics that were not directly impacted in
1414 either of these manners, but these metrics were each batch normalized or agnostic, yet also
1415 hinged on the assumption that reference labels should carry no additional within group variation
1416 (i.e. ASW) (**ED Table 2g**).

1417 We also assessed the impact of batch confounded cell states when applied only to a
1418 single cluster that we split it into three, through underdispersed templated cell-state spike-ins as
1419 described above. Given that most of these metrics provide global averages, across a dataset,
1420 these effects were blunted, except in the case of metrics looking for the region of "worst
1421 performance" such as minimum cluster ASW and maximum kBET-cell type (**Supplemental Fig.**
1422 **2e,f**).

1423 Additionally, the labels used by typical benchmarks were arrived at by humans using
1424 their preferred pipelines and analytic preferences. Therefore, using the labels generated by
1425 these individuals sets their given approach as equivalent to the ground-truth. This also suggests
1426 the possibility that processing steps taken during generation of reference labels can impact the
1427 apparent performance of a batch-correction method independent of accuracy, in a manner that
1428 rewards similarity to the approach used to generate refence labels. To test this, we compared
1429 results based on cluster label assignment approaches: 1) Louvain modularity based on a kNN
1430 built from the cell-cell Euclidean distances of the cell-cell Spearman correlation matrix; 2)
1431 Louvain modularity on a kNN built from Euclidean distances based on a dimensionally reduced
1432 version of the dataset (the first 30 principal components).

1433 Indeed, some commonly used metrics including cLISI, entropy, minimum ASW were
1434 significantly different based on the processing method used for generating the reference labels,
1435 systematically favoring similarity over the processing method used for reference label

1436 generation (**Supplemental Fig. 2g**; **ED Table 2**). Conversely other metrics did not show a
1437 global mean shift, but in some cases, changing the clustering pipeline used for reference labels
1438 created un-correlated metrics (particularly for the kBET-cell-type, and several others that
1439 measure within-cluster spread, which yielded non-correlated results, simply by changing this
1440 minor step in processing when reference labels were generated) (**Supplemental Fig. 2g**; **ED**
1441 **Table 2**). While not the focus of this work, the change in benchmarking metrics dependent on
1442 latent dimension mapping fits with recent findings where operating on low dimensional
1443 projections can yield representations that are not reflective of true cell-cell similarity/distance[59].



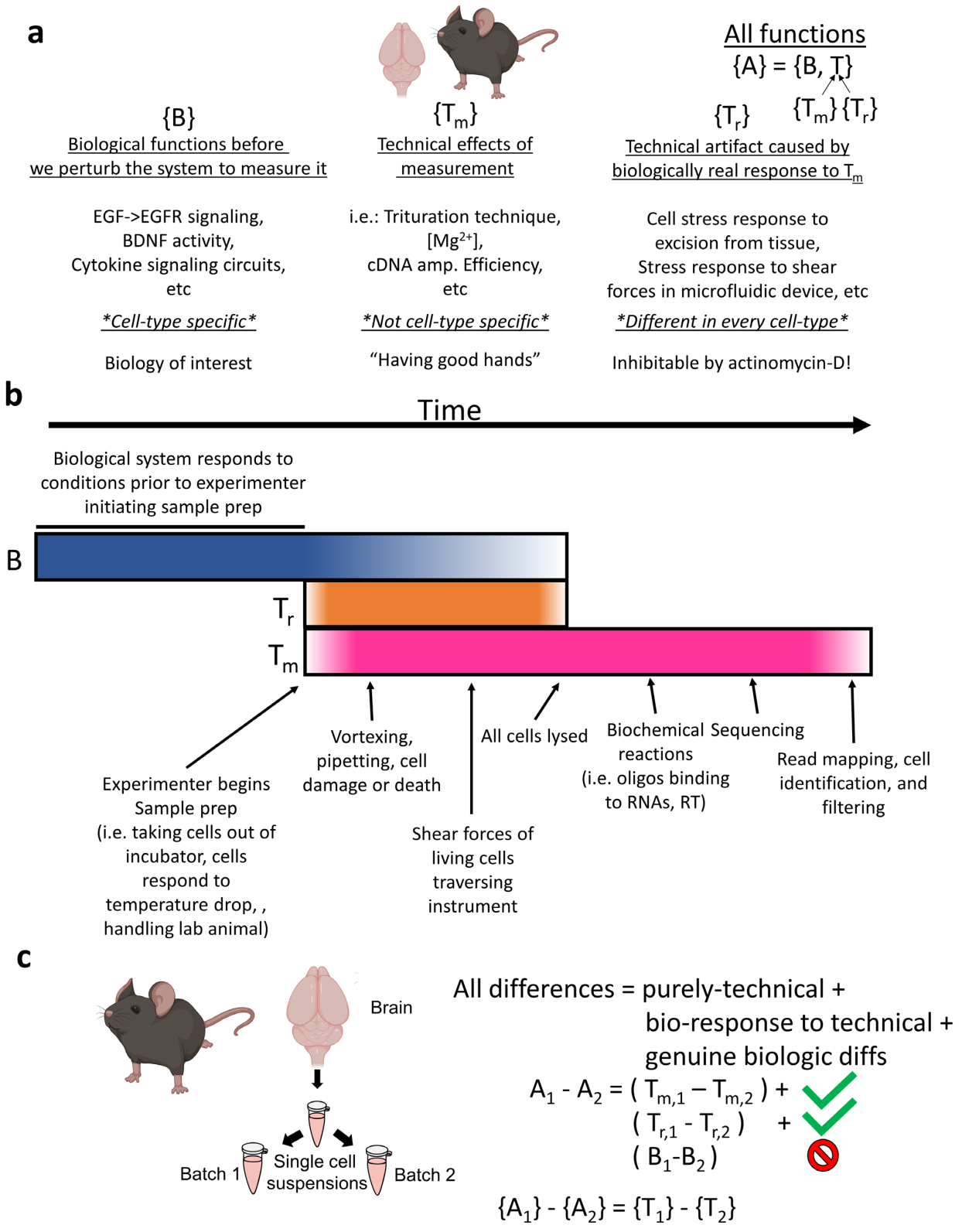
Supplemental Figure 2: Application of standard ML metrics for benchmarking batch-correction that use externally provided labels assume perfect knowledge of biology by

1447 **the experimenter. a**, A schematic example of three batches that differ broadly from each other
1448 based on a shared source of variation, but the source of difference may fall into the set of
1449 biological functions, technical functions associated with measurement (T_m), or the biological
1450 response to processing, which can also be considered a technical artifact, and not of interest
1451 (T_r). We examined the effect of using different labels as ground-truth that either keep or discard
1452 this source of variation between batches. **b**, A scatter plot of rank-mean expression and log-
1453 standardized variance, colorized by dispersion status. Using human PBMCs, we used under-
1454 dispersed genes as a "reference bank" in lieu of simulation to ensure that distributions match
1455 biological distributions; these low-/mid-/high-reference banks were used as templates to spike-in
1456 cell-state signal in all annotated clusters. **c**, We compared the effects of using the cluster labels
1457 from pre-cell-state spiked signal and labels that were "split" based on the spiked-in immune
1458 state differences. Significant differences and correlations of standard ML and batch-correction
1459 metrics comparing the without-cell-state labels or with-cell-state-annotated labels as reference.
1460 In many cases, results were significantly different or uncorrelated (salmon color), indicating that
1461 these metrics frequently require perfect knowledge of the underlying biology, otherwise the
1462 unannotated sources of variation will dramatically impact results. **d**, A selection of commonly
1463 used metrics examining the difference between using cell-state-annotated and cell-state-
1464 unannotated reference labels shows that many of these metrics systematically advantage status
1465 quo Class-2 methods. **e**, Similar results are found when cell states differ only in a single cluster
1466 (MAPK/ERK/immune signaling spiked into T-cells only), but results are notably blunted because
1467 many of these methods average over the entire dataset. **f**, However, metrics that examine the
1468 worst performing cluster show a similar pattern as with **d**. **g**, Significant differences and
1469 correlations between when reference labels were generated using a LDM pre-processing step
1470 vs using high-dimensional Spearman correlation pre-processing.
1471

1472 While we agree with the broad consensus that these metrics in fact capture *some*
1473 attributes of preservation of human-annotated biological signal, the above examples clearly
1474 demonstrate that these metrics can also incentivize erasure of non-annotated attributes of cell-
1475 identity, and can incentivize consistency with the pipeline choice during generation of reference
1476 labels. Lastly, others have also used less supervised methods to annotate conservation of
1477 biological signal, such as conservation of cell cycle or highly variable genes[60]. These metrics
1478 benchmark preservation of *within-batch* biological variation rather than *across-batch* differences:
1479 an important, albeit orthogonal question to the one investigated here.
1480
1481

1482 ***Supplemental Note-2: How can we understand what is biological and what is technical***
1483 ***when these are typically intrinsically confounded?***

1484 We began attempting to develop a functional framework to describe the often intrinsically
1485 confounded attributes of biologically meaningful variation and purely technical variation. In our
1486 framework (**Fig. 1**), we accept the premise that our observations must belong to one of these
1487 two classes: biological, B and technical, T. Notably however, within the technical set, others
1488 have previously described an interaction term in which the biological system responds to
1489 processing[21, 22]. We therefore split the technical set into two non-overlapping subsets that
1490 combine to encompass all technical effects T: 1) Those that are strictly technical effects of
1491 measurement and processing (T_m), which may include things such as enzyme activity, buffer
1492 composition differences, GC content, internal priming, etc; and 2) those that capture the
1493 biological response to processing (T_r) (**Supplementary Fig. 3a,b**). Others have previously
1494 combatted T_r using actinomycin-D to prevent active transcription as early as possible in
1495 processing[21, 22].
1496



Supplementary Fig. 3: Conceptual basis for set-theoretic framework to classify biological and technical functions. a, Schematic description of the sets of effects at play in a typical

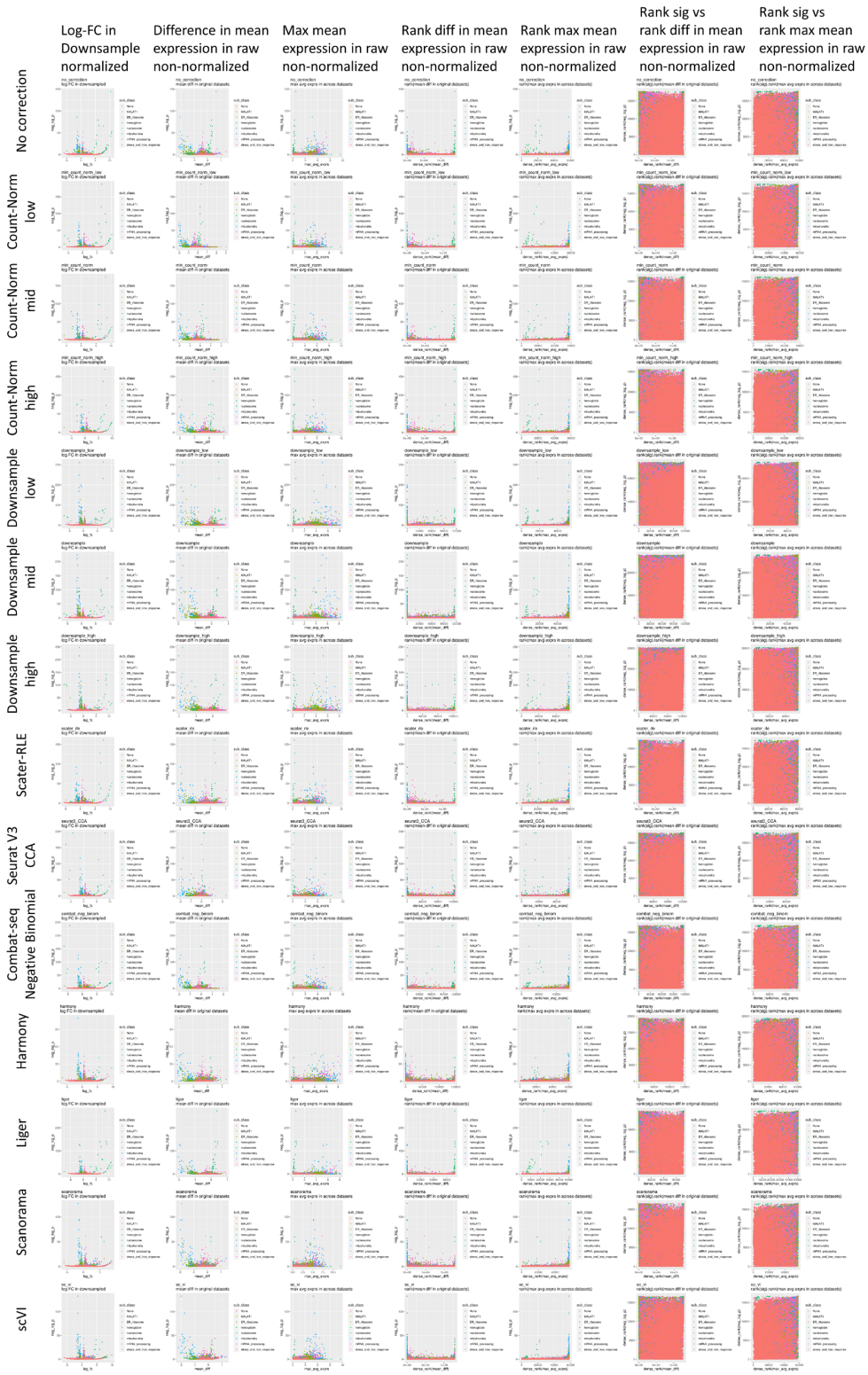
1500 scRNAseq experiment. **b**, A schematic timeline of when each set of effects will be acting in a

1501 manner that alters the observed count matrix. **c**, Using this functional framework, it becomes

1502 apparent that technical-replicates do not differ in the biological set, meaning that all difference in

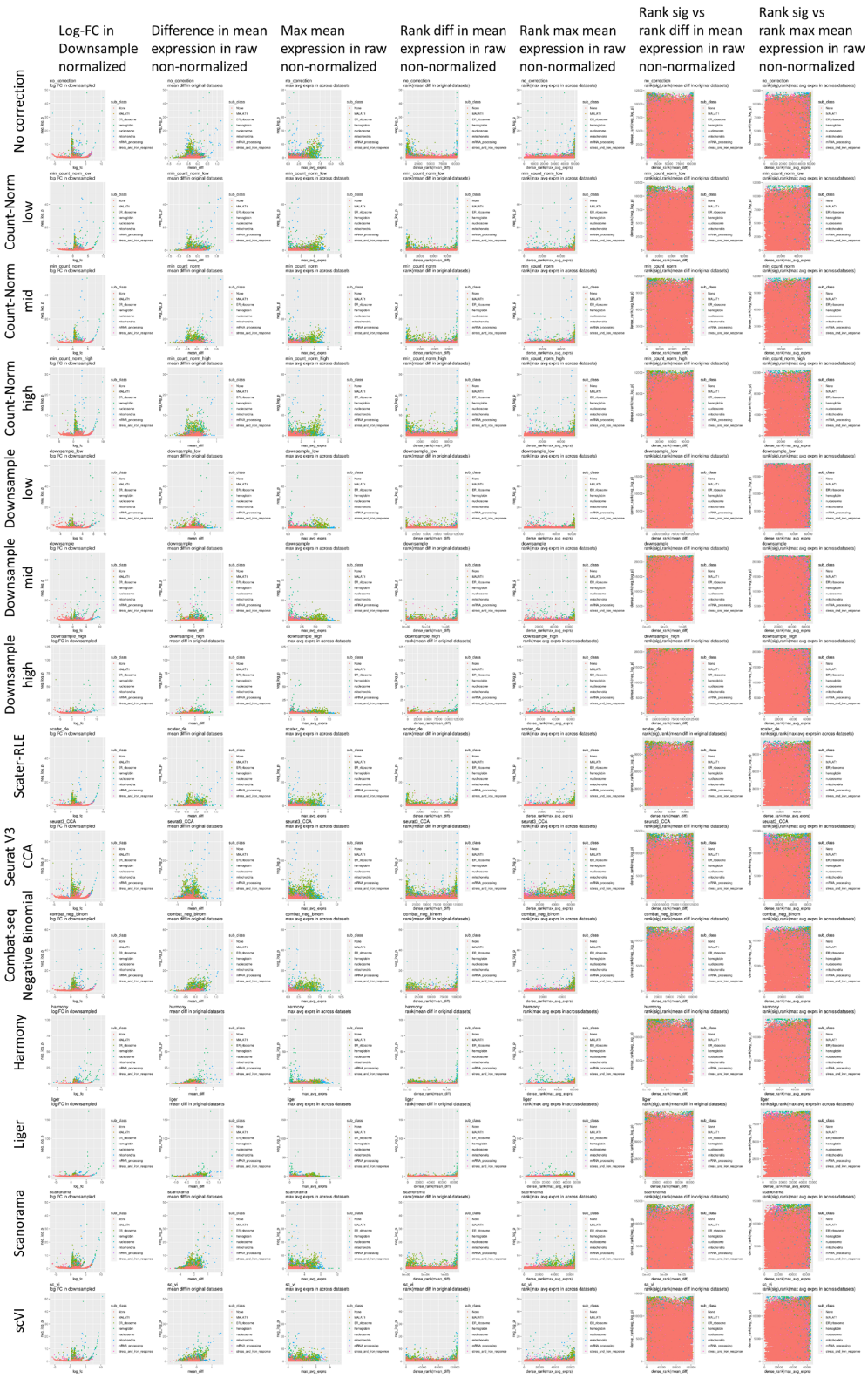
1503 the observed matrix can be traced back to the technical set.

1504 With this framework in hand, it became apparent that if we can obtain samples which are
1505 known to *only* differ by technical effects, and our original biology of interest is identical (i.e. true
1506 technical-replicates – not biological-replicates), then we can directly measure the downstream
1507 effects of the set of functions contained within T (**Supplementary Fig. 3c**).
1508 We therefore used three sets of technical-replicates in which two pipette draws from the
1509 same exact single cell suspension were run in parallel, as these must only differ biologically by
1510 random Poisson sampling of the samples taken from the single cell suspension. We utilized
1511 10x's technical-replicates all within their version 3 chemistry: Mouse brain, mouse heart, and
1512 human PBMCs. Note that the results described here may be different for other chemistry
1513 versions and other technologies.
1514 As discussed in the main text (**Fig. 1**), we found 6 categories of genes that appeared to
1515 be contained within T_m that were differentially expressed in a high percentage of clusters that
1516 were also highly conserved across tissue. These included genes encoding nuclear targeted
1517 proteins or lncRNA such as mRNA processing machinery such as capping and splicing,
1518 nucleosome binders such as High Mobility Group Box genes (HMGBs), which are both
1519 nucleoproteins and an extracellular stress related cytokines[61], and the lncRNA MALAT1. We
1520 also observed genes relating to the other large structural components of the cell as members of
1521 T_m that were differentially detected across technical-replicates in a high fraction of cells
1522 including ER associated ribosomes and mitochondria (**Supplementary Fig. 4,5**). Interestingly,
1523 in the two solid tissue preps we examined, hemoglobin and related genes were also
1524 differentially detected, yet in PBMCs where red blood cells (RBCs) are physically separated
1525 from the cells of interest during sample preparation, these genes were not differentially detected
1526 (**Fig. 1g**, **Supplementary Fig. 6**). This suggests that RBC depletion may help technical
1527 consistency even in solid tissue.

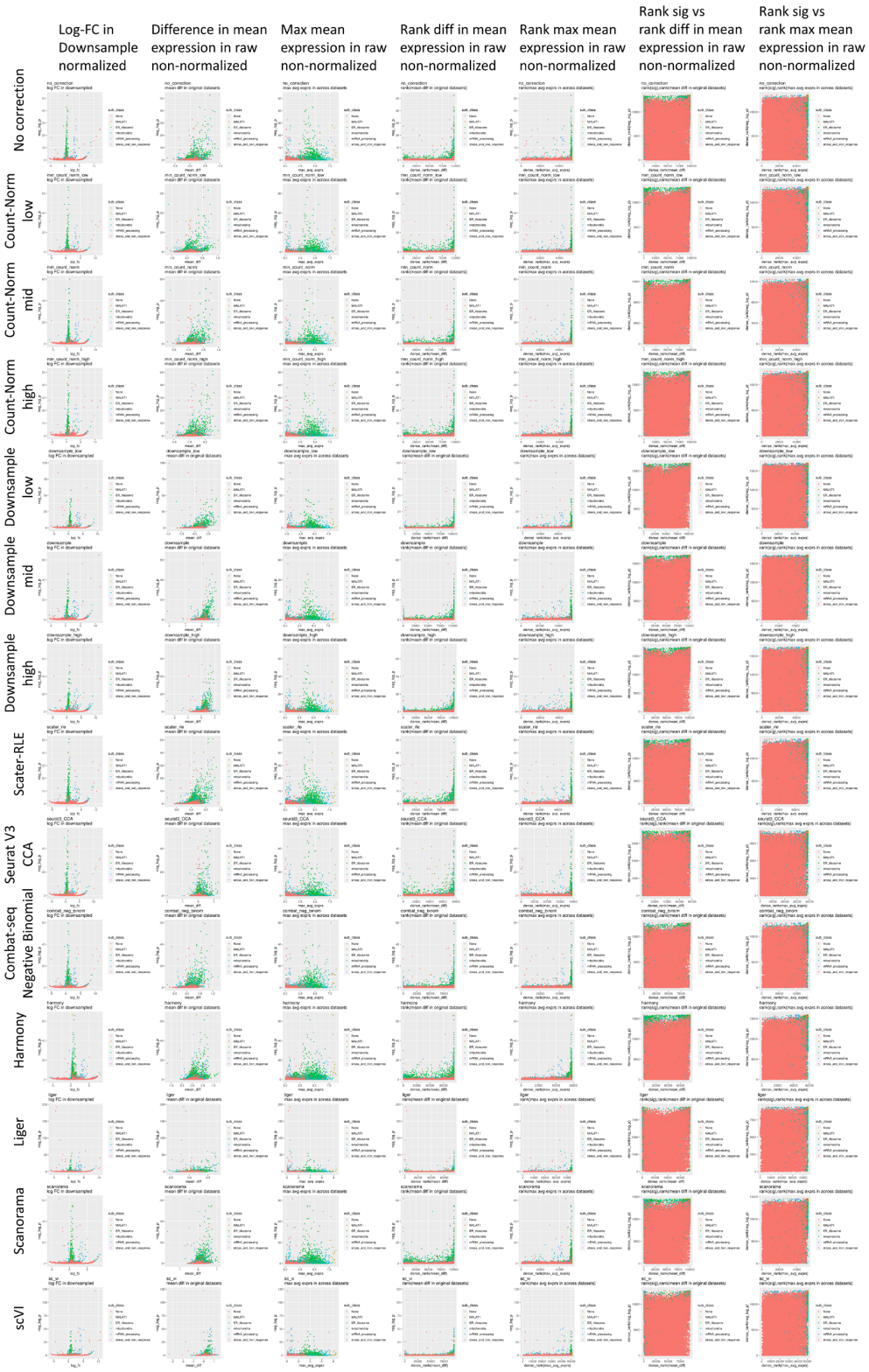


Supplemental Fig. 4: WCAB volcano plots and absolute expression differences in mouse brain technical-replicates. In the first column, volcano plots showing average log fold change

1531 (FC) (x-axis) and -log10(p-value) for differential expression within clusters across batch (WCAB)
1532 colorized by technical effect sub-class. This column's average log-FC are those calculated on
1533 the downsampled datasets quantifying the difference in relative composition rather than
1534 absolute counts; these were used for DE analyses. We assessed whether this was a technical
1535 artefact of these genes being highly expressed by examining whether *all* highly expressed
1536 genes were detected as DEGs. As shown by the subsequent columns, while higher expression
1537 led to higher *sensitivity* and the *ability* to detect DEGs, high expression on its own did not
1538 determine significance as shown by the many highly expressed genes that traversed the
1539 spectrum of significant and non-significant. This points towards a change in measured
1540 composition within these pathways, rather than a general trend attributable to differences in total
1541 UMI reads.



Supplemental Fig 5: WCAB volcano plots and absolute expression differences in mouse heart technical-replicates. See legend for **Supplemental Fig. 4**.



1545

1546 **Supplemental Fig 6: WCAB volcano plots and absolute expression differences in human**

1547 **PBMC technical-replicates.** See legend for **Supplemental Fig. 4**.

The fact that we observed genes that were highly expressed in each technical replicate that were not differentially detected indicates that the simple property of high expression was not the *cause* of this differential detection (**Supplemental Figs. 4-6**). However, as with all statistics, when dropout from shallow sampling is a factor, greater sampling depth increases the power to detect differences. In other words, greater expression levels increase power and sensitivity (due to a greater probability of detection), but were not the single determinant of significant differences.

While the above findings should primarily be taken as measures of the technical effects for 10x's Chromium V3 chemistry, this approach can be taken to further describe technical effects in other chemistries and platforms. Additionally, more work should be done using technical replicates across other tissues and species to build upon these results.

Supplemental Note-3: Proof 1: What are the theoretical implications of the functional framework for Class-1, Class-2, and Class-3 correction algorithms?

To extend our experimentally determined descriptions of the underlying elements of T, we further sought to understand whether this these different sets may or may not be altered during the transformations applied to the observed data using different batch-correction approaches. We first begin with the definitions required to understand how different classes of algorithms operate.

<u>Given the definitions:</u>

Dimensionality parameters:

1) *b*: the number of batches
2) *i* and *j*: Iterators that operate over *b*
3) *n*: the total number of cells observed
4) n_i: the number of cells in a given batch *i* (with *i* ranging from 1 to *b*)
5) *m*: the number of genes

Types of expression matrices:

1) $\boldsymbol{S}$: The initial *m* by *n* dimensional data matrix that can be thought of as the "base-transcriptome" for all cells.
2) $\boldsymbol{D}$: An *m* by *n* dimensional matrix that represents the ground truth transcriptomic data matrix, whose values can be generated through the application of all biologically derived functions in the set B (defined below); this is the matrix we are interested in measuring perfectly.
3) $\boldsymbol{D^i}$: The subset of the ground truth transcriptomic matrix that belongs to each batch *i*.

4) $\boldsymbol{O}$: The observed data matrix across all batches, representing the empirically determined version of $\boldsymbol{D}$, after application of the technical effects introduced to make the observation of $\boldsymbol{D}$.

5) $\boldsymbol{O^i}$: The set of observed data matrices, each of *m* by n_i dimensions for each batch *i*, that can be described as the output of the functions contained in T and B: $\boldsymbol{O}$=*A*($\boldsymbol{S}$).

6) $\boldsymbol{D_{c,g}}$: A subset of cells (from the *n* columns in $\boldsymbol{D}$) and or genes (from the rows of $\boldsymbol{D}$) that are assumed by the experimenter, or selected by an algorithm, to be invariant across a subset of batches.

7) $\boldsymbol{C}$: An *m* x *n* matrix that represents the batch corrected data, intended to accurately represent $\boldsymbol{D}$.

Sets of functions to transform expression matrices:

1) B^i: The theoretical set of biologically encoded functions, for each batch *i*, that reflect the complex non-linear biological processes that transform the baseline expression levels represented by $\boldsymbol{S}$ to the states the cells are in when they are profiled represented by $\boldsymbol{D}$. That is, $B^i(\boldsymbol{S}) = \boldsymbol{D}^i$. These are the functions we are trying to learn by performing the experiments giving rise to the cells being profiled. These functions encode the cell-lineage, -type, and -state prior to sample processing.

2) T^i: The set of technical effects for each batch *i* that impact the observed measures of the true transcript.

3) T_m^i: $\{T_m^i \in T^i\}$; $T_m^i = T^i | T_r^i$: The subset of T^i technical effects that are not encoded by the cells' stress or other physiology derived responses to processing (T_r^i defined below), but are technical effects from the act of measurement. These include factors such as transcript biases in PCR amplification, internal priming while using an oligo-dT primer, the effects of minor variations in buffer composition due to pipette error, diffusion of

small transcripts out of the cell after fixation for methods such as split-seq[62], read mapping errors, and more.

4) T_r^i: $\{T_r^i \in T^i\}$; $T_r^i = T^i | T_m^i$: The subset of T^i technical effects that are biologically real, caused by transcription, or other physiological changes, but are attributable to the cells' response to sample collection and preparation. This may be cell stress, but can also include of forms of processing induced transcription, or transcript degradation, etc. This can also be interpreted as the set of functions that capture the interactions between T_m^i and B (linear and nonlinear alike). Because our central focus as biologists is to understand B, rather than how our processing affects the cell, these are typically viewed as uninteresting, and in some cases have been prevented through the use of actinomycin-D, preventing the application of these functions by inhibiting transcription during sample processing[22]. However, because these are processing artifacts, they will not be of biologic interest since they reflect the experimenter's action on the basal state system B that we are trying to study. Notably, others have already reported that the cellular responses to differences to processing T_r differs per cell-type[22].

5) A^i : The union of all biologically real functions in B^i, whose outputs are passed to the set of functions contained in T for each batch i: $\{T^i, B^i\}$. Given B represents the biological functions giving rise to the transcriptome up until technical artifacts are introduced, these functions are composed as $A^i(\boldsymbol{S}) = T^i\left(B^i(\boldsymbol{S})\right)$.

6) T^{-1}: The true inverse function of T. This can be interpreted as the "perfect" batch correction algorithm.

7) $\widehat{T^{-1}}$: A batch correction function that takes as input $\boldsymbol{O}$, attempting to correct for the effects induced by functions in T, returning a corrected m x n transcriptome ($\boldsymbol{C}$) that is intended to accurately represent $\boldsymbol{D}$.

With the above definitions that describe the series of processes that impact baseline states of cells in response to experimental conditions and the isolation and processing of the cells for sequencing, we can categorize the different algorithms for correcting for batch effects into 3 classes:

1. **Class 1** $\widehat{T^{-1}}$ batch correction algorithms: Algorithms that attempt to directly model or mask the technical effects encoded by the functions in T. While these algorithms require an explicit understanding of T, they are not selection of a subset of cells and genes that are biologically equivalent across batches as these algorithms do not make any assumptions on biological equivalence across subsets of batches.
    a. Model T^{-1} without assumptions on equality or inequality of B or subsets of B;
    b. The parameterization of $\widehat{T^{-1}}$ must be known or learned *a priori*, and then for a given dataset, parameter estimates are determined to come up with the specific $\widehat{T^{-1}}$ to apply on the dataset;
    c. Some of the parameters may be convolved with the biological signal, so that some loss of the true biological signal may happen, but only to the extent that these signals covary.
2. **Class 2** $\widehat{T^{-1}}$ batch correction algorithms: Algorithms that attempt to calculate the inverse of T by learning the relationships between $\boldsymbol{O}^{i}_{c,g}$ and $\boldsymbol{O}^{j}_{c,g}$ across all batch *i* and *j* comparisons. These algorithms do not explicitly model the functions in T and B, but rather they identify subsets of genes and cells between batches in an unsupervised way, and assume the $\boldsymbol{D}^{i}_{c,g}$ matrices are equivalent to the $\boldsymbol{D}^{j}_{c,g}$ and thus by definition will remove differences between A^{i} and A^{j}. As a consequence of this strong equivalence assumption, if the $\boldsymbol{D}^{i}_{c,g} \neq \boldsymbol{D}^{j}_{c,g}$, then these batch correction procedures will condition out true biological signal (see proof below). Identify regions (cell clusters) of matched

topology, assume within those B^i and B^j are equal, and model T^{-1} as taking you back to a common $\boldsymbol{D}$ among these equal regions.

a. Because T^{-1} is determined by looking at biologically equivalent regions, the T^{-1} will be fit in a way that forces the biological signals to be the same, even if they are in fact different;

b. To see this, suppose $O^i_{c,g}$ and $O^j_{c,g}$ are determined to be comprised of *c* cells of the same type and state across the set of *g* genes, and then assumed to reflect the same biology, but suppose the true underlying state is such that we $D^i_{c,g} \neq D^j_{c,g}$;

c. **<u>Proof 1</u>**: With $O^i_{c,g} = T^i\left(B^i(S_{c,g})\right)$ and $O^j_{c,g} = T^j\left(B^j(S_{c,g})\right)$, the application of the inverse would give us that:

   i. $\left[T^i\left(B^i(\boldsymbol{S}_{c,g})\right)\right]^{-1} = \left[T^j\left(B^j(S_{c,g})\right)\right]^{-1}$

   ii. $(B^i)^{-1}\left((T^i)^{-1}(\boldsymbol{O}^i_{c,g})\right) = (B^j)^{-1}\left((T^j)^{-1}(\boldsymbol{O}^j_{c,g})\right)$

   iii. $(B^i)^{-1}(\boldsymbol{D}^i_{c,g}) = (B^j)^{-1}(\boldsymbol{D}^j_{c,g})$

   but in order for this last result (iii) to hold, given that contrary to the assumption of equivalence we have that $\boldsymbol{D}^i_{c,g} \neq \boldsymbol{D}^j_{c,g}$, $(B^i)^{-1}\left(B^i(\boldsymbol{S})\right) \neq \boldsymbol{S}$, so that the inverse fit by the algorithm would be with respect to a different biological function than that which gave rise to $\boldsymbol{D}^i_{c,g}$ (and $\boldsymbol{D}^j_{c,g}$);

d. Unlike class 3 approaches which are semi-supervised, in class 2 algorithms one is letting the algorithm pick the regions that are homologous and then acting to minimize the differences between them. This is frequently performed using a mutual nearest neighbors (MNN) algorithm.

3. **Class 3** $\widehat{T^{-1}}$ batch correction algorithms: Similar to class 2 algorithms, but perform a semi-supervised selection of c and g, cells and genes, which are assumed to share the same biological cell-lineage, -type, and -state. In practice, this takes the form of manual annotation of cell-identity labels in various datasets that are used to train an algorithm that performs an end-to-end learning of the transformation function to merge the shared labels for *c*.

For the class 2 and 3 algorithms, distance functions that measure the distance between the "matched" areas of topology are used to learn the inverse function by minimizing the distance between the areas of matched topology. The precise implementation of these distance functions vary by each algorithm. For example, Harmony performs a linear model correction on the batch factor, minimizing the least squares of the results of fuzzy clustering of the principle components identified in a PC analysis of the data restricted to the genes selected by a feature selection algorithm, with cell equivalent locations identified by gradient descent in the linear model's correction[10]. LIGER uses non-negative matrix factorization as the dimensionality reduction technique and then again uses least squares to find the minimum distances, weighting for both full dataset reconstruction accuracy and minimized across dataset distance[11, 63]. Part of Seurat's IntegrateData function includes simple subtraction of the difference between mutual nearest neighbor (MNN) "anchor" cells across datasets, $\boldsymbol{O}^{i}_{c,g} - \boldsymbol{O}^{j}_{c,g}$, which they call the 'integration matrix[64].' While the specific pipeline implementation of these local distance functions, and their implementation of $\widehat{T^{-1}}$ differs across methods, they are conceptually similar: estimate the inverse of T by minimizing the local distance between the regions of the dataset that the algorithm assumes are biologically equivalent, and apply a locally smoothed transformation that minimizes those distances.

It is worth noting the relation that the above has to traditional batch correction using linear models to regress batch using bulk data. In large part, these also assume biological equivalence across batch, like Class-2 algorithms in the context of scRNAseq; however they rely on the balancing of known biological factors across batch (age, sex, disease/treatment status, etc). This means that $B^i(\boldsymbol{S}) \approx B^j(\boldsymbol{S})$ due to supervised allocation of B across B^i and B^j symmetrically, while an additional offset for each batch is therefore likely to be caused by technical effects. However, as shown by the proof above, this would not be a safe assumption if each biological replicate were its own batch, because we would be unable to rely on the assumption of mean equality across batch due to balancing biological effects. With the advent of multiplexing single-cell omics however [2, 62, 65, 66], these algorithms may be modifiable in the future to rely on the same supervised balancing of B that traditional bulk batch correction algorithms use. However, as with traditional batch correction, this relies on high quality balancing of effects, and could also result in misleading results if important biological factors are not balanced across batch, whether those biological factors are known to the experimenter or not.

Supplemental Note-4: PMD characterization and proof of invariance to number of clusters

PMD metric definition

Definitions

- O_{actual}: An m-by-n matrix of observed cell counts, across the m identified clusters and n batches.
- E_{actual}: An m-by-n matrix of expected cell counts under the assumption (or hypothesis in the context of a statistical test) that the distribution of clusters is the same across batches. This m-by-n matrix is tabulated in a manner similar to the "expectation matrix" in the context of a χ^2 test-of-independence.
- O_{max}: An n-by-n matrix with the diagonal set to the number of cells in each of the *n* batches and all other elements of the matrix set to 0; that is, for each batch (i) with i = 0..n, O_{max}[i,i] = batch_size[i] and O_{max}[i,j] = 0 when $i \neq j$. O_{max} represents the theoretic maximum level of asymmetry across batches, under the hypothetical that each batch was comprised of a single cluster that contained all of that batch's cells.
- E_{max}: An n-by-n matrix of expected max cell counts across the n batches under the assumption of each batch being comprised of a single cluster.

Using the observed and expected contingency matrices (similar to χ^2 test-of-independence), we found that the sum of the absolute value of the difference between expected and observed matrices were equivalent, regardless of the size of *m*, the number of clusters observed within each batch ($\Sigma|\mathbf{O}_{actual} - \mathbf{E}_{actual}|$), so long as relative cluster composition by batch was conserved (**ED Fig. 2a,b**). A proof that the PMD metric has the property of invariance to the number of clusters is provided in a subsequent section (**Proof 2**).

This equivalence property allowed the creation of a hypothetical contingency table, in which the entirety of each dataset fell into a single cluster that was specific for this batch ($\mathbf{O}_{max}$). $\mathbf{O}_{max}$ can be thought of as, what a contingency matrix would look like, given no overlap in cluster composition by batch (given the input batch sizes), with only a single cluster per batch. Thus creating a hypothetical contingency matrix that maximizes the possible asymmetry by batch, yet mirrors batch sizes of the input datasets. Because there is no dependence on the number of clusters within batch on the outcome of $\Sigma|\mathbf{O} - \mathbf{E}|$ (**ED Fig. 2a**; **Proof 2**), we can calculate this for the maximum asymmetric hypothetical contingency matrix ($\Sigma|\mathbf{O}_{max} - \mathbf{E}_{max}|$) as well, and take the ratio of the actual difference relative to maximum possible (**ED Fig. 2b**):

$$\text{Raw PMD} = \Sigma|\mathbf{O}_{actual} - \mathbf{E}_{actual}| \,/\, \Sigma|\mathbf{O}_{max} - \mathbf{E}_{max}|$$

This yields the Raw PMD metric, that is bounded between 0 and 1, that quantifies how similar cluster composition is across batches relative to the version of the input that maximizes the possible asymmetry by batch. A result that equals 0 indicates that the input ($\mathbf{O}_{actual}$) showed batches where cluster abundances were *exactly* equivalent ($\mathbf{O}_{actual} \approx \mathbf{E}_{actual}$; **ED Fig. 2c**). Note that because of random Poisson sampling, it is likely that a real-world dataset would only *approach* zero, but that there would be fluctuation above zero even when there was no real difference – simply attributable to random sampling.

When no clusters appear in more than one batch, a dataset would maximize the Raw PMD value at 1 (**ED Fig. 2d**). Because of the previously mentioned equivalency of $\Sigma|\mathbf{O} - \mathbf{E}|$ regardless of the number of clusters, $\Sigma|\mathbf{O}_{actual} - \mathbf{E}_{actual}|$ will be equivalent to $\Sigma|\mathbf{O}_{max} - \mathbf{E}_{max}|$, thus making the ratio equal to 1. Unlike the lower bound however, Poisson sampling under this scenario does not yield a distribution because all clusters have a 0 percent chance of appearing in multiple batches.

Characterization of the Raw PMD metric and creation of the final PMD metric

1770 To characterize the linearity and other properties of the Raw PMD function (**ED Fig. 2**),
1771 we simulated two batches containing 10 clusters each, with variable numbers of cells in each
1772 batch as noted in the legend of **ED Fig. 2e** (10 iterations each). In combination with varying the
1773 number of cells per batch, we also incrementally changed the amount of overlap in the two
1774 batches. This was done by decreasing the number clusters that were in batch 1 that also
1775 appeared in batch 2. In all simulations, each batch still maintained 10 clusters, but progressively
1776 increasing the number of clusters that were specific to a given dataset. The most dissimilar
1777 scenario therefore contained a total of 20 clusters (2 batches, 10 clusters each that were all
1778 batch specific). The most similar scenario contained a total of 10 clusters (2 batches, 10 clusters
1779 each that were all shared across both batches evenly). This characterization is implemented in
1780 the do_full_pmd_characterization() function in the PercentMaxDiff R package that we released
1781 (https://github.com/scottyler89/PercentMaxDiff).
1782 In all simulated cases, the relationship between the number of non-shared clusters was
1783 linear with the Raw PMD value (**ED Fig. 2e**) as calculated with the equation in **ED Fig. 2b**.
1784 Interestingly however, the slope and intercept of these lines appeared to depend on the number
1785 of cells in the datasets. We plotted histograms (**ED Fig. 2f**) to characterize the distributions of
1786 Raw PMD at the intercept of each line in **ED Fig. 2e**; this showed that when there was no actual
1787 difference in underlying cluster abundance (2 batches, 10 clusters total, all with equal relative
1788 abundance), the distributions of Raw PMD appeared to be Poisson distributed. This fits with the
1789 underlying discrete sampling errors that generated the cluster by batch contingency tables.
1790 Importantly, the peak of these distributions (lambda parameter of the Poisson function)
1791 appeared to be dependent on the input dataset sizes, shown in different colors in **ED Fig. 2f**.
1792 This indicates that the Raw PMD function, while being bounded between 0 and 1 and equivalent
1793 under *idealized* circumstances regardless of the number of cells or clusters, it is affected by
1794 random Poisson sampling. This is particularly true when underpowered, as noted by the higher
1795 Raw PMD when comparing two batches with only 100 cells each, but 10 cell types. At this point,

the background noise from Poisson sampling increases the observed Raw PMD, even when the underlying distributions in cluster abundance across batch are equivalent (**ED Fig. 2f**).

An ideal metric would allow all input datasets, regardless of number of clusters or number of cells, to scale linearly with the degree of overlap between the input datasets. We therefore sought to better characterize any dependence of the Raw PMD on 1) the number of clusters and 2) the number of cells. Specifically, for each dataset in these simulations, we performed Poisson sampling Monte Carlo simulations under the hypothetical that there are no differences in relative abundance of clusters by batch (as with the Y-intercept of the lines shown in **ED Fig. 2e**). This allowed an empirical observation of the input dataset's Y-intercept under the "unpatterned by batch" scenario, while also capturing the effect of the input sample sizes and number of clusters given that these are maintained in the simulations.

This is similar to calculating the expected matrix, as with a χ^2 test-of-independence, but through simulation that matches the underlying cluster abundances and number of cells rather than just calculating the *idealized* expected matrix; this allowed the characterization of how Poisson sampling alters the Raw PMD when the underlying cluster distributions by batch are equivalent, but matching the input batch sizes. Using this background, we calculated the observed Raw PMD for each Monte Carlo simulation, under the hypothetical of no-pattern by batch, and fit the above described Poisson distributions, calculating the lambda parameter using the fitdistr function from the R library MASS.

Using 100 null datasets of each given input in the prior simulation (**ED Fig. 2e**), we observed that the lambda parameter distributions showed a clear dependence on both the total number of clusters and the size of the input datasets **ED Fig. 2g**. This can be intuitively understood as increasing the chance of random Poisson sampling away from *perfect* equivalence between batches, simply due to random sampling. Notably, this lambda parameter background is the intercept of the lines shown in **ED Fig. 2e**. Furthermore, due to the overall linearity of the shown lines, this intercept can be used to calculate the slope of the lines in **ED**

Fig. 2e as well, thus allowing a modification of the Raw PMD, that incorporates and corrects for the background distributions caused by random Poisson sampling after fitting the distribution and calculating the best fit lambda parameter (**ED Fig. 2h**):

PMD = (Raw PMD – lambda) / (1 – lambda)

Indeed, applying this correction to Raw PMD, allowed the creation of the final PMD metric which shows no dependence on input size of the batches or total number of clusters (**ED Fig. 2i**). Thus, the final PMD metric has an upper bound equal to 1, whose lower distribution is centered around 0, and scales 1-to-1 linearly with increasing batch asymmetry.

Proof 2: Proof and example of cluster invariance with PMD equation

To prove whether the PMD is invariant to the number of clusters, let $C_{i,j}$ denote the clusters from $i = 1..n$ detected (*n* clusters in total) across the batches B_j for $j = 1..b$ (b batches in total). Let $n(C_{i,j})$ denote the number of cells in cluster $C_{i,j}$, and $n(B_j) = \sum_i n(C_{i,j})$ (the sum of the cell counts across all clusters identified in batch *j* – number of cells in the batch). Thus, the total number of cells over all batches and clusters is given by:

$$N = \sum_{i=1}^{n} \sum_{j=1}^{b} n(C_{i,j}) = \sum_{j=1}^{b} n(B_j)$$

With this nomenclature, the PMD can be re-written as:

PMD Definition:

$$PMD = \frac{\sum_{j=1}^{b} \sum_{i=1}^{n} \left| n(C_{i,j}) - E(C_{i,j}) \right|}{\sum_{j=1}^{b} \sum_{k=1}^{b} \left| n(B_{j,k}) - \frac{n(B_k)}{N} n(B_k) \right|}$$

$$= \frac{\sum_{j=1}^{b} \sum_{i=1}^{n} \left| n(C_{i,j}) - \frac{n(B_j)}{N} n(C_i) \right|}{\sum_{j=1}^{b} \sum_{k=1}^{b} \left| n(B_{j,k}) - \frac{n(B_k)}{N} n(B_k) \right|},$$

where $E(C_{i,j})$ is the expected cell count for batch *j* in cluster *i* under the assumption of no association between batch and cell counts in the clusters, $n(C_i) = \sum_{j=1}^{m} C_{i,j}$ (number of cells in

cluster *i* across all batches), and $n(B_{j,k}) = I_{j,k} \cdot n(B_j)$ with $I_{j,k} = 1$ when $j = k$ and *0* otherwise.

For the denominator of the *PMD,* all of the terms are independent of the clusters and so by definition is invariant with respect to the number of clusters. We therefore need only to prove that the numerator is invariant to the number of clusters, given equal composition across batches.

To show that the numerator is also invariant, we need to show that the numerator delivers the same value for a cluster identified by one algorithm, as the value delivered for multiple clusters other algorithms may identified as representing that one cluster. The only constraint is that the proportion of cells across the different batches is preserved. Given this, without loss of generality, suppose that for cluster C_i we subdivide it into *l* subclusters:

$$C_{i_1}, \dots, C_{i_l}$$

with the condition

$$\frac{n(C_{i_k,j})}{\sum_{j=1}^{b} n(C_{i_k,j})} = \frac{n(C_{i_k,j})}{n(C_{i_k})} = \frac{n(C_{i,j})}{n(C_i)},$$

for $j = 1..b$ and $k = 1..l$. We then have that

$$n(C_{i_k,j}) = n(C_{i_k}) \frac{n(C_{i,j})}{n(C_i)}.$$

It then follows that

$$\sum_{k=1}^{l} \left| n(C_{i_k,j}) - \frac{n(B_j)}{N} n(C_{i_k}) \right| = \sum_{k=1}^{l} \left| n(C_{i_k}) \frac{n(C_{i,j})}{n(C_i)} - \frac{n(B_j)}{N} n(C_{i_k}) \right|$$

$$= \sum_{k=1}^{l} \left| \frac{n(C_{i_k})}{n(C_i)} \left(n(C_{i,j}) - \frac{n(B_j)}{N} n(C_i) \right) \right|$$

$$= \frac{\sum_{k=1}^{l} n(C_{i_k})}{n(C_i)} \sum_{k=1}^{l} \left| \left(n(C_{i,j}) - \frac{n(B_j)}{N} n(C_i) \right) \right|$$

$$= \sum_{k=1}^{l} \left| \left(n(C_{i,j}) - \frac{n(B_j)}{N} n(C_i) \right) \right|,$$

since $\sum_{k=1}^{l} n(C_{i_k}) = n(C_i)$. Thus we have that regardless of the number of clusters that subdivide any given cluster, so long as the proportion of cell counts across batches in the subdivided clusters is conserved, the *PMD* score will not change. Note also that subclusters in this context does not necessarily imply or require any hierarchical lineage relationships between clusters, but rather are simply representing different potential subdivisions of a dataset.

Example

Given two clustering results (R and R'), for two batches with 50% shared cluster composition, in which the shared cells are called either one cluster or two clusters, and third clustering result in which the relative composition was altered after cluster splitting (R''), yielding slightly more asymmetry across batches:

Clustering result 1 (R):

	B1	B2
C1	100	100
C2	0	100

Clustering result 1 (R'):

	B1	B2
C1	25	25
C2	75	75
C2	0	100

Clustering result 1 (R''):

	B1	B2
C1	25	75
C2	75	25
C2	0	100

Yielding the expected matrices for the numerator (E_{actual}):

E_{actual} (R):

	B1	B2
C1	66.66	133.3
C2	33.33	66.66

E_{actual} (R'):

	B1	B2
C1	16.66	33.33
C2	50	100
C2	33.33	66.66

E_{actual} (R''):

	B1	B2
C1	33.33	66.66
C2	33.33	66.66
C2	33.33	66.66

And the difference between O and E_{actual}

O-E_{actual} (R):

	B1	B2
C1	33.33	-33.33
C2	-33.33	33.33

O-E_{actual} (R'):

	B1	B2
C1	8.33	-8.33
C2	25	-25
C2	-33.33	33.33

O-E_{actual} (R''):

	B1	B2
C1	-8.33	8.33
C2	41.66	-41.66
C2	-33.33	33.33

We can see that the numerator will not change between R and R' because R:C1's values are simply split between R':C1 and R':C2, and will therefore sum to the same value. However, this is not true for R'' whose relative composition was changed with this subdivision.

When we complete the calculation of the PMD numerator, we observe:

Sum(abs(O-E_{actual}) (R):	Sum(abs(O-E_{actual}) (R'):	Sum(abs(O-E_{actual}) (R''):
133.3333	133.3333	166.6666

Showing that with R and R', whose cluster composition was held constant, give the same PMD numerator value. Lastly, dividing by the universal denominator, as calculated independently from the clustering results from O_{max} and E_{max}, yields the final PMD_{raw} metrics for these idealized examples.

PMD_{raw} (R):	PMD_{raw} (R'):	PMD_{raw} (R''):
133.3333/266.666 = 0.5	133.3333/266.666 = 0.5	166.6666/266.666 = 0.625

It is important again to note that actually observing ideal results from Poisson sampling is exceedingly rare, making it important to apply the lambda correction as described in **ED Fig. 2**, thus accounting for Poisson sampling noise.

Characterization of PMD, χ^2 statistic, χ^2 statistic -log10(p-value), Cramer's V, average inverse Simpsons index, and average count normalized Shannon's entropy at the extremes of similarity and difference between batches

Analysis of contingency tables has been thoroughly investigated for many different purposes throughout history. Some of these methods include χ^2 test-of-independence, it's -log10(p-value), Cramer's V, average inverse Simpsons index, and average count normalized Shannon's Entropy. However, in our specific use case, we sought a metric with very specific properties that could linearly quantify column similarity in a manner independent of differences in column totals, yielding the same result when relative row-wise composition is held constant between is the columns, even when row-wise dimensionality can be variable.

χ^2 and its corresponding p-value has been a standard metric for quantifying the independence of two sets of variables and their co-occurrence. However, in most cases the χ^2 or its p-value alone are the statistics of interest. Here however, with preliminary testing, we found some non-linearities, and non-robustness to differences in contingency table dimensionality and that differences in column totals impacted results.

Cramer's V statistic is a previously published approach that builds on χ^2 to create a correlation-like metric based on counts of discrete classes in a manner similar to PMD[67]. A prior report indicated that this statistic demonstrated a high level of correctable bias, especially with contingency tables of dimension greater than 2x2, as will usually be the case in a cluster x batch contingency table.[68] However, in initial testing, we found that the bias corrected Cramer's V was still non-linear across the spectrum of similarity.

1920 Diversity and information theoretic metrics can also intersect with this problem in the
1921 form of Inverse Simpsons index[69] and Shannon's entropy[70]. These metrics were originally
1922 designed to quantify diversity and information contained within strings respectively; however,
1923 since their introduction have become widely used across many fields of study.

1924 To demonstrate specific scenarios in which the PMD metric provides an advantage over
1925 other metrics such as the χ^2 statistic, its -log10(p-value), bias corrected Cramer's V (as
1926 implemented in rcompanion[71]), inverse Simpsons index (average across clusters), and count
1927 normalized Shannon's Entropy (average across clusters), we designed a series of comparisons
1928 of all six metrics under different scenarios. An ideal metric would have all of the following
1929 characteristics:

1930 1. **Robustness to differences in power**: An ideal metric would be unaffected by the total
1931 number of cells in the batches. For example, two datasets of equal size (1,000 cells
1932 each) that were sampled from the same distribution of cell clusters, should be identified
1933 as equivalent to each other. The returned metric should not be different if these datasets
1934 were 5,000 cells each. This allows direct interpretation on the degree of overlap between
1935 batches, regardless of their size.
1936 2. **Robustness to differences in batch size**: An ideal metric for comparing how similar or
1937 different two batches are to each other in terms of their cluster composition, would be
1938 robust to differences in the size of two batches. For example, in the datasets used here,
1939 the dataset of 1,000 brain cells integrated with 10,000 brain cells should yield the same
1940 result as an integration of 1,000 brain cells with another dataset of 1,000 brain cells – in
1941 both cases, the desired metric would indicate that the batches come from the same
1942 cluster composition – unaffected by differences in batch sizes.
1943 3. **Invariance to the number of clusters found within a given batch**: Batch-correction
1944 algorithms are expected to leave different levels of within-batch variation and across-
1945 batch variation, yet, with real-world datasets, we do not know the ground-truth number of

1946 clusters. However, when integrating datasets that came from the same single cell
1947 suspension, we know that cluster composition should be the same. Additionally, we can
1948 make reasonable assumptions that intestinal epithelial cells should not be found in the
1949 same clusters as neuronal cells. However, we cannot claim to know the ground-truth
1950 number of clusters within these datasets. To be able to benchmark batch-correction
1951 algorithms, which leave different levels of variation in a dataset, without making
1952 assumptions on the number of clusters, a metric is required that would report the degree
1953 to which cluster composition was similar across batches, while still allowing the number
1954 of clusters to vary, as this is determined by unbiased algorithms. For example, if two
1955 batch-correction algorithms were used, on intestine and brain datasets, and both found
1956 no batch overlap, but algorithm one found three brain derived clusters, while the other
1957 found 4 brain derived cell clusters, an ideal metric would indicate that in both cases, the
1958 methods were equivalent in yielding no batch overlap. As formally proven above, PMD is
1959 invariant to the number of clusters found, however, we will also show this with Monte
1960 Carlo simulations (**Supplemental Fig. 7,8**).

1961 4. **Linearity across the spectrum of batch similarity:** For two batch-correction
1962 algorithms to be fairly compared to each other, there must be a consistent mathematical
1963 relationship between the degree to which batches are similar/dissimilar, and the reported
1964 metric. In the particular case where the relationship between the percentage of overlap
1965 in cluster composition and the reported metric is linear, the difference between having
1966 50% and 60% of cells derived from the same cluster composition would be the same
1967 difference in the reported metric as 80% to 90% of cells coming from shared cluster
1968 composition. This gives immediately interpretable results.

1969

1970 We therefore sought to benchmark PMD against the more traditional metrics such as χ^2
1971 statistic, its -log10(p-value), the less frequently used χ^2 statistic derived bias corrected Cramer's

1972 V, inverse Simpson's index, and Shannon's entropy, under different idealized or adverse
1973 situations to assess for all of these properties. In each case, we calculated these metrics for a
1974 simulation of integrating two different batches, with differing numbers of clusters, and differing
1975 degrees of overlap between batches.

1976 Importantly, what it means to have a given percentage overlap in cluster composition
1977 can come in two different modes. The first mode, and our first simulation paradigm
1978 (**Supplemental Fig. 7a**), is by sharing different numbers of clusters. For example, with two
1979 batches of four equally abundant clusters each could share no clusters (eight clusters total, four
1980 in each batch). They could share one out of the four clusters that appear in each batch (25%
1981 similar composition, 75% non-overlapping composition), etc. The second mode by which two
1982 batches can have varying degrees of overlap in cluster composition, is by sharing all, or some
1983 clusters, but having different relative abundance of the more-shared or less-shared clusters
1984 (**Supplemental Fig. 8a**). In our implementation of this second paradigm, we allocated a single
1985 cluster to batch 1, and vary the relative abundance of this cluster in batch 2, while batch 2's
1986 remaining clusters are unique to it. In the case of 0% overlap, 0% of the second batch's cells
1987 were allocated to cluster #1 (the cluster that harbors all of the cells from batch 1), and 100% of
1988 batch 2's cells are allocated to clusters number 2 to n with equal probability. In the case of 25%
1989 overlap, 25% of batch 2's cells are allocated to the first cluster, while the remaining 75% are
1990 distributed evenly across the remaining batch 2 specific clusters. Similarly, in the case of 50%
1991 overlap, 50% of batch 2 is allocated to cluster #1, and the remaining 50% of cells are allocated
1992 evenly to clusters unique to batch 2, etc. Note that in this second simulation paradigm, when
1993 there is 100% overlap between batches 1 and 2, each cell from batch 2 has a 100% probability
1994 of being allocated to cluster #1, thus being equivalent to having only a single cluster across both
1995 batches.

1996 We utilized both of these simulation paradigms (**Supplemental Fig. 7a,8a**) to
1997 benchmark PMD against the previously mentioned status quo metrics, to evaluate each of these

1998 metric's adherence to the above listed four desirable traits of an ideal metric for comparing
1999 cluster composition after batch-correction, integration and unsupervised clustering.
2000 For our first Monte Carlo simulation in this comparison, we created two fully equivalent
2001 batches, with equal abundance (1000 cells each), and equal probabilities of sampling each
2002 cluster. We further simulated equivalent batches that harbored differing number of clusters (still
2003 all equivalent across batch, such that there was no pattern of clusters across batch). As
2004 expected from our theoretical work, PMD was robust to changes in dimensionality of the input
2005 matrix; however, with increasing number of clusters, the χ^2 statistic rose linearly the number of
2006 clusters in the batches (**Supplemental Fig. 7b**), thus demonstrating that the χ^2 statistic gives
2007 linear background signal when batches are equivalent, increasing with the number of clusters in
2008 each batch. Similarly, Cramer's V showed a slight trend upwards. This is likely a phenomenon
2009 similar to the Raw PMD metric with a loss of power due to random sampling. Note that this
2010 pattern goes away when there are only a single cluster simulated (**Supplemental Fig. 8b**).
2011 Additionally, in the special case of one cluster, Cramer's V cannot compute as it requires ≥2x2
2012 contingency table. Interestingly, under the first simulation paradigm inverse Simpson's index
2013 and Shannon's entropy show negative trends with increasing number of clusters, but no trend
2014 under the second simulation paradigm (**Supplemental Fig. 7b,8b**).
2015 Next, we performed a Monte Carlo simulation in which batches were completely
2016 orthogonal in their cluster composition, with no overlap; still simulating a variable number of
2017 clusters within each batch, but with no overlap (100% different), again equivalent in batch sizes.
2018 As expected, because of the invariance property, PMD was robust to different numbers of
2019 clusters, when there was no overlap in cluster composition between batches (**Supplemental**
2020 **Fig. 7c**). The χ^2 statistic however, lacked the invariance property, particularly when moving from
2021 1 cluster per batch to 4 clusters per batch, but then plateaued. The -log10(p-value of the χ^2 test-
2022 of-independence however bottomed out with the lowest possible p-value due to floating capacity
2023 errors at -log10(p-value) = ~308. Interestingly, Cramer's V was negatively correlated with the

number of clusters despite batches being completely different in each circumstance. The same pattern of results was observed under the second simulated paradigm (**Supplemental Fig. 8c**). Interestingly, similar to PMD, in this scenario inverse Simpson's index and Shannon's entropy, showed no relationship with the number of clusters, consistently yielding 1 or 0 respectively under both simulation paradigms (**Supplemental Fig. 7c,8c**).

We next assessed the effect of differences in batch sizes, where one batch was larger than the other. When cluster composition was the same, PMD was centered around zero, regardless of the number of clusters or differences in batch size (**Supplemental Fig. 7d**), whereas the χ^2 statistic was highly significant with one cluster, but dropped to non-significant with more than one cluster. The degree to which the one cluster results were significant by the χ^2 statistic was dependent on the dataset sizes as shown by the difference between dotted and solid lines (**Supplemental Fig. 7d**). Similarly to **Supplemental Fig. 7b** Cramer's V showed some level of background above zero likely due to sampling error, as with Raw PMD. Inverse Simpson's index and Shannon's entropy showed different slopes dependent on the differences in dataset size, and Shannon's entropy showed a global downward trend with increasing number of clusters.

Under the second simulation paradigm, where the number of clusters was only expanded in the second batch, this simulation was equivalent to having only a single cluster in total that held all cells from both batches. Consistent with the prior simulation, PMD was 0 (indicating no difference), while the χ^2 statistic and its corresponding p-values indicated highly significant differences by batch in a manner dependent on batch sizes (**Supplemental Fig. 8d**). Again, in the special case of a single cluster present, Cramer's V fails to compute. Inverse Simpson's index differed by size of the datasets, but under this simulation scenario, Shannon's entropy was invariant with differing number of clusters or batch size (**Supplemental Fig. 8d**).

We next assessed a scenario similar to the above (**Supplemental Fig. 7e,8e**), but now with no overlap in cluster composition. PMD, inverse Simpson's index, and Shannon's entropy

2050 gave consistent results with PMD=1 (completely different), inverse Simpson's index=1, and
2051 Shannon's entropy=0, regardless of input batch sizes (**Supplemental Fig. 7e**). The χ^2 statistic
2052 however was fully dependent on the size of the input datasets. This indicates that differences in
2053 cell-filtering through variable QC pipelines, or other sources will directly impact the ability to use
2054 the χ^2 statistic or its -log10(p-value) for comparisons. Cramer's V was dependent on both the
2055 sizes of the datasets and the number of clusters, despite the composition of clusters across
2056 batches being completely different in all cases. Furthermore, the invariance property of PMD
2057 and robustness to input dataset sizes demonstrate that it can be used, even to compare
2058 completely different dataset integrations, such as brain-heart concordance, relative to intestine-
2059 PMBC concordance. The same results were seen under the second simulation paradigm
2060 (**Supplemental Fig. 8e**).
2061
2062
2063 ***Quantifying the linearity of PMD, χ^2 statistic, and χ^2 statistic -log10(p-value), and***
2064 ***Cramer's V across the spectrum of similarity***
2065 We sought to quantify the linearity of these metrics, at intermediate levels of batch
2066 similarity, rather than only at the extreme poles (completely different and identical cluster
2067 composition). Linearity is important because it will enable the satisfaction of many parametric
2068 frequentist statistics assumptions without the need for additional transformations. For example,
2069 that the difference between 0 and 0.5 be equivalent to the difference between 0.5 and 1.0.
2070 To test the property of linearity across the spectrum of similarity, we first created
2071 simulations in which batches were of equal size, with differing levels of overlap as determined
2072 by the presence or absence of shared clusters (**Supplemental Fig. 7a**). PMD gave consistent
2073 results, in which the PMD readout matched the simulated percentage of clusters that appeared
2074 in a batch specific manner (**Supplemental Fig. 7f**). For example, when 25% of clusters
2075 appeared in a batch specific manner, PMD was centered around 0.25, etc (**Supplemental Fig.**

2076 **7f**, upper left panel). This concordance manifested in full linearity across the spectrum of
2077 overlap under this simulation paradigm for PMD (**Supplemental Fig. 7f**, lower left panel).
2078 Interestingly, this pattern held true for the χ^2 statistic as well (**Supplemental Fig. 7f**, middle-left
2079 panels). However, the χ^2 -log10(p-value) was non-linear, and was inversely proportional to the
2080 number of clusters (**Supplemental Fig. 7f**, middle panels). Similarly, Cramer's V was non-linear
2081 across the spectrum of similarity (**Supplemental Fig. 7f**, middle-right panels). Inverse
2082 Simpson's index and Shannon's entropy were both non-linear and varied with number of
2083 clusters (**Supplemental Fig. 7f**; right panels). Importantly, however, many of these patterns
2084 changed under our second simulation paradigm.

2085 Using the second simulation paradigm in which the percent overlap was determined by
2086 the relative abundance of a single potentially shared cluster, we again found that PMD was
2087 robust to the number of clusters and linear across the spectrum of batch similarity
2088 (**Supplemental Fig. 8f**, left panels). The χ^2 statistic however showed an exponential pattern
2089 across the spectrum of similarity that was consistent regardless of the number of clusters
2090 (**Supplemental Fig. 8f**, middle-left panels) and the -log10(p-values) were slightly non-linear as
2091 well, in a manner dependent on the number of clusters (**Supplemental Fig. 8f**, middle-left lower
2092 panel). Under these circumstances, Cramer's V was largely linear (unlike the first simulation
2093 paradigm) (**Supplemental Fig. 8f**, middle-right lower panel), however, as previously mentioned,
2094 the special case of one cluster fails to compute. Inverse Simpson's index and Shannon's
2095 entropy both collapse towards one and zero respectively as the number of clusters increased,
2096 independently of the actual batch similarity (**Supplemental Fig. 8f**, right top panels). This
2097 resulted in very unstable estimates of similarity when compared to the actual similarity
2098 (**Supplemental Fig. 8f**, right bottom panels).

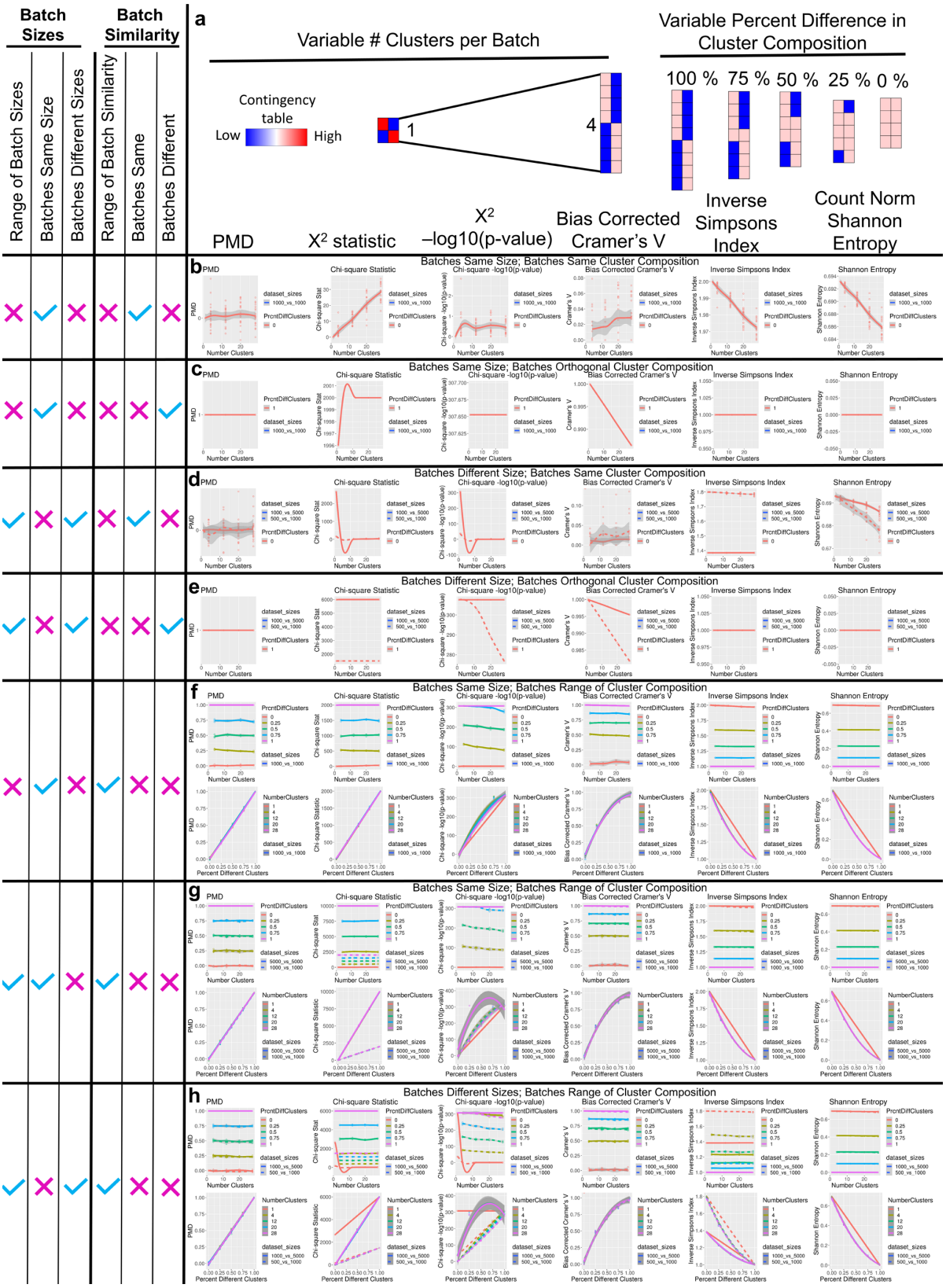
2099 Next, we simulated a range of batch similarity, with identical batch sizes, but testing the
2100 effect of power under two scenarios – integrating two batches with 5,000 cells each (solid lines),
2101 compared to integrating two batches with 1,000 cells each (dashed lines). PMD was linear

2102 across the spectrum of similarity and robust to different levels of power under both simulation
2103 paradigms (**Supplemental Fig. 7g**, left panels). Interestingly, the χ^2 statistic was again linear
2104 within the two simulated batch sizes (two batches, 5,000 cells each and two batches 1,000 cells
2105 each), however, the slope of these lines was different, in which the simulated integrations with a
2106 larger number of cells had a greater slope compared to the smaller simulations (**Supplemental**
2107 **Fig. 7g**, middle-left panel). With the χ^2 test-of-independence -log10(p-value), we observed
2108 floating point errors, as the p-value approaches zero, that resulted in non-linearity with dataset
2109 similarity in the larger dataset (**Supplemental Fig. 7g**, middle-left panel). Cramer's V was
2110 robust to dataset sizes, but was non-linear across the spectrum of similarity (**Supplemental Fig.**
2111 **7g**, middle-right panel). Inverse Simpson's index and Shannon's entropy were again non-linear
2112 (**Supplemental Fig. 7g**, middle-right panels).

2113 Interestingly, under the second simulation paradigm, the χ^2 statistic was exponential in
2114 relation to batch similarity, instead of being linear when batches as with the first simulation
2115 paradigm (**Supplemental Fig. 8g**, lower middle panel). Cramer's V was robust to dataset size
2116 and linear, because batch sizes were identical, but failed under the special case of one cluster
2117 (**Supplemental Fig. 8g**, middle-right panels). Similarly to the first simulation paradigm, PMD,
2118 was again robust to dataset size, number of clusters, and linear across the spectrum of
2119 similarity (**Supplemental Fig. 8g**, left panels), while inverse Simpson's index and Shannon's
2120 again entropy dropped to one and zero respectively, when ground truth difference was
2121 intermediate (**Supplemental Fig. 8g**, right top panels), resulting in highly variable and non-
2122 linear estimates (**Supplemental Fig. 8g**, right bottom panels).

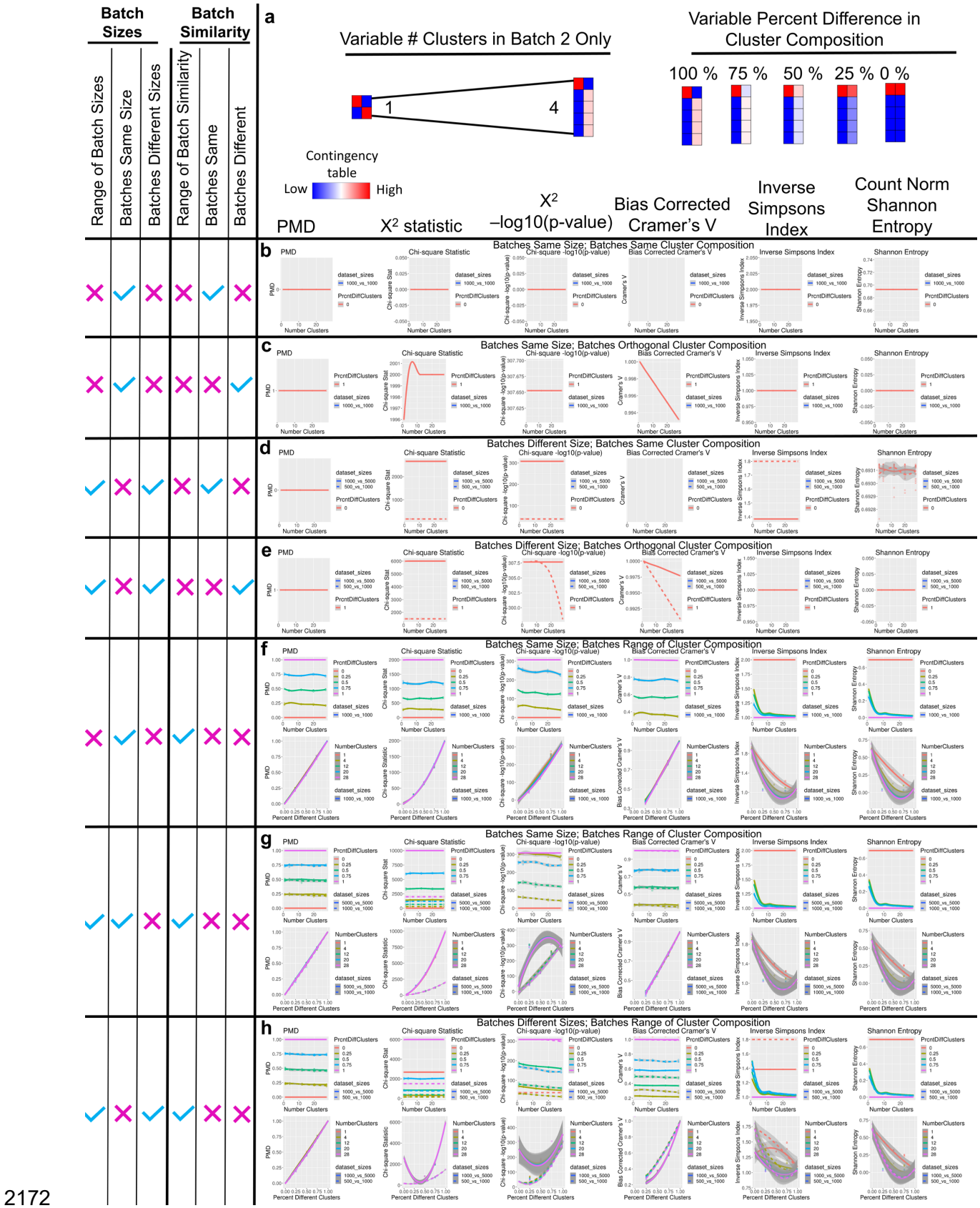
2123 Lastly, we performed the same simulations as in (**ED Fig. 7g,8g**, but with datasets that
2124 were of different sizes within each integration (solid lines: batch1:1,000 cells, batch2:5,000 cells;
2125 dashed lines: batch1:500 cells, batch2:1,000 cells). PMD was again robust to differences in
2126 batch size, differences in power (as indicated by overlapping solid and dashed lines), was
2127 unaffected by the number of clusters, and operated linearly with batch similarity under both

2128 simulation paradigms, (**Supplemental Fig. 7h,8h**, left panels). The χ^2 statistic and its -log10(p-
2129 value) operated linearly only under the first simulation paradigm, but in a manner that was
2130 dependent on the number of clusters and power, while again under the second simulation
2131 paradigm, the χ^2 statistic and its -log10(p-value) were non-linear with batch similarity, dependent
2132 on the number of cells, the difference in batch size, and in the case of -log10(p-value),
2133 decreased with increasing number of clusters (**Supplemental Fig. 7h,8h**, middle-left panels).
2134 Under the first simulation paradigm, Cramer's V was non-linear with percent of clusters that
2135 were unique to a given batch, but was robust to different dataset sizes (**Supplemental Fig. 7h**,
2136 middle-right panels). Under the second simulation paradigm, Cramer's V was both non-linear
2137 and dependent on dataset sizes under this simulation paradigm (**Supplemental Fig. 8h**,
2138 middle-right panels). Surprisingly, the pattern of non-linearity in Cramer's V was different under
2139 the two simulation paradigms (**Supplemental Fig. 7h,8h**, middle-right-lower panels). Inverse
2140 Simpson's index and Shannon's entropy followed the same patterns of collapse towards one
2141 and zero for intermediate similarity dependent on the number of clusters resulting in highly
2142 variable estimates (**Supplemental Fig. 7h,8h** right panels).
2143 Overall, these results demonstrate that PMD is 1) invariant to the number of clusters, 2)
2144 robust to the size of input datasets, 3) robust to differences between batch sizes, and 4)
2145 operates linearly across the spectrum of similarity in batches when simulated as
2146 presence/absence of shared clusters across batches (first simulation paradigm) or relative
2147 abundance of a shared cluster (second simulation paradigm). The χ^2 statistic, its -log10(p-
2148 value), bias corrected Cramer's V, inverse Simpson's index, and Shannon's entropy did not
2149 consistently share these properties robustly to all simulated circumstances.


Batch Sizes
Batch Similarity
Range of Batch Sizes
Batches Same Size
Batches Different Sizes
Range of Batch Similarity
Batches Same
Batches Different
a
Variable # Clusters per Batch
Contingency table
Low
High
1
4
Variable Percent Difference in Cluster Composition
100 %
75 %
50 %
25 %
0 %
PMD
X² statistic
X² –log10(p-value)
Bias Corrected Cramer's V
Inverse Simpsons Index
Count Norm Shannon Entropy
b Batches Same Size; Batches Same Cluster Composition
c Batches Same Size; Batches Orthogonal Cluster Composition
d Batches Different Size; Batches Same Cluster Composition
e Batches Different Size; Batches Orthogonal Cluster Composition
f Batches Same Size; Batches Range of Cluster Composition
g Batches Same Size; Batches Range of Cluster Composition
h Batches Different Sizes; Batches Range of Cluster Composition
Number Clusters
Percent Different Clusters

2150

Supplemental Fig. 7: Extended Data Figure 9: Characteristics of PMD, χ^2 test-of-independence statistic, χ^2 -log10(p-value), Cramer's V, inverse Simpsons index, and count normalized Shannon's entropy. a, Schematic of the first simulation paradigm to vary batch similarity. Batches were comprised of the same number of clusters, but varied by the number of clusters that were batch specific clusters. **b,** PMD, χ^2, χ^2's -log10(p-value), bias corrected Cramer's V, inverse Simpson's index, and Shannon's entropy (y-axes) for the scenario in which batches have varying number of clusters (x-axis), batches contain the same number of cells, and batches share the same cluster composition. **c**, Batches are fully orthogonal, sharing no overlap in cluster composition, but batches contain the same number of cells with varying number of clusters (x-axis). **d**, Batches are identical, but with different batch sizes (indicated by solid & dashed lines). Note that Cramer's V does not compute due to small contingency table size. **e**, Batches are different sizes (solid and dashed lines), and orthogonal cluster composition. **f-h**, (top row) Panels depicting the relationship between the simulated number of clusters (x-axis), and the corresponding metric (y-axis). Color indicates the ground-truth batch overlap. (Bottom row) x-axis is the ground-truth percent differences between batches, while colors indicate the number of clusters. In all panels, solid or dotted lines correspond to the indicated dataset sizes. **f**, Batches are the same size, over a range of cluster compositions. **g**, Similar to **f**, Batches are the same size, over a range of similarities, but now simulated with two different batch sizes (B1 & B2:1000 or 5000 cells each; solid or dotted lines). **h**, Batch sizes are each different within integration tasks, with variable scales of difference shown by solid or dotted lines.


Batch Sizes
Batch Similarity
Range of Batch Sizes
Batches Same Size
Batches Different Sizes
Range of Batch Similarity
Batches Same
Batches Different
a
Variable # Clusters in Batch 2 Only
1
4
Contingency table
Low
High
Variable Percent Difference in Cluster Composition
100 %
75 %
50 %
25 %
0 %
PMD
X2 statistic
X2 –log10(p-value)
Bias Corrected Cramer's V
Inverse Simpsons Index
Count Norm Shannon Entropy
b
Batches Same Size; Batches Same Cluster Composition
c
Batches Same Size; Batches Orthogonal Cluster Composition
d
Batches Different Size; Batches Same Cluster Composition
e
Batches Different Size; Batches Orthogonal Cluster Composition
f
Batches Same Size; Batches Range of Cluster Composition
g
Batches Same Size; Batches Range of Cluster Composition
h
Batches Different Sizes; Batches Range of Cluster Composition
Number Clusters
Percent Different Clusters

2172

2173 **Supplemental Fig 8: Benchmarking PMD against status quo metrics with altered cluster**

2174 **composition by relative abundance. a**, The second simulation paradigm schematic for

2175 benchmarking performance with changing cluster relative abundance, rather than

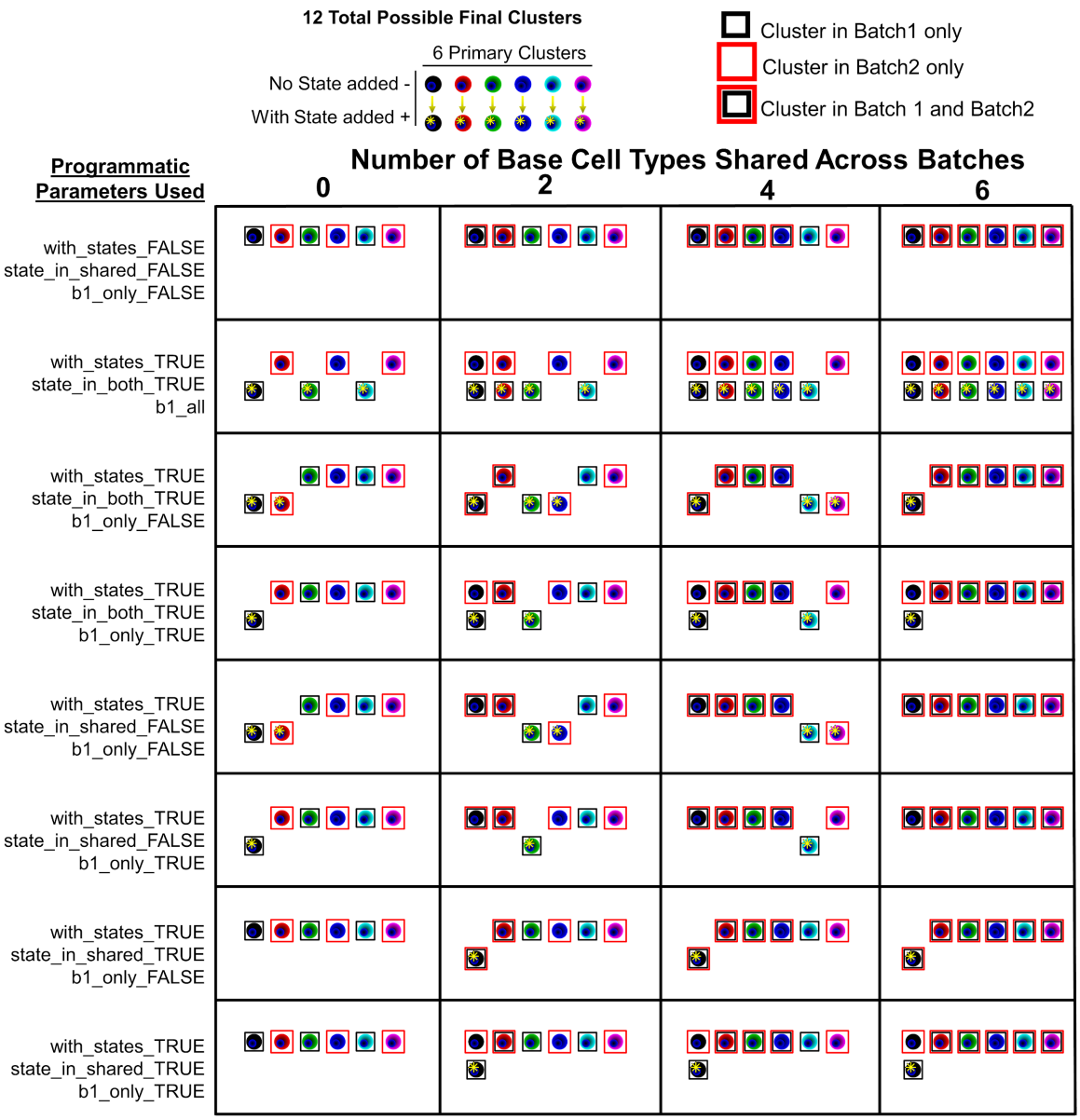
2176 presence/absence of shared clusters as in **Supplemental Fig. 7**. **b-h**, See legend for

2177 **Supplemental Fig. 7**.

2178 **Supplemental Note-5: Simulations of mixed expression programs**

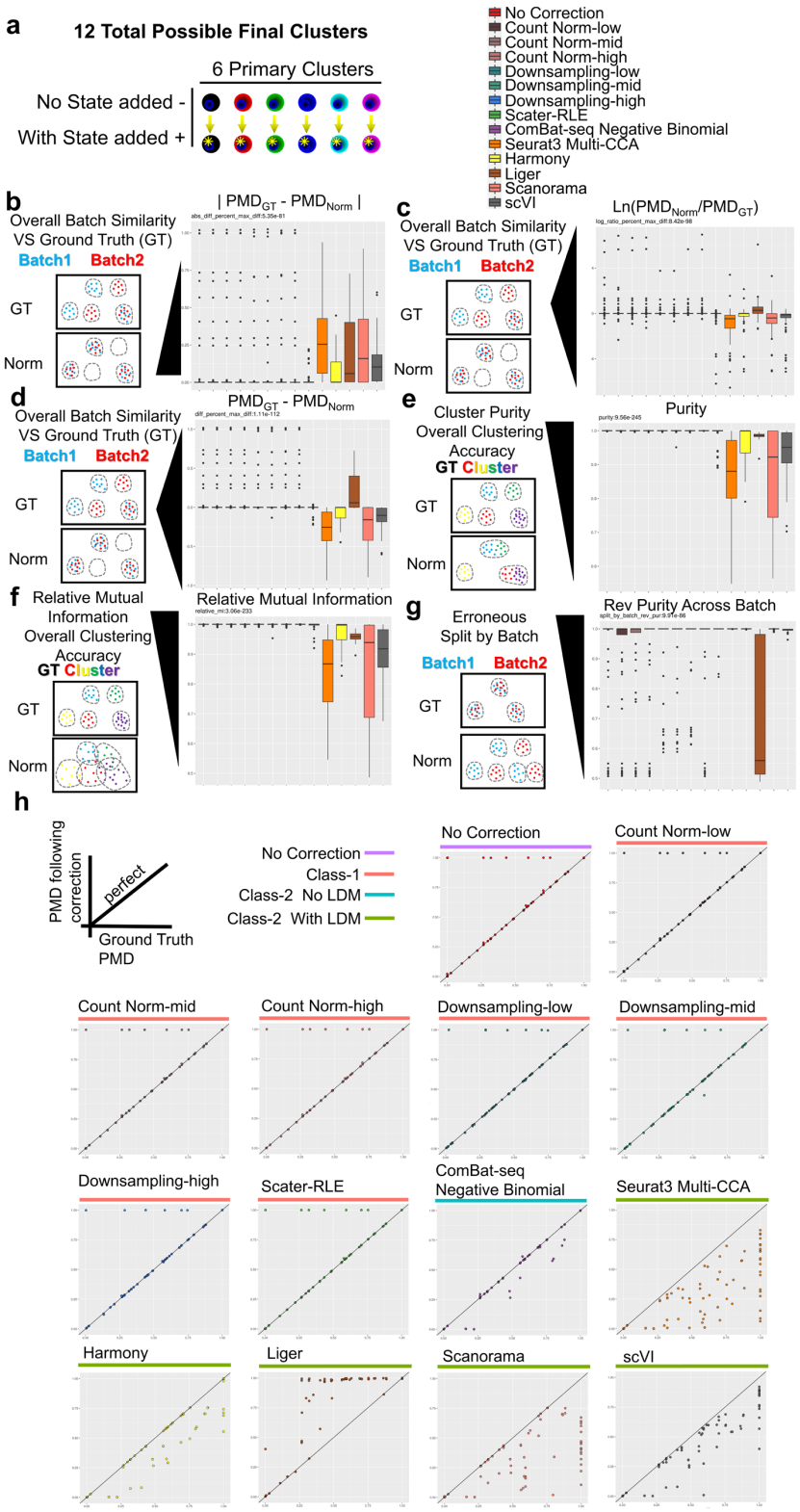
2179 We simulated a mixture-model of negative binomial distributed gene expression
2180 profiles,[50, 72] to test the effect of batch confounded gene-expression program mixtures, where
2181 batches differ in sequencing depth and are variably confounded with biology (See **Methods** for
2182 details).

2183 We simulated six cell-types, with a cell-state optionally added, some of which appear in
2184 a batch confounded manner (**Fig. 6a**; **Supplemental Fig. 9**). Overall, we simulated 8 different
2185 scenarios in which cell-states were added in different arrangements across batches, in
2186 combinations with varying levels of clusters appearing in both batches, or in a batch specific
2187 manner; this totaled 32 simulation scenarios, each simulated in triplicate (96 batch integration
2188 tasks total) (**Supplemental Fig. 9**). To analyze these simulations, we employed 9 different
2189 quantitative metrics including those that measure global cluster accuracy after correction, the
2190 ability of algorithms to successfully remove the depth aspect of the batch effect, and the efficacy
2191 of algorithms in maintaining multiple sources of biologic variation, including when confounded by
2192 batch (**Box 1f**; **ED Table 4**).



Supplemental Figure 9: All simulated scenarios. We simulated 8 different parameter combinations (rows), with differing degrees of overlap in base cell-type presence in both or a single batch (columns). In all combinations, this generated 32 simulation combinations that were repeated in triplicate for 96 different scRNAseq batch-correction and integration tasks; note that some of these scenarios converged to similar schematics. The noted simulation parameters are also reported in **ED Table 4**.

2201 In general, Class-1 algorithms tended to provide more accurate results using measures
2202 of $PMD_{observed}$ distance to $PMD_{ground\text{-}truth}$ (**Supplemental Fig. 10a-f**) and using cluster purity, and
2203 relative mutual information. Liger however was a Class-2 algorithm outlier. It appeared to have
2204 a, threshold of ground-truth PMD (PMD≈.25-.30) at which point its behavior changed; when
2205 batches were more different than this threshold, they would not be aligned (resultant "corrected"
2206 PMD≈1), explaining its periodically different behavior compared to other Class-2 algorithms
2207 (**Supplemental Fig. 10h**).
2208 We next sought to directly answer the question: do integration methods inappropriately
2209 erase batch confounded cell-state information? To this end, we included simulations that had a
2210 base cell-type (with no cell-state added) in batch-1, while in batch-2, we simulated the same
2211 base cell-type, but with an added cell-state. With clusters identified as the mixture of all
2212 expression programs (cell-lineage, cell-type, cell-state), these two clusters should be identified
2213 as separate. We therefore quantified the purity of these populations, to measure if clusters are
2214 inappropriately merged relative to the ground-truth. Indeed, Class-1 normalization approaches
2215 showed high purity in this context, while Class-2+LDM approaches did not, indicating that
2216 normalization approaches preserve cell-state differences across batch (F=205.9, *P*=2.49e-78
2217 main effects; *P*<1e-7 in all post-hoc comparisons; 1-way ANOVA/TukeyHSD; **Fig. 6b**). Similarly,
2218 if the largest source of variation (arbitrarily called cell-type variation) was different, yet shared
2219 the smaller source of variation (cell-state), these populations were often also merged by class-2
2220 algorithms (F=73.95, *P*=8.31e-38 main effects; *P*<1e-7 in all post-hoc comparisons; 1-way
2221 ANOVA/TukeyHSD; **Fig. 6c**).



Supplemental Figure 10: General accuracy of methods and efficacy in batch effect in simulations. a, Twelve possible clusters were used in simulations; they were generated from

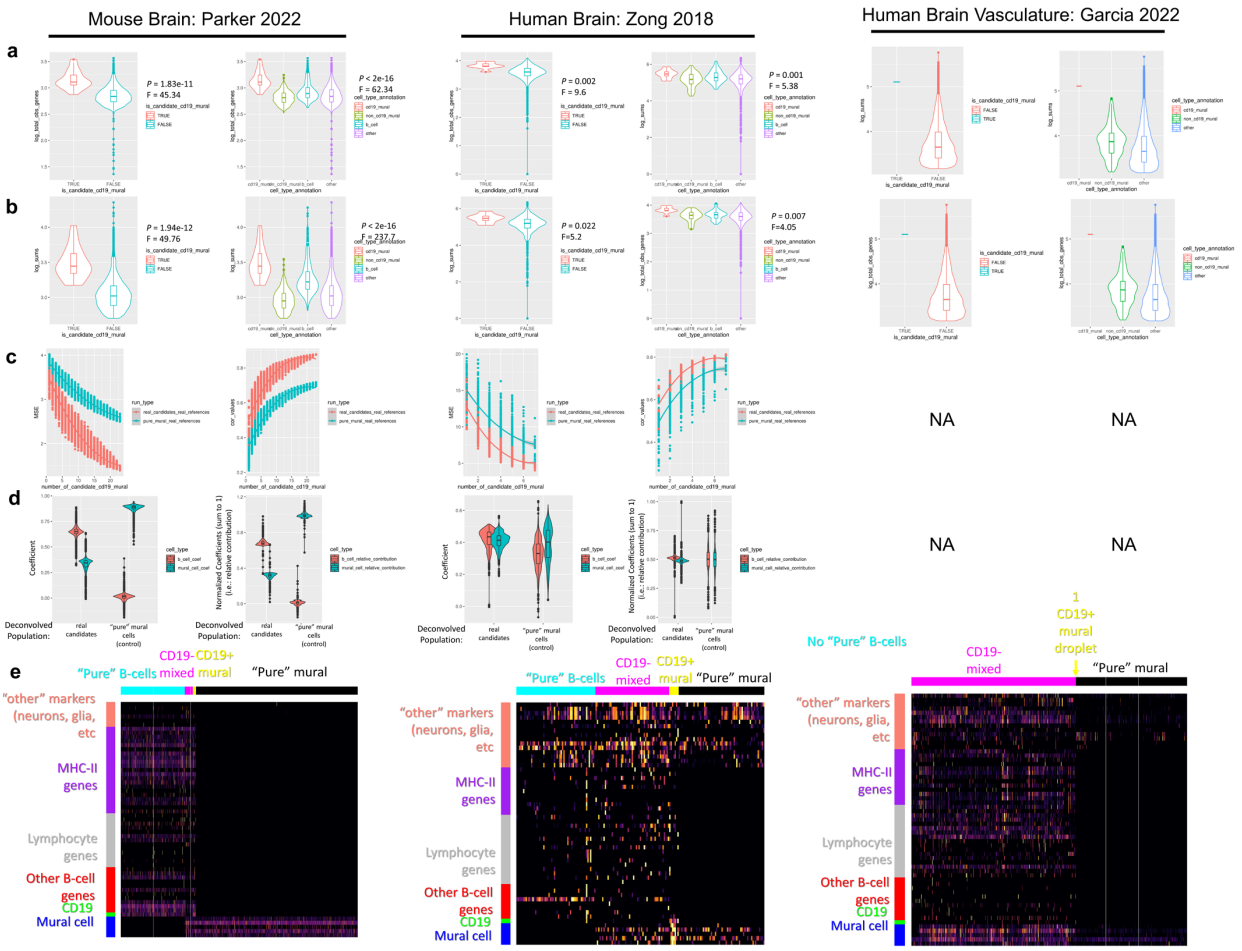
2225 six primary clusters, with a cell-state optionally added (**Supplemental Fig. 6**). **b-g,** Boxplots and
2226 schematics illustrating the possible failure-mode are shown for six metrics. Boxplots are the
2227 typical display of median, with box around second-third quartiles, whiskers to the first and fourth
2228 quartiles, with outliers defined as 1.5x the inner quartile range. **b**, The absolute difference in
2229 ground-truth (GT) observed PMD indicates how different the dataset similarity is after batch-
2230 correction compared to the known ground-truth. **c**, Similar to **b**, but the log-ratio PMD
2231 $\ln(PMD_{observed} / PMD_{ground\text{-}truth})$. Higher indicates that an algorithm tends to leave residual batch
2232 effects (batches appear more different than they are in ground-truth). Lower indicates that an
2233 algorithm tends to make batches more similar than they are in the ground-truth. **d**, Difference
2234 between the ground-truth and observed PMD ($PMD_{ground\text{-}truth} - PMD_{observed}$); interpretation is the
2235 same as **c**. **e**, Purity measures the quality of clustering results, penalizing when separate
2236 ground-truth clusters are erroneously merged together. **f**, Mutual information quantifies overall
2237 clustering accuracy using ground-truth known cluster identities as reference. **g**, Reverse purity
2238 quantifies cluster splitting; when subset for clusters that only appear in both batches, this metric
2239 penalizes cluster splitting specifically attributable to the batch effect. **h**, Scatterplots showing
2240 ground truth PMD (x-axis) against the observed PMD following application of the noted
2241 correction method.
2242

2243 ***Supplemental Note-6: Analysis to determine if droplets positive for CD19 and mural cell***
2244 ***markers were actually multiplets or multi-lineage contaminated droplets***

2245 Given that human brain samples are obtained at variable time intervals post-mortem,
2246 cell-death related signal, and cell-fragments/-debris may be present. In mice however, samples
2247 are perfused and rapidly harvested, reducing cell death, fragments and the potential for ambient
2248 RNA. We sought to determine whether the candidate CD19+/mural-like droplets show
2249 characteristics of being doublets or harboring across lineage cell fragments. Indeed, in all
2250 datasets, CD19-mural-cell-marker co-positive droplets exhibited both a higher number of
2251 observed genes and total counts (**Supplemental Fig. 11a,b**), metrics previously shown as
2252 indicative of doublets[54]. Moreover, the transcriptomes of these candidate cells were more
2253 correlated with a 50:50 transcriptome mixture model of B-cells and "pure" mural cells, compared
2254 to the mural-droplets that were negative for *Cd19* expression; this 50-50 mixture was also
2255 supported by the mean squared error (MSE) as the best fitting model. Note that these
2256 correlations and MSE calculations were performed holding out the markers that were used to
2257 identify those populations, thus A) preventing data leakage and self-fulfilling prophecy, and B)
2258 confirming that the B-like properties of the CD19+ mural-like droplets were *globally* more similar
2259 to B-cell transcriptomes rather than only being positive for CD19.

2260 To directly measure the relative contribution of the "pure" mural and B-cell-like droplet
2261 transcriptomes to explain the candidate CD19/mural population, we fit a linear mixed model of
2262 the average log(count+1) transcriptomes of the reference "pure B-cell" (negative for mural
2263 markers) and "pure mural cell" transcriptomes. In the mouse dataset, this model clearly shows
2264 that, the candidate CD19/mural-cells fit best with a ~70% contribution from B-cells and 30%
2265 contribution from mural cell transcriptomes. Our methodologic positive control, fitting "pure"
2266 mural cells (as a held-out population not used for reference construction), worked as expected
2267 with ~0% contribution from B-cells and ~100% contribution from mural cells in the mouse
2268 dataset (**Supplemental Fig. 11d**).

2269 Notably however, when we performed this same analysis on the Zhong 2018 human
2270 dataset originally used to identify the putative CD19+/mural-cells, we observed similar
2271 coefficients between the candidate populations as well as when fitting a held out selection of the
2272 "pure" population of mural cells. A possible reason for this could be that that even the "pure"
2273 populations of mural cells (CD19 negative, B-cell marker negative, mural marker positive) were
2274 themselves also mixed in origin (**Supplemental Fig. 11d**). Fittingly, when we examined marker
2275 genes for other cell-types, we found that essentially all observed droplets were contaminated
2276 with marker genes across many lineages, indicating the presence of highly contaminated
2277 droplets in the original dataset (**Supplemental Fig. 11e**).
2278 We further investigated the pathways of genes that were significantly correlated with
2279 CD19. Interestingly, using the full dataset, CD19 was correlated with genes enriched for both
2280 lymphocytes/immune system related pathways (-log10(p)=12.72; **ED Table 5a**) and
2281 vasculature/tube morphogenesis pathways (-log10(p)=12.97,9.07; **ED Table 5a**); however,
2282 when "pure" B-cells were removed from the dataset (Zong 2018[73]), CD19 *remained* correlated
2283 with the "immune system" (-log10(p)=8.05; **ED Table 5a**), indicating that from a global
2284 perspective, CD19 transcripts *continued* to track with the co-occurrence of other immune-related
2285 transcripts, fitting with a model of cell-fragment contaminations. When the candidate
2286 CD19+/mural-droplets were removed, the CD19 correlation with vascular and tube-
2287 morphogenesis was no longer significant, switching the CD19 correlations primarily to those
2288 correlated with immune related pathways, similar to those found with CD79A/B correlations (**ED**
2289 **Table 5b**), further supporting the hypothesis that this small population was indeed a mixture
2290 population.
2291 Lastly, examination of a recently released human brain vasculature atlas (Garcia 2022)
2292 showed only a single droplet that was CD19+/mural-marker co-positive droplet (**Supplemental**
2293 **Fig. 11a-e**)[43].
2294



Supplemental Figure 11: Droplets positive for CD19 and mural cell markers in the Zong 2018 dataset are consistent with cell-type mixtures rather than a novel population of mural cells. a, Violin plots showing the total number of observed genes in the "candidate" CD19+/mural-cell-marker+ double positive droplets and either all other cells, or the relevant possible mixture sources (B-cells, mural cells, and others). **b**, Similar to (**a**), violin plots showing the total number of counts rather than number of observed genes. **c**, Assessment of mean squared error (MSE) and Spearman correlation values determining how well a 50:50 mixture model of the average "pure" B-cell and "pure" mural cell transcriptomes fit with either the real candidates or a subset of "pure" mural cells (which were not included in the reference mixture model) shows that a 50:50 mixture model of the full transcriptome fits the data best (with the exclusion of CD19 and other markers used to identify these populations, to prevent data

2307 leakage), supporting that the candidate CD19+/mural cells are a global transcriptome mixture of
2308 mural and B-cells as opposed to a population of “pure” mural cells. Note where there were no
2309 “pure” B-cell-like droplets identified making a mixed model impossible, indicated by “NA.” **d**,
2310 Violin plots showing raw and normalized coefficients fitting a linear mixed model to “pure” B- and
2311 mural-cell references. Candidate CD19+ mural cells and methodologic positive controls (“pure”
2312 mural droplets) were held out from references to prevent data leakage. In the Zong 2018
2313 dataset, *all* populations were consistent with a transcriptome mixture, suggesting broad
2314 contaminations rather than strict doublets, and that even the reference “pure” populations were
2315 not strictly lineage-pure. **e**, Heatmaps of lineage marker genes in the Zong 2018 human dataset
2316 confirm that nearly all droplets are lineage contaminated, while the mouse dataset appeared
2317 comparatively non-mixed. This fits with the mixture model results (**d**), in which the mouse
2318 mixture model had clearly distinct “pure” references, where the reference “pure” mural cells fit
2319 100% with other mural cells, and 0% with B-cells, whereas in the Zong 2018 dataset, all cells fit
2320 well with a mixture model due to the transcriptomes of each lineage already being mixed. The
2321 human brain vasculature atlas did not contain a convincing CD19+/mural-droplet population (1
2322 droplet).

2323