Segmentation aware probabilistic phenotyping of single-cell spatial protein expression data

Yuju Lee[1,2], Edward L. Y. Chen[1], Darren C. H. Chan[1,3], Anuroopa Dinesh[1], Somaieh Afiuni-Zadeh[1], Conor Klamann[4], Alina Selega[1,5], Miralem Mrkonjic[6], Hartland W. Jackson[1,3,7,✉], Kieran R. Campbell[1,2,3,5,7,8,✉]

[1] Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto ON M5G 1X5, Canada
[2] Department of Computer Science, University of Toronto, Toronto ON M5S 1A8, Canada
[3] Department of Molecular Genetics, University of Toronto, Toronto ON M5T 3A1, Canada
[4] Data Sciences Institute, University of Toronto, Toronto ON M5G 1Z5, Canada
[5] Vector Institute, Toronto, ON M5G 1M1, Canada
[6] Department of Laboratory Medicine & Pathobiology, University of Toronto, Toronto ON M5S 1A8, Canada
[7] Ontario Institute of Cancer Research, Toronto, ON M5G 1M1, Canada
[8] Department of Statistical Sciences, University of Toronto, Toronto ON M5S 3G3, Canada

✉ hjackson@lunenfeld.ca, kierancampbell@lunenfeld.ca

Abstract

Spatial protein expression technologies can map cellular content and organization by simultaneously quantifying the expression of >40 proteins at subcellular resolution within intact tissue sections and cell lines. However, necessary image segmentation to single cells is challenging and error prone, easily confounding the interpretation of cellular phenotypes and cell clusters. To address these limitations, we present STARLING, a novel probabilistic machine learning model designed to quantify cell populations from spatial protein expression data while accounting for segmentation errors. To evaluate performance we developed a comprehensive benchmarking workflow by generating highly multiplexed imaging data of cell line pellet standards with controlled cell content and marker expression and additionally established a novel score to quantify the biological plausibility of discovered cellular phenotypes on patient-derived tissue sections. Moreover, we generate spatial expression data of the human tonsil – a densely packed tissue prone to segmentation errors – and demonstrate cellular states captured by STARLING identify known cell types not visible with other methods and enable quantification of intra- and inter- individual heterogeneity. STARLING is available at https://github.com/camlab-bioml/starling.

Introduction

Highly multiplexed imaging technologies such as Imaging Mass Cytometry (IMC)[1] are a set of state-of-the-art assays that allow simultaneous quantification of up to approximately 40 protein targets at sub-cellular resolution in two dimensions[2]. This enables delineation of high dimensional single-cell phenotypes and cellular communities in both archival patient tissue sections as well as tissues from animal models and cell lines. Together, their applications have uncovered cellular phenotypes and spatial architectures that stratify patient outcomes and predict disease course across a range of pathologies including breast[3,4], lung[5], melanoma[6], and brain[7] cancers, along with other pathologies such as diabetes[8].

A major step in the analysis of multiplexed image cytometry is the process of cell segmentation that identifies where in the original (pixel-level) data cells are. Commonly, the data is subsequently summarized to the single-cell level by taking the mean expression of markers over all pixels belonging to a given cell along with cell size and/or morphology information and applying standard data analysis pipelines developed for suspension cytometry such as visualization[9], clustering[10,11], and cell type assignment[12]. For the process of segmentation, multiple algorithms have been proposed ranging from hand-crafted machine learning pipelines[3] to deep learning methods including CellPose[13] and Mesmer[14].

However, it is increasingly being recognized that cell segmentation is an inherently error-prone process due to resolution limitations, three-dimensional tissue architecture with cells overlapping in the z-plane, and algorithmic imperfections. These issues result in signal spillover between adjacent cells[15,16], over-segmentation resulting in doublets, and cellular projections being mis-assigned[17]. Upon summarization to the single-cell level, this results in cellular phenotypes that are frequently the average of two or more actual cells, which in turn leads to cells or clusters that co-express implausible protein expression combinations. For example, CD3 and CD20 should be exclusively expressed on T and B lymphocytes respectively[18], but cell populations that co-express them are frequently found in imaging mass cytometry datasets[3,4].

Despite these challenges, few computational methods for highly multiplexed imaging address these limitations. Classical clustering methods for suspension CyTOF frequently re-used for multiplexed imaging[3,4] such as PhenoGraph[11] and FlowSOM[10] all assume pure single cell expression as a starting point. Multiple strategies have attempted to augment resolution to improve segmentation, but segmentation errors occur even in high resolution imaging

methods[19,20]. CellSighter[17] is a recently developed method for multiplexed imaging that attempts to circumvent segmentation errors by operating on the raw imaging data with data augmentation, but is used for supervised classification of cell types given a labeled training set rather than *de novo* discovery of cell phenotypes as is standard in a clustering workflow. Conversely, doublet detection and removal have received significant attention in single-cell transcriptomics, with multiple methods[21–24] and benchmarking works available[25]. To-date, no methods have been proposed to learn denoised cellular phenotypes while accounting for cell segmentation errors in highly multiplexed imaging.

To help tackle this, we present STARLING (SegmenTation AwaRe cLusterING), a novel probabilistic clustering method specifically designed to account for the fact that single-cell expression profiles from multiplexed imaging may represent the combination of two or more actual cell populations. We introduce the concept of a plausibility score that quantifies the quality of a given clustering to infer biologically meaningful cellular phenotypes and introduce a further benchmarking framework to quantify the ability of methods to detect segmentation errors from multiplexed imaging. We further generate IMC data of cell pellets of controlled cell lines to quantify the ability of different methods to re-infer cellular clusters when the input cell types are both controlled and known. Finally, we generate IMC data from 16 regions of interest (ROIs) from the human tonsil across 3 donors — a secondary lymphoid organ densely packed with immune cells — and demonstrate how the STARLING immunophenotypes detect spatial single-cell communities and highlight intra- and inter-donor variability. An open source Python package implementing the STARLING method is available at https://github.com/camlab-bioml/starling.

Results

STARLING: a probabilistic model for segmentation error aware clustering of highly multiplexed imaging data

Multiple artifacts give rise to segmentation errors in multiplexed imaging data, including doublets where a segment covers two or more cells, spatial spillover, 3D tissue effects, and cellular projections[16,17]. The consequence is that when the error involves cells from different cell types — previously described as heterotypic doublets[24] — clusters are formed due to the composition of different cells within a segment that are not “real” cell types in the source biological material (**Fig. 1A**). This can be shown by clustering data from multiple recent publications giving rise to

cell types that co-express biologically implausible marker combinations (**S. Fig. 1**). The consequence is that outside the few scenarios where implausible clusters may be manually identified, it is often difficult to trust the cellular phenotypes identified from such experiments without follow-up validation, and even more so trust every cell label within the thousands or millions of cells in a dataset. At worst, novel cell types or spatial features may be identified that are in fact purely due to segmentation artifacts.

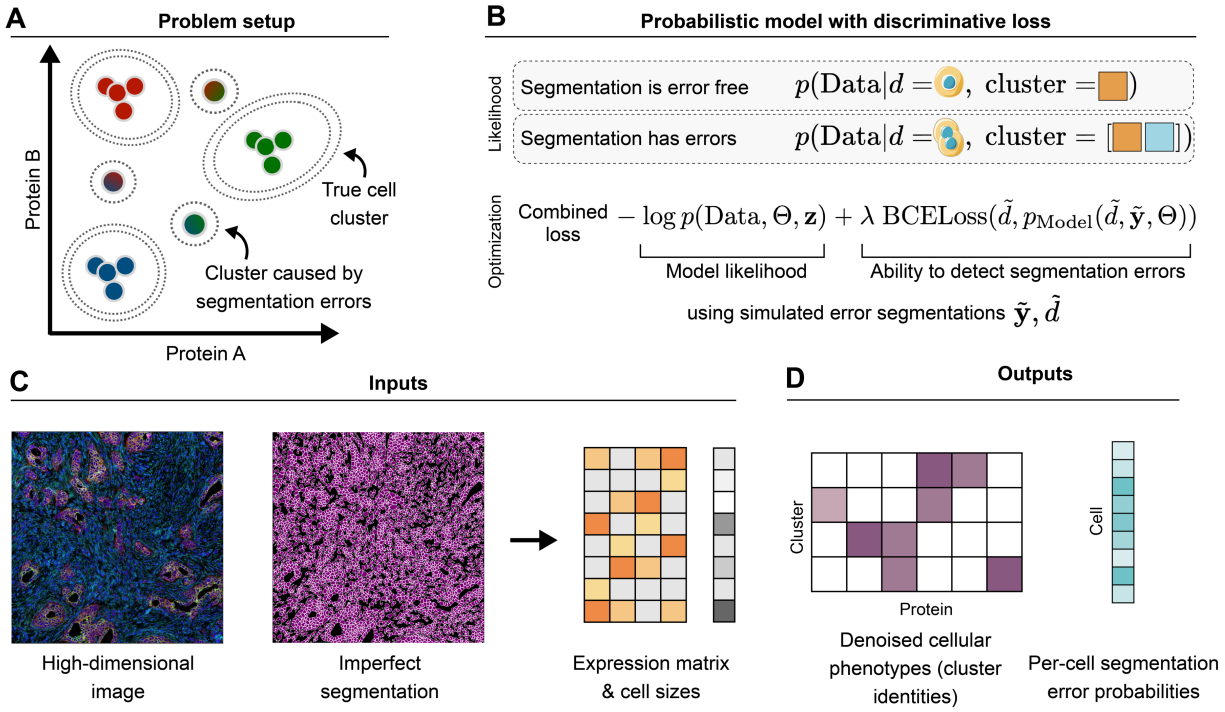

Figure 1 A Segmentation errors in highly multiplexed imaging experiments induce clusters formed by the composition of multiple true cell types. **B** Our probabilistic model STARLING for clustering highly multiplexed imaging data models whether each cell is observed segmentation error free, and if not models the composition of cell types leading to the observed cell. The loss function jointly maximies the likelihood of the data and the ability to discriminate on-the-fly simulated segmentation errors. **C** The input to STARLING includes a high dimensional multiplexed image and imperfect segmentation that is summarized to an expression matrix with cell sizes. **D** The output includes both the denoised cell types along with a per-cell estimate of segmentation errors.

To help address this, we developed STARLING (**S**egmen**T**ation **A**waRe c**L**uster**ING**), a novel probabilistic clustering method to identify true cell phenotypes in the presence of segmentation errors. STARLING specifically models a per-cell probability of a segmentation error, along with the true cluster identities of the underlying cells segmented together as part of the error. The joint likelihood of the data, cluster identities and assignments, and segmentation error probabilities are then optimized along with a discriminative loss that regularizes the model to be able to both accurately predict errors and orient which clusters are likely true compared to error-induced (**Fig. 1B** and **Methods**). As input, STARLING requires only the high-dimensional image

107 along with imperfect cell segmentation that are summarized to a per-cell expression profile
108 along with physical cell size in pixels (**Fig. 1C**) as is commonly output by a range of pipelines for
109 highly multiplexed imaging data[26,27]. After optimization of the loss, STARLING outputs both the
110 denoised cluster identifies representing the cellular phenotypes of the underlying cells in the
111 input biological material, as well as a per-cell segmentation error probability that may be useful
112 for follow-up visualization and analysis (**Fig. 1D**).

113 Scoring the biological plausibility of clusters inferred from multiplexed images

114 We first sought to validate that the clusters returned by STARLING showed more biologically
115 realistic cellular phenotypes than existing strategies. Given true cell types are *a priori* unknown
116 and benchmarking using simulated data can introduce unrealistic assumptions, we built upon
117 previous work[16,28] that uses the co-expression of marker proteins to evaluate the quality of the
118 inferred cell clusters. Specifically, we generalized this idea to create the *plausibility score* (**Fig.
119 2A**) that takes an *a priori* known set of protein pairs that either exhibit mutually exclusive
120 expression or conditional co-expression. For example, the T lymphocyte marker CD3 should
121 never be co-expressed with the B lymphocyte marker CD20; conversely, CD3 should never be
122 expressed in the absence of the pan-leukocyte marker CD45. By curating a list like this for each
123 dataset's antibody/antigen panel, for a given clustering we can simply count the proportion of
124 centroids that fall in the expected regions using a given threshold, and designate a normalized
125 version of this as the plausibility score (**Methods**).

126 To validate this plausibility score we exploited the fact that cells isolated with no neighbours
127 should have fewer segmentation errors than those in densely packed regions with many
128 neighbours. Using three common clustering algorithms (FlowSOM[10], PhenoGraph[11], Kmeans)
129 we clustered three multiplexed imaging datasets — two published breast cancer atlases[3,4] and a
130 human tonsil dataset generated for this study (see **Methods**) — separately for cells that have
131 neighbours and those that do not. We then computed the plausibility score for each set of
132 clustering using marker pairs derived for each dataset's antibody panel (**S. Table 1**). Across all
133 datasets and clustering algorithms, the plausibility score is significantly higher for cells with no
134 neighbours than those with neighbours (**Fig. 2B**), validating its utility as an independent
135 measure of biologically realistic cellular clusters from highly multiplexed imaging data.

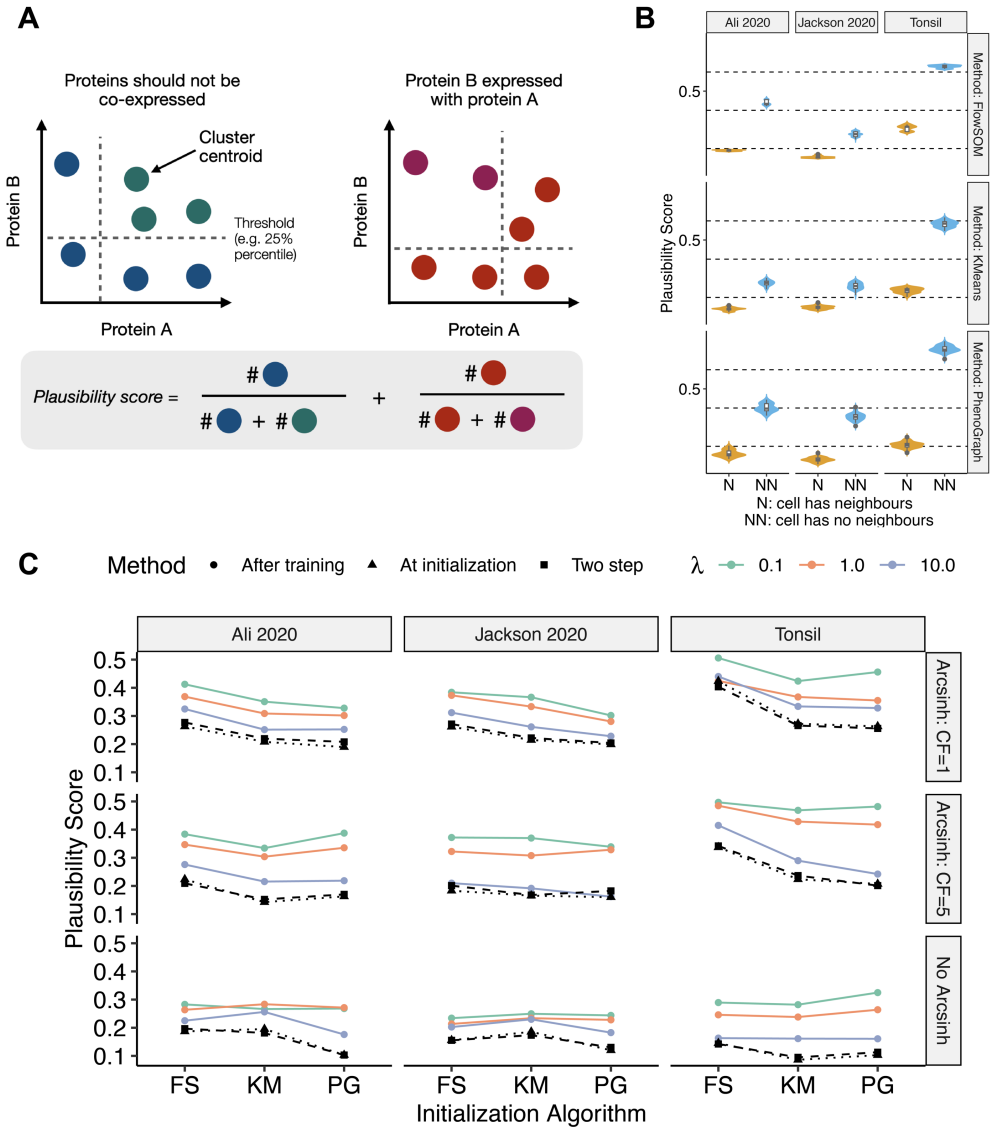

Figure 2 A Pairs of proteins known to not be co-expressed (left) or show conditional expression (right) are collated for each dataset, and the *plausibility score* defined as the proportion of centroids that lie in the expected regions under some threshold of expression. **B** To validate the plausibility score on real data, we took three IMC datasets and separated cells into those that have neighbours (N) or no neighbours (NN). We clustered each with three clustering algorithms and computed the plausibility score on the centroids. Cells with no neighbours and therefore likely fewer segmentation errors show significantly higher plausiblity scores than those with neighbours. **C** The plausibility score quantified across three datasets, three normalization scenarios, and three hyperparameter settings. The black triangles represent the score at initialization using three popular clustering algorithms (FS: FlowSOM, KM: K-means; PG: PhenoGraph) and coloured lines after training. The black squares represent the two step approach of first removing likely doublets.

Next, we benchmarked the plausibility score of clusters found by STARLING compared to the three established clustering algorithms on the three highly multiplexed imaging datasets. To comprehensively disentangle the factors associated with benchmarking performance, we

initialized STARLING using each of the three clustering algorithms, and compared the initial plausibility score (equivalent to clustering using the existing methods) to the post-STARLING clustering plausibility score. In addition, we varied the λ hyperparameter in STARLING that controls the trade-off between optimizing the model likelihood and ability to detect synthetic segmentation errors as well as different data pre-processing strategies (via archsinh normalization) commonly used in IMC data analysis[12]. We also considered a two-step approach where first a random forest model is trained to detect doublets and cells with a high probability of being a doublet are removed before the data is clustered using the baseline methods (**Methods**). The results in **Fig. 2C** demonstrates that for all clustering algorithms, hyperparameter combinations, and data normalization strategies, STARLING outperforms both the baseline initializations as well as the two-step doublet removal strategy.

Assessing the ability of STARLING to detect cell segmentation errors

Next, to assess whether segmentation errors quantified by STARLING are reliable we adopted a multi-faceted strategy. First, we exploited the fact that MESMER[14] — a state of the art segmentation algorithm for multiplexed imaging — can perform both whole-cell and nuclear segmentation. We segmented each image in each dataset twice and designated cells where more than one nuclei was found in the nuclear-only segmentation despite only a single cell being found in the whole-cell segmentation as likely to contain a segmentation error. We then quantified STARLING's segmentation error probability for cells likely to contain a segmentation error compared to otherwise. The results in **Fig. 3A** shows a clear increase in segmentation error probability for cells whose nuclear segmentation conflicts with the whole-cell segmentation. Note that given the multiple ways a segmentation error can occur beyond doublets we do not expect the segmentation error probability to be zero in cells where the nuclear segmentation matches the whole-cell segmentation.

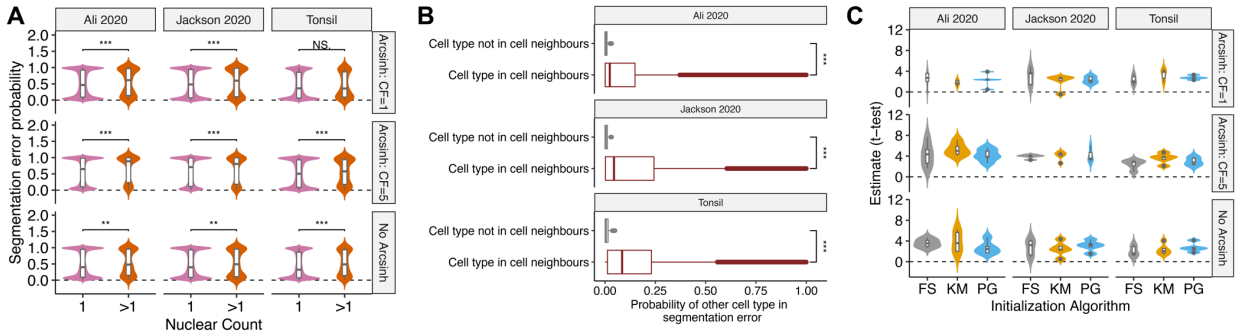

Figure 3 A. The segmentation error probability computed by STARLING based on whether each whole-cell segmented by MESMER contains one or more segmented nuclei across three datasets and three normalization schemes. **B**. The probability that the mix of cell types identified by STARLING to be involved in a segmentation error is significantly higher when those cell types involve the cell neighbours. **C**. A T-statistic comparing the convex area to total area of each cell for cells predicted to contain segmentation errors by STARLING vs those predicted to be segmentation error free.

Secondly, we quantified the ability of STARLING to detect spatial spillover despite not explicitly being given spatial information about the cells beyond size. Specifically, for each cell across the three datasets we identified the most likely cluster in the absence of segmentation errors as well as the most likely cluster of each neighbouring cell. Next, we took STARLING's per-cell Γ matrix that quantifies the probabilities of all possible pairings of clusters that contributed to the segmentation errors involved in a given cell (**Methods**). For each cell, we quantified the value of this matrix for the pair of clusters corresponding to that cell and its neighbours and contrasted this with the pair of clusters corresponding to that cell and all clusters not present in the cell's neighbours. We found significantly higher probability of a cluster contributing to a cell's segmentation error when a neighbouring cell was of that cluster type than otherwise (**Fig. 3B**). This implies STARLING implicitly detects spatial spillover artifacts despite not being explicitly provided information about neighbouring cells or their expression profiles.

Finally, we reasoned that cells likely to involve segmentation errors would have altered morphology. Cell segments corresponding to doublets (i.e. two cells segmented into one) or with cellular projections will be more likely to have a concave morphology where the ratio of convex area to total area is increased compared to cells with high quality segmentations (**S. Fig. 2**). Therefore, across the three datasets, normalization strategies, and initialization algorithms, we separated cells predicted to involve segmentation errors from those predicted to be error-free and computed the per-cell ratio of convex to total cell area for each. Despite STARLING not having access to information on cellular morphology, contrasting these two values using a t-test demonstrates a significantly increased convex to total area ratio in cells predicted to have segmentation errors (**Fig. 3C**). Overall these results imply that in addition to the cellular

187 phenotypes quantified by STARLING being biologically plausible, the segmentation error
188 probabilities represent a robust measure of segmentation quality that may serve for useful
189 downstream analysis and interpretation.

190 Quantifying cell phenotype inference using controlled cell types

191 Next, we evaluated the inference of cellular phenotypes in a controlled setting. We created a
192 cell pellet from four cell lines, each known to be positive for only one major lineage marker
193 (CD3+ Jurkat, CD20+ RAMOS, CD68+ THP1, PanCK+ BxPC3, see **Methods**). This dataset
194 acts as a control as each resulting cluster should be positive for only one of the major lineage
195 markers, with any crossover due to segmentation errors. The cell lines were mixed in equal
196 proportions, stained with an antibody panel able to discriminate different cell types (**Fig. 4A**),
197 and subjected to IMC (**Fig. 4B**) resulting in high-dimensional quantification of the resulting cell
198 lines with spatial resolution (**Fig. 4C**). After applying a standard analysis pipeline of
199 segmentation using Mesmer[14] and clustering with PhenoGraph[11], we identified 13 distinct cell
200 types showing heterogeneous expression profiles (**Fig. 4D**). Using a procedure to assign
201 clusters to cell types based on single marker positivity (**Methods**), we were able to confidently
202 assign 9/13 clusters to the cell lines of origin, with 4/13 labeled as "Unsure" due to multi-marker
203 positivity.

204 Next, we applied STARLING using the PhenoGraph clustering for initialization. Using the same
205 cell type assignment scheme as before, we were subsequently able to map 12/13 clusters as
206 belonging to a given cell type (**Fig. 4E**). Quantifying the difference from the highest most
207 expressed lineage marker to the second highest across all clusters showed an increase in
208 discrimination for all but two of the clusters (**S. Fig. 3**), further demonstrating STARLING's ability
209 to recapitulate ground truth cell types in the dataset. This finding was robust to the precise
210 threshold used to discriminate cell types (**S. Table 2**). Finally, we examined the spatial
211 distribution of imaged cell types for both PhenoGraph (**Fig. 4G**) and STARLING (**Fig. 4H**) for the
212 same region of interest as **Fig. 4C**. We noted that cells assigned under PhenoGraph as
213 "unsure" largely corresponded to those identified as having segmentation errors by STARLING.
214 To quantify this, we subsetted to all cells assigned to an unsure cluster using PhenoGraph
215 across the entire image, and examined their assignments following STARLING (**Fig. 4H**). For
216 3/4 of the initially uncertain clusters the most common STARLING assignment was to a
217 segmentation error, with segmentation errors being the second most likely assignment for the

218

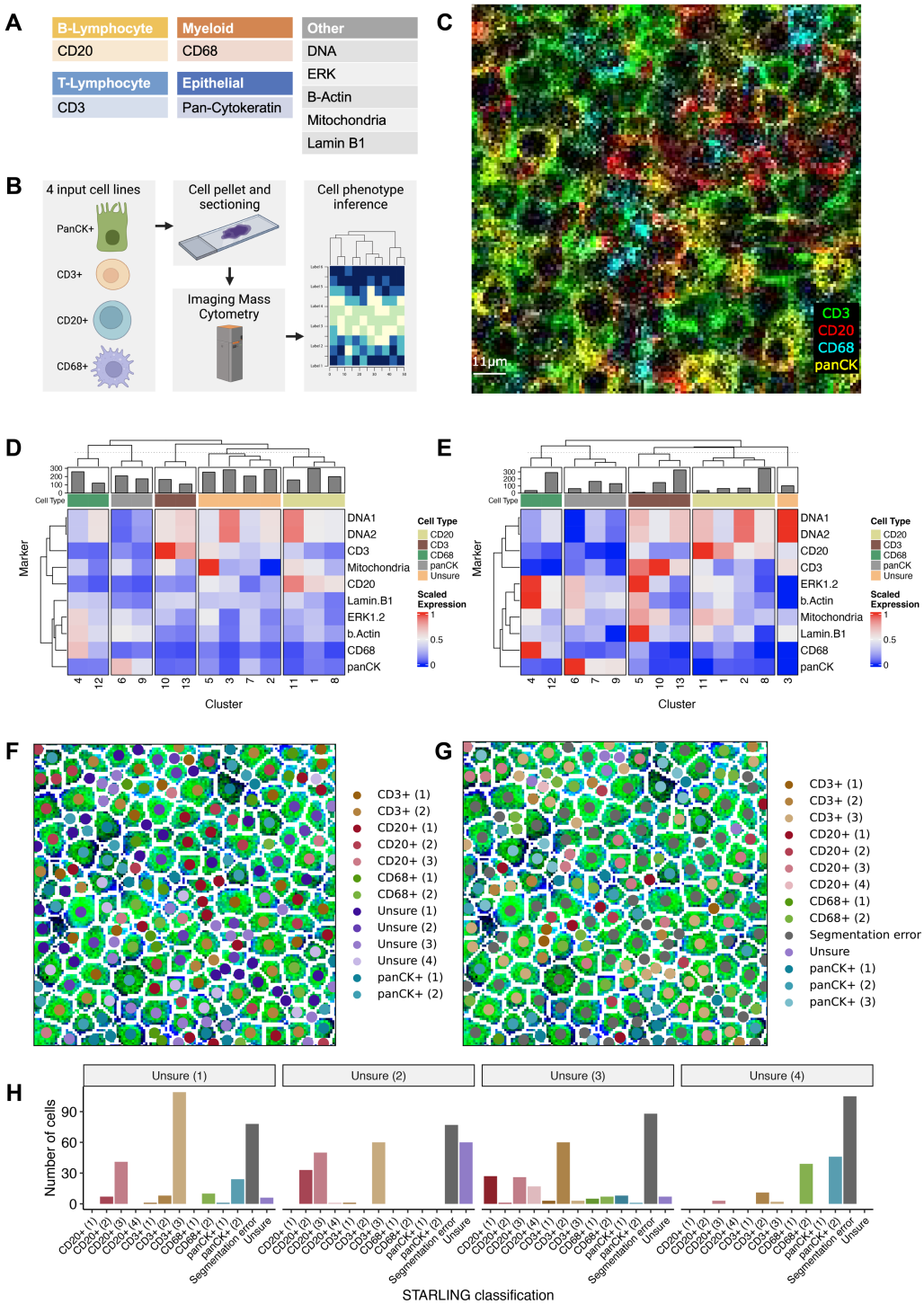

Figure 4 A Antigens targeted by the antibody panel used to profile the cell pellet data. **B** Overall workflow of IMC profiling of four cell lines to generate a controlled dataset for benchmarking. **C** Example ROI in resulting image coloured by major lineage markers. **D** Cell clusters and cell type interpretations using a standard PhenoGraph workflow. **E** Cell clusters and cell type interpretations after applying STARLING. **F** Same ROI as (**C**) coloured by PhenoGraph cell types. **G** The ROI coloured by STARLING cell type assignments. **H** The number of cells previously classified as Unsure (1-4) assigned to each cell type by STARLING.

220 remaining cluster. This demonstrates both that segmentation errors may induce their own cell
221 clusters in multiplexed imaging data and that STARLING may successfully guide users between
222 biologically “real” cellular phenotypes compared to those induced primarily by segmentation
223 errors.

224 Robustly deconvoluting tissue architecture and heterogeneity in human tonsils

225 To demonstrate the ability of STARLING to uncover insights into tissue architecture and intra-
226 and inter-sample cellular heterogeneity in densely packed tissues, we profiled over 240,000
227 cells of the human tonsil across 16 ROIs and 3 donors using an immune-focused antibody panel
228 (**Fig. 5A, S. Table 3** and **Methods**). Visualization of the resulting ROIs using canonical markers
229 clearly demonstrated known anatomical structures of the tonsil such as CD20+ germinal centers
230 and surrounding epithelium (**Fig. 5B**). Applying STARLING to the segmented single-cell
231 expression profiles uncovered a total of 25 clusters (**Fig. 5C** and **Methods**), which we
232 successfully categorized into nine distinct cell types consistent with major immune and stromal
233 lineages. This included extensive B and T cell heterogeneity consistent with the major cell
234 populations in lymphoid tissue, including naive and activated B and T cell populations, distinct
235 IFNγ+ and PD-1+ B cell populations, and multiple myeloid populations including macrophages,
236 monocytes, and neutrophils. Notably, interpretation of cell clusters was improved compared to
237 baseline methods (**S. Fig. 4**).

238 Next, we quantified how cell type abundance varies between ROIs from the same donor (intra-
239 donor heterogeneity) and between ROIs from different donors (inter-donor heterogeneity). By
240 quantifying the proportion of each cluster in each ROI (**Fig. 5D**), we found several clusters
241 enriched in a donor-specific manner. For example, Donor 1 is specifically enriched for clusters
242 2, 7, and 13, corresponding to specific populations of IFNγ+ B cells and IFNγ+ naive and
243 activated T cells. In contrast, some clusters such as 4 and 6 (corresponding to B and other
244 CD45+ cells) are specific to two ROIs from two distinct donors. However, when collapsing this
245 proportion analysis to major lineage type (**Fig. 5E**), the frequencies are largely consistent across
246 cell types showing enrichment for B and T cells as expected in the tonsil, with several
247 exceptions. This highlights the potential of STARLING coupled with highly multiplexed imaging
248 technologies to reverse out cell type heterogeneity in densely packed tissues.

249

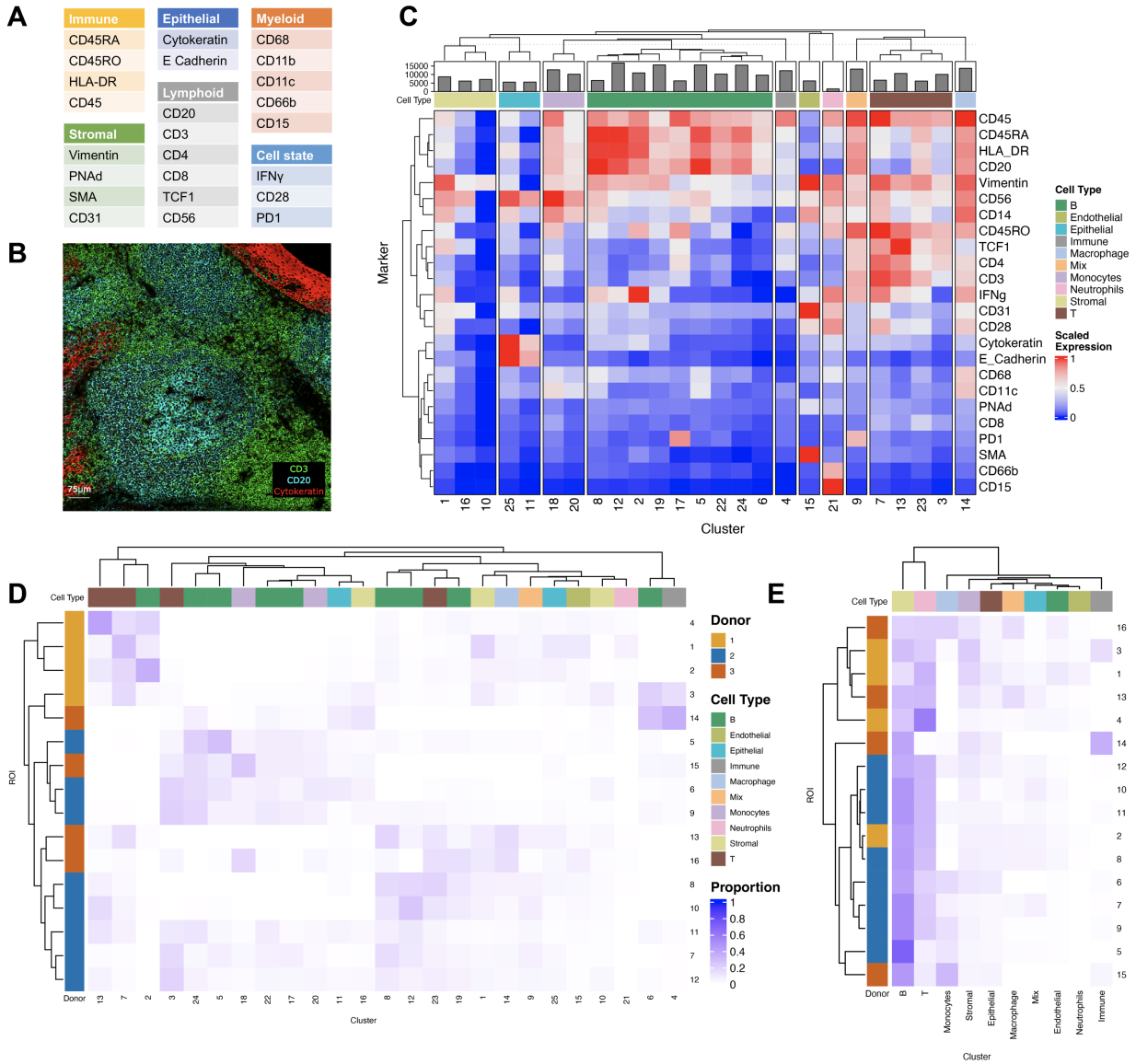

Figure 5 A. Antigens targeted in antibody panel designed for tonsil tissue. **B** Example ROI with CD3, CD20, and Cytokeratin identifies a B cell rich germinal center with peripheral T cells and tissue epithelium. **C** Mean marker expression for cluster identities discovered by running STARLING simultaneously across all 16 ROIs. **D** Cluster proportions identified for each cluster in ROI highlights intra and inter-sample heterogeneity. **E** Cluster proportions for each ROI summarized to the cell type level.

Finally, we sought to understand how the spatial architecture of the cell types present in the tonsil varied within and between donors. We implemented a spatial enrichment analysis to identify which cell clusters are closely mixed or highly excluded spatially across the images (**Methods**). While general patterns were evident consistent with the tissue of origin such as T-B exclusion due to the spatial architecture of germinal centers, we observed extensive variability in spatial enrichment both within and between donors (**Fig. 6A**). We first checked the

consistency of this spatial enrichment under two strategies for assigning cells: removing cells with a high segmentation error probability before the enrichment analysis, or using STARLING's maximum likelihood estimate of the cell cluster assuming no segmentation error and therefore including all cells. We found the two estimates to be highly correlated (**Fig. 6B**, $\rho=0.977$, $p<0.001$), demonstrating we can robustly quantify spatial enrichment even in the presence of segmentation errors. Finally, we explored specific examples of this enrichment. For example, ROI 1 (**Fig. 6C** left) is identified as an outlier for neutrophil infiltration into the stroma compared to all other ROIs as validated by specific marker expression (**S. Fig. 5**). Conversely, ROI 11 (**Fig. 6C** right) is identified as having the most extreme compartmentalization between T and B cells.

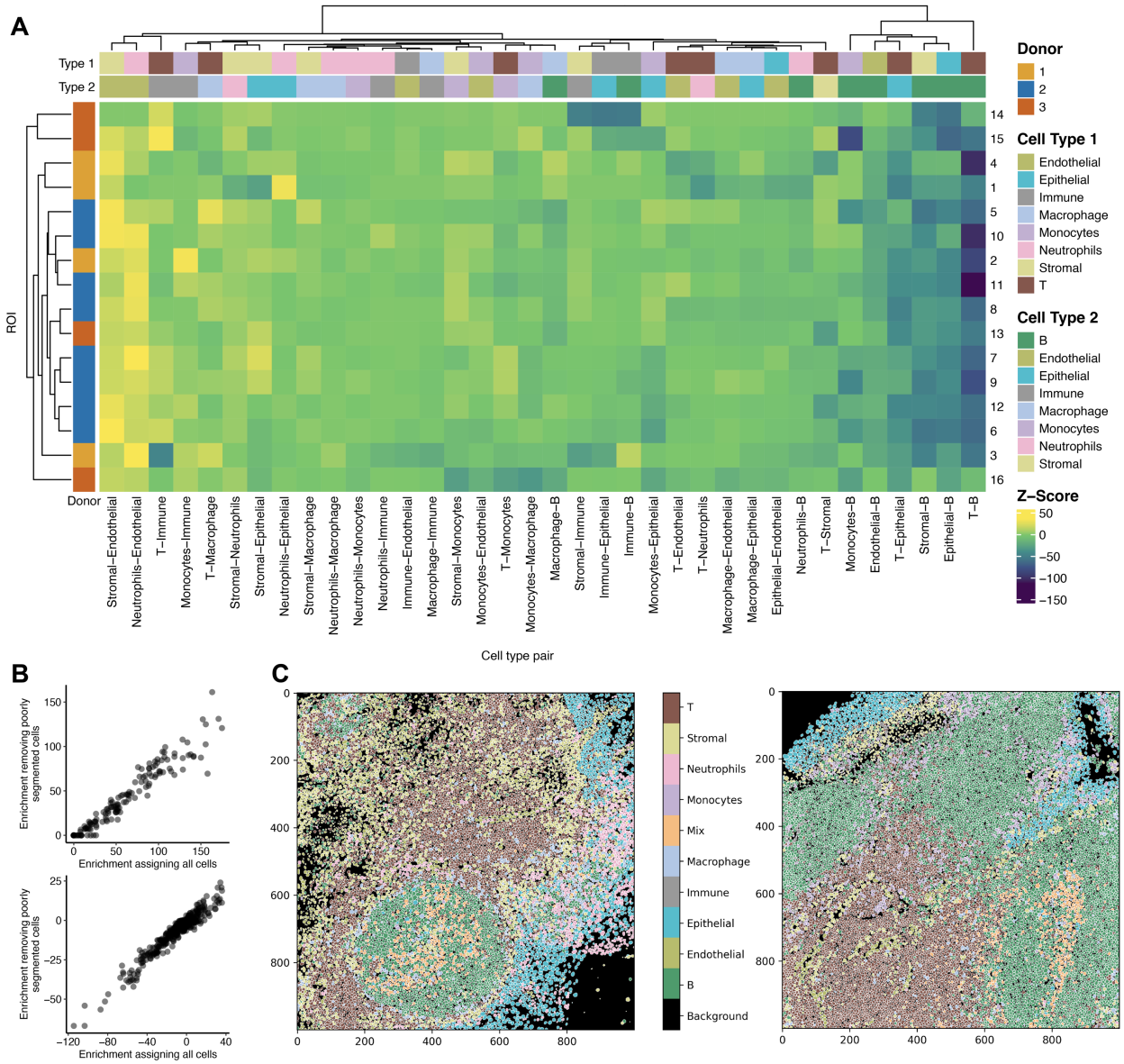

Figure 6 A Spatial enrichment of cell type pairs across all tonsil ROIs. **B** Comparison of enrichment values within the same cell type (top) or between cell types (bottom) for two strategies of cluster assignment, either using all cells assigned to their most likely segmentation-error-free clusters or by removing cells with likely segmentation errors before assignment. **C** Images of ROIs 1 (left) & 11 (right) coloured by inferred cell type.

Discussion

While there is much effort in the community to improve the process of cell segmentation from multiplexed imaging[29], fundamental physical limitations such as image resolution (>1um), signal spillover, and 3D tissue slicing effects mean a perfect segmentation is likely impossible. Therefore, we have introduced an algorithm that explicitly acknowledges the input segmentation

273 likely contains errors and uses knowledge of how these imperfections likely impact the
274 quantified expression profiles to effectively denoise the data and learn the underlying cellular
275 phenotypes. Part of the difficulty in approaching this problem was both the lack of ground truth
276 and prior work in this area, which necessitated the development of a complete benchmarking
277 workflow including the generation of control data as well as the creation of a plausibility score to
278 quantify the quality of inferred cellular phenotypes.

279 In developing STARLING, we sought to strike a balance between practical performance and
280 model complexity, which leads to several limitations that may serve as future directions for
281 development. Firstly, beyond the size of each cell we do not explicitly use spatial information in
282 the image. The number of possible reasons for a segmentation error in highly multiplexed
283 imaging meant we opted for an approach that modeled the consequences on cellular
284 expression rather than explicit spatial relationships giving rise to segmentation errors. While this
285 had the advantage that we could use the consequently held-out spatial information as part of
286 the benchmarking setup, future methods may wish to explicitly build this into the model.
287 Secondly, our model is mis-specified with respect to the true data generating process, assuming
288 each segmentation error comes from the pairwise combination of cell clusters before averaging
289 over all possible pairings to compute the final segmentation error probability (see **Methods**). We
290 opted for this as models closer to the true data generating distribution were too flexible and
291 required additional prior information in preliminary testing and our empirical analyses above
292 show this mis-specified approach achieves strong performance in the absence of further *a priori*
293 information. However, future methods may wish to explicitly consider more complex but better
294 specified models incorporating additional prior information.

295 Beyond STARLING, we believe there is huge scope for segmentation-error aware methods for
296 highly multiplexed imaging and even spatial transcriptomics. For example, though not explored
297 here, dimensionality reduction methodologies such as UMAP[30] are an important step in the data
298 analytic workflow of multiplexed imaging data. Segmentation error aware dimensionality
299 reduction could improve the ability of researchers to both interpret the data as well as integrate
300 diverse datasets across conditions and tissues. Similarly, we previously published
301 methodology[12] for automated assignment of cell types in highly multiplexed imaging, where
302 each cell is assigned to an *a priori* specified cell type based on marker over-expression.
303 Variants of such methodology that are explicitly aware of segmentation errors would no doubt
304 improve cellular quantification.

305 # Methods

306 ## Cell pellet preparation

307 1 000 000 cells from each cell line (Jurkat, RAMOS, THP1, BxPC3) were pooled together and
308 washed three times in cold PBS (4 °C). Cells were suspended in 67 µL human plasma and 116
309 µL bovine thrombin and allowed to coagulate for 10-15 minutes on ice. The resulting pellet was
310 transferred into a biopsy capsule and fixed in 10% NBF for 12-24 hours, followed by 3 washes
311 in cold PBS, 1 wash in 50% EtOH, and 1 wash in 70% EtOH for 30 minutes each. Samples
312 were stored in fresh 70% EtOH in a fridge until paraffin embedding and microtome sectioning
313 into 4 µm sections.

314 ## Cell pellet and tonsil staining

315 Tonsil sections were formalin-fixed and paraffin embedded, and obtained from the surgical
316 pathology unit at Mount Sinai Hospital, Toronto, Canada (REB# 20-0178-E). For IMC staining,
317 cell pellet and tonsil sections were deparaffinized by first baking for 1 hour at 60 °C, followed by
318 dewaxing in a series of 3 xylene solutions for 10 minutes each and then rehydration in a graded
319 series of alcohol (ethanol:deionized water 100:0, 100:0, 96:4, 90:10, 80:20, 70:30) at 5 min
320 each. Heat-induced epitope retrieval was conducted in a pH 9.2 Tris-EDTA buffer at 95 °C for
321 30 minutes. Samples were allowed to cool and blocked using a 3% BSA and 5% horse serum in
322 0.1% TBST (Tris buffered saline (TBS) containing 0.1% Tween-20) blocking/staining buffer for 1
323 hour at room temperature. Unconjugated primary antibodies were stained overnight at 4 °C,
324 followed by conjugated secondary antibodies for 1 hour at room temperature and conjugated
325 primaries overnight at 4 °C, with 15 minute TBS washes after each stain. All antibodies were
326 diluted in the blocking/staining buffer. Iridium staining (^{191}Ir/^{193}Ir) was done at 500 nM in TBS for
327 5 minutes. Slides were washed for 15 minutes in TBS, dipped into ddH2O, and air dried before
328 the IMC acquisition. Information for antibodies used to stain cell pellet and tonsil sections,
329 including metal tag, clone, company, catalog and lot numbers, are provided in **S. Table 3 & 4.**

330 ## Imaging mass cytometry

331 Images were acquired using a Hyperion or Hyperion XTi Imaging System (Standard Biotools).
332 Chosen regions of interest on the cell pellet or tonsil sections were acquired using laser-ablation

333 in a rastered pattern at 400 or 800 Hz, and preprocessing performed using commercial
334 acquisition software (Standard Biotools).

335 Data pre-processing for tonsil data

336 Tonsil IMC data were preprocessed and segmented using an in-house integrated flexible
337 analysis pipeline (ImcPQ) available at https://github.com/JacksonGroupLTRI/ImcPQ. The
338 analysis pipeline is implemented in Python. Briefly, data were converted to TIFF format and
339 segmented into single cells using the pipeline to classify pixels based on a combination of
340 antibody stains to identify membranes/cytoplasm and nuclei (**S. Table 5**). The stacks were then
341 segmented into single-cell object masks.

342 Data preprocessing for cell pellet data

343 Cells were segmented using the Mesmer model in DeepCell v0.12.3 using an average of the
344 two DNA channels as the nuclear channel and an average of the four cell marker channels (b-
345 Actin, CD20, CD3, CD68) for the cytoplasm/membrane channel. Arcsinh normalization with a
346 cofactor of 5 was applied to all channels prior to averaging. Single-cell expression for each
347 marker was quantified as the mean pixel value for each channel. Clustering was performed with
348 Scanpy v1.9.5 and Phenograph v1.5.7 using Phenograph Leiden clustering on default
349 parameters, followed by STARLING. Clusters were assigned to the cell type associated with the
350 highest normalized expression cell type marker if the difference with the next highest cell type
351 marker is at least 0.15.

352 Data re-processing for public datasets

353 In this study, the Basel cohort (Jackson et al., 2020) and the Metabrics cohort (Ali et al., 2020)
354 were selected. We utilized 358 and 548 images from the Basel and Metabrics cohorts,
355 respectively, along with 16 unpublished Tonsil images. We employed Mesmer (Deepcell
356 v0.12.4) for generating single-cell masks. For images from Basel and Metabrics, control
357 channels and the markers H3K27me3, pHH3, and totHH3 were excluded. We designated DNA1
358 and DNA2 as nuclear channels, while the other markers were categorized as
359 cytoplasm/membrane channels. In the case of Tonsil images, only DNA1 and DNA2 were
360 included as nuclear channels (excluding TCF1), and the other markers were considered as
361 cytoplasm/membrane channels. After defining the marker types, we calculated the average

362 nuclear and cytoplasmic expression levels at each location in each image. This resulted in two
363 distinct channels as required for Mesmer input parameters: one representing the nuclear
364 channel and the other representing the cytoplasm channel. Lastly, we determined the single-cell
365 expression for each protein by calculating the average expression levels within the boundaries
366 of each mask, inclusive.

367 **Expression matrix**

368 In addition to utilizing the segmentation masks provided by Basel and Metabrics cohorts, we
369 generated our own segmentation masks using MESMER. Unlike the Basel and Metabrics
370 cohorts, the Tonsil images lacked panel spillover corrections, requiring additional processing
371 steps. To address the spillover issues, we obtained the spillover matrix from Jackson's lab,
372 which was experimentally generated. Subsequently, we applied the CATALYST R package for
373 the necessary corrections [15].

374 Next, we utilized MESMER [14], a cell segmentation method for generating cell masks from
375 multiplexed imaging data. This method is recognized for its accuracy, comparable to human
376 performance, without the need for manual annotation, making it seamlessly integrated into our
377 pipeline. To obtain the masks, we separately normalized (min-max) the cytoplasm/membrane
378 and nuclear channels. Subsequently, we calculated averaged images from these channel types,
379 as required for MESMER's input. Following this, the two resulting images underwent another
380 round of normalization (min-max) before being used in MESMER, generating masks that
381 represent single cells.

382 To create a cell's expression, we mapped its masked locations to the same locations in the IMC
383 image to compute the cell expression by averaging expressions for each measured protein
384 within each masked area (border inclusive). This process was repeated for all masks in each
385 ROI (i.e., image), resulting in a cell expression matrix of N cells and P proteins.

386 **Plausibility score**

387 We introduced plausibility scores (PS) to turn to prior knowledge of cell type markers into a
388 quantitative measure of clustering performance. Each marker for each cluster is categorized as
389 either non-expressed (if the average expression is in the lower 25% of all clusters) or expressed
390 otherwise. Next, we collate a priori known sets of markers that should either never be co-

expressed (e.g. T Lymphocyte marker CD3 and B Lymphocyte marker CD20) in the tissue of origin, or should always show conditional co-expression (e.g. CD3 conditional on CD45).

For each pair and each scenario, we next define “allowed” regions. For markers never expected to be co-expressed, this corresponds to the three quadrants that touch one of the axes, i.e. the regions where neither marker is expressed or only one is. For the conditional co-expression scenario where marker 1 is only expressed if marker 2 is, this region corresponds to the quadrants where (i) neither marker is expressed, marker 2 is expressed but not marker 1, or marker 1 and 2 are both expressed. The plausibility score is then simply computed as the proportion of clusters that fall into the allowed regions, averaged over all marker pairs for which information exists about co- or mutually exclusive expression.

To assess Plausibility Scores (PS), we divided cells into two distinct categories: neighboring and non-neighboring cells. A cell was classified as a neighboring cell if at least a portion of its border was adjacent to another cell. For each category, we implemented clustering algorithms—PhenoGraph, FlowSOM, and Kmeans—on their respective expression matrices, using the default settings for these clustering methods. To further evaluate the effectiveness of PS, we experimented with arcsinh transformation on the raw data and as well as using cofactors of 1 and 5. Additionally, each unique parameter setting was run 50 times, with a subsample of 10,000 cells each time. It is important to note that in cases where a category had fewer than 10,000 cells, such as with the Tonsil images, we upscaled the data with replacement.

STARLING

For a given cell expression matrix measuring N cells and P markers, highly multiplexed imaging quantifies a vector of cell sizes $\mathbf{S} \in \mathbb{R}_+^N$ and a matrix of protein expression $\mathbf{Y} \in \mathbb{R}_+^{N \times P}$. Here, we augment a mixture model to include further latent variables representing whether each cell has been segmented with errors, and if so which underlying clusters of cells contribute to the segmentation error. Let $z_n \in 1, \ldots, C$ be the cluster indicator of cell n for one of C clusters. We introduce a binary variable d_n that signifies whether cell n is a mis-segmented cell, and $\gamma_n \in \{0,1\}^C \times \{0,1\}^C$ denotes the possible pairings of clusters to which segmented cell n belongs in the case $d_n = 1$.

Next, we Introduce cluster centroids $\mu_c \in \mathbb{R}_+^P$. Our model assumes that conditioning on the cell containing a segmentation error and the cluster identities of each true cell being $\gamma_c = [c, c']$, assuming a Gaussian or Student-T likelihood results in additive means and variances with averaged mean expression profiles $(\mu_c + \mu_{c'})/2$. This allows us to specify the following generative probabilistic model:

$$d_n \sim \text{Bernoulli}(\delta); z_n \sim \text{Categorical}(\pi); \gamma_n \sim \text{Categorical}(\tau)$$

$$y_n | d_n = 0, z_n = c \sim T(\mu_c, \mathbf{\Sigma}_c)$$

$$y_n | d_n = 1, \gamma_i = [c, c'] \sim T(\frac{\mu_c + \mu_{c'}}{2}, \frac{\Sigma_c + \Sigma_{c'}}{2})$$

Note that the generative model here is mis-specified with respect to the true data generating distribution: the segmentation error scenario is assumed to come from equal contributions from two clusters, where in reality a segmentation error can result from any number of clusters being averaged under any proportions. We opted for this formulation as preliminary model development suggested inference under a more faithful model was too flexible and would require significant prior knowledge in order to fix what the “true” cell types were which is undesirable. Furthermore, when we compute segmentation error probabilities, we average over every possible pairing of segmentation errors weighted by their relative probability, which explains the ability to capture general segmentation errors as demonstrated above.

We further incorporate cell size information into the model. We assume in the case of a singlet, the cell size scatters around the average cell size for that cluster as $s_n | d_n = 0, z_n = c \sim T(\psi_c, \omega_c)$, where ψ_c and ω_c denote the cluster cell sizes and variances. However, in the case of a wrongly segmented cell, the cell size may lie in-between the maximum size of either cluster in the case of maximum z-plane overlap, to the sum of the cluster sizes in the case of no overlap. Fortunately, we can analytically integrate out this unknown cell size and defining $L = \max(\psi_c, \psi_{c'})$ and $U = \psi_c + \psi_{c'}$ gives:

$$s_n | d_n = 1, \gamma_n = [c, c'], v \sim N(v, \omega_c + \omega_{c'}) = \frac{1}{2(U - L)} \left[ERF \frac{U - s}{\sqrt{2}(\omega_c + \omega_{c'})} - ERF \frac{L - s}{\sqrt{2}(\omega_c + \omega_{c'})} \right]$$

To further constrain the model, we noted that segmentation error cells may be cheaply synthesized by averaging the expression of existing cells as is common when training scRNA-seq doublet detection methods[25]. Therefore, rather than just maximize the likelihood of the model as described above, we add a discriminator to the loss that ensures the resulting probabilities $p(d_n|\cdot)$ as calculated by the model can accurately identify synthetic imperfect cells vs. singlets (the overall dataset of which we denote $\tilde{\mathcal{D}}$). Therefore, the loss function becomes:

$$\mathrm{LOSS}(\theta) = -\ell_m(\theta) + \lambda H\left(p(d_n|\theta, \tilde{\mathcal{D}}), \tilde{\mathcal{D}}\right)$$

where $\ell_m(\theta)$ is the log-likelihood function from the model, $H(.)$ is a Binary Cross-Entropy (BCE) function applied to the model predicting the synthesized data, and λ is a fixed hyper-parameter that controls the balance between Log-Likelihood from the model and Binary Cross-Entropy (BCE) from the simulated data (default λ=1). Higher λ values accentuate the importance of the synthetic data (singlet detection), while lower values emphasize the input data (data likelihood) more.

Moreover, we set the segmentation error rate and γ for the synthetic data (see synthetic data) and as well as the segmentation error prior to be 40% and uniform prior for singlet (π) and segmentation error clusters (τ). STARLING is trained using Adam optimization with a learning rate of 1e-3 and minibatching (size 512) for both CPU and GPU enabled machines. For each dataset, we executed STARLING five times with random initialization, subsampling 10,000 cells each time to evaluate the model performance. After identifying the optimal set of model parameters, STARLING was then applied to each cohort without subsampling. This process was repeated five times, selecting the run with the lowest likelihood score for further analysis.

Benchmarking with plausibility score

To evaluate cluster centroid quality, we employed compared the plausibility scores (see above) for the centroids computed by STARLING compared to three common clustering algorithms: Kmeans (as implemented in the sklearn package v0.24.2), PhenoGraph (as implemented in the scanpy package v1.7.2) and FlowSOM (v0.1.1), all using default settings. For each baseline method, we further employed a two-step approach, where a random forest algorithm (implemented using the sklearn package v0.24.2) trained to discriminate singlet cells vs. two cells pooled together, then re-predicting whether each originally segmented cell was likely a

singlet, and removing cells with singlet probability < 0.5. For each analysis, we conducted an experiment involving a subsample of 10,000 cells for the three datasets, repeated five times for each parameter combination. We utilized both the provided masks and those obtained by MESMER, resulting in a total of 720 distinct model variations for each data source. For each experiment, the number of clusters (k) was initially determined using PhenoGraph, and this value was then applied consistently in Kmeans, FlowSOM and STARLING to ensure fair comparisons.

Benchmarking for segmentation error detection

To benchmark STARLING's ability to detect segmentation errors, we computed $p(d_n = 1|\theta, X)$ for every cell as the probability of a segmentation error. We segmented each dataset using MESMER v0.12.4 in both "whole cell" and "nucleus" mode. We denoted each segmented whole cell as containing a likely segmentation error if multiple cells segmented in nucleus only mode were found in a single cell segmented in whole cell mode. To confidently map cells, we further restricted the set of whole cells denoted as likely segmentation errors to those where at least one nuclear only segmentation took up >30% of the whole cell segmentation area.

To correlate STARLING's ability to associate segmentation errors with adjacent cells, we fit the model to estimate cluster assignments for every cell. For every cell n, we found the segmentation error free cluster identity (i.e. the max of $p(z_n = c|d_n = 0, X)$) for both that cell, and all of its immediate neighbours denoted $c'_1, ..., c'_M$ for *M* touching cells. Next, we extract the *C*x*C* matrix Γ for each *n* denoted G (where we drop the *n* subscript for future clarity). Next, we let *v* be the cth row of G, i.e. $v = G[c, :]$ — the set of segmentation error probabilities for cell *n* for the mixing between the assigned cell type *c* and all others. Then, let v_a be the elements of v belonging to the identities of the touching cells/neighbours, i.e. $v_a = v[c'_1, ...c'_M]$ and let v_b be the other elements of *v*, i.e. $v_b = v[\sim [c'_1, ..., c'_M]]$. We repeated this for every cell, and the difference between v_a and v_b (i.e. mixing probabilities for touching vs non-touching cells).

Tonsil analysis

STARLING was fitted using mini-batching (batch size 512) and C=25 clusters on all cells from the tonsil data ROIs for 5 different initializations, with the final model chosen by the fit that minimized the log-likelihood score. For enrichment analysis we employed the Python Squidpy

package v1.2.0 to generate z-scores for a cell-type-by-cell-type square enrichment matrix. The diagonal and off-diagonal z-scores represent the z-scores of cell types within the same and different pairs of cell types, respectively. Each Region of Interest (ROI) was analyzed independently, and the results for the two cases were presented separately. Segmentation-error-free cell phenotypes analyzed with STARLING may be interpreted in two ways. Firstly, cells with high probability of segmentation error (i.e. $p(d_n = 1|\theta, X)$ greater than 50%) may be removed, then the most likely cluster assigned (*Method 1*). Alternatively, we can ask STARLING for the "best guess" assuming a cell has no segmentation error (i.e. $p(z_n = c|d_n = 0)$), which we refer to as *Method* 2.. We assigned cells using both of these methods, with the subsequent enrichment analysis comparing them shown in **Fig. 6B**.

Data and code availability

The STARLING Python package is available at https://github.com/camlab-bioml/starling. Code to reproduce the analyses in this paper is available at https://github.com/camlab-bioml/starling_plots.

Acknowledgements

This work is funded by NSERC Discovery Grants RGPIN-2020-04083 to KRC and RGPIN-2021-03404 to HWJ, and a CIHR project grant PJT-175270 to KRC. Both KRC and HWJ acknowledge support from the Canada Research Chairs program and the Canadian Foundation for Innovation. This work was further supported by a University of Toronto Data Sciences Institute Research Software Development Support.

Ethics approval

The data generation and analysis in this manuscript is covered under REB #20-0178-E at Mount Sinai Hospital, Toronto ON, Canada.

Contributions

Conceptualization: YL, AS, HWJ, KRC. Data analysis: YL, DC, SA, KRC. Software development: YL, CK, KRC. Data generation: EC, AD, MM. Manuscript writing and review: YL, AS, HWJ, KRC.

References

1. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
2. de Souza, N., Zhao, S. & Bodenmiller, B. Multiplex protein imaging in tumour biology. *Nat. Rev. Cancer* (2024) doi:10.1038/s41568-023-00657-4.
3. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).
4. Ali, H. R. *et al.* Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat Cancer* **1**, 163–175 (2020).
5. Sorin, M. *et al.* Single-cell spatial landscapes of the lung tumour immune microenvironment. *Nature* **614**, 548–554 (2023).
6. Moldoveanu, D. *et al.* Spatially mapping the immune landscape of melanoma using imaging mass cytometry. *Sci Immunol* **7**, eabi5072 (2022).
7. Karimi, E. *et al.* Single-cell spatial immune landscapes of primary and metastatic brain tumours. *Nature* **614**, 555–563 (2023).
8. Damond, N. *et al.* A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell Metab.* **29**, 755–768.e5 (2019).
9. Amir, E.-A. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
10. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and

549 interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015).
550 11. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells
551 that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
552 12. Geuenich, M. J. *et al.* Automated assignment of cell identity from single-cell multiplexed
553 imaging and proteomic data. *Cell Syst* **12**, 1173–1186.e5 (2021).
554 13. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for
555 cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
556 14. Greenwald, N. F. *et al.* Whole-cell segmentation of tissue images with human-level
557 performance using large-scale data annotation and deep learning. *Nat. Biotechnol.* **40**,
558 555–565 (2022).
559 15. Chevrier, S. *et al.* Compensation of Signal Spillover in Suspension and Imaging Mass
560 Cytometry. *Cell Syst* **6**, 612–620.e5 (2018).
561 16. Bai, Y. *et al.* Adjacent Cell Marker Lateral Spillover Compensation and Reinforcement for
562 Multiplexed Images. *Front. Immunol.* **12**, 652631 (2021).
563 17. Amitay, Y. *et al.* CellSighter: a neural network to classify cells in highly multiplexed images.
564 *Nat. Commun.* **14**, 4302 (2023).
565 18. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, e27041 (2017).
566 19. Lu, P. *et al.* IMC-Denoise: a content aware denoising pipeline to enhance Imaging Mass
567 Cytometry. *Nat. Commun.* **14**, 1601 (2023).
568 20. Baars, M. J. D. *et al.* MATISSE: a method for improved single cell segmentation in imaging
569 mass cytometry. *BMC Biol.* **19**, 99 (2021).
570 21. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis
571 of single-cell RNA-seq data with Bioconductor. *F1000Res.* **5**, 2122 (2016).
572 22. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell
573 Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281–291.e9 (2019).

574 23. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-
575 Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329–337.e4
576 (2019).
577 24. Bernstein, N. J. *et al.* Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-
578 Supervised Deep Learning. *Cell Syst* **11**, 95–101.e5 (2020).
579 25. Xi, N. M. & Li, J. J. Benchmarking Computational Doublet-Detection Methods for Single-
580 Cell RNA Sequencing Data. *Cell Syst* **12**, 176–194.e6 (2021).
581 26. Schapiro, D. *et al.* MCMICRO: a scalable, modular image-processing pipeline for
582 multiplexed tissue imaging. *Nat. Methods* **19**, 311–315 (2022).
583 27. Windhager, J. *et al.* An end-to-end workflow for multiplexed image processing and analysis.
584 *Nat. Protoc.* **18**, 3565–3613 (2023).
585 28. Alina Selega, K. R. C. Multi-objective Bayesian Optimization with Heuristic Objectives for
586 Biomedical and Molecular Data Analysis Workflows. *Transactions on Machine Learning*
587 *Research* (2023/3).
588 29. Israel, U. *et al.* A Foundation Model for Cell Segmentation. *bioRxiv* 2023.11.17.567630
589 (2023) doi:10.1101/2023.11.17.567630.
590 30. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat.*
591 *Biotechnol.* (2018) doi:10.1038/nbt.4314.

592