Automated assignment of cell identity from single-cell multiplexed imaging and proteomic data

Michael J. Geuenich[1,2,*], Jinyu Hou[1,3,*], Sunyun Lee[1,3,*], Hartland W. Jackson[1,2,5], and Kieran R. Campbell[1,2,3,4,5,6,†]

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada
[2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[3]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada
[4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[5]Ontario Institute of Cancer Research, Toronto, Ontario, Canada
[6]Vector Institute, Toronto, Ontario, Canada
[*]Authors contributed equally
[1]kierancampbell@lunenfeld.ca

Abstract

The creation of scalable single-cell and highly-multiplexed imaging technologies that profile the protein expression and phosphorylation status of heterogeneous cellular populations has led to multiple insights into disease processes including cancer initiation and progression. A major analytical challenge in interpreting the resulting data is the assignment of cells to *a priori* known cell types in a robust and interpretable manner. Existing approaches typically solve this by clustering cells followed by manual annotation of individual clusters or by strategies that gate protein expression at predefined thresholds. However, these often require several subjective analysis choices such as selecting the number of clusters and do not automatically assign cell types in line with prior biological knowledge. They further lack the ability to explicitly assign cells to an unknown or uncharacterized type, which exist in most highly multiplexed imaging experiments due to the limited number of markers quantified. To address these issues we present Astir, a probabilistic model to assign cells to cell types by integrating prior knowledge of marker proteins. Astir uses deep recognition neural networks for fast Bayesian inference, allowing for cell type annotations at the million-cell scale and in the absence of previously annotated reference data across multiple experimental modalities and antibody panels. We demonstrate that Astir outperforms existing approaches in terms of accuracy and robustness by applying it to over 2.1 million single cells from several suspension and imaging mass cytometry and microscopy datasets in multiple tissue contexts. We further showcase that Astir can be used for the fast analysis of the spatial architecture of the tumour microenvironment, automatically quantifying the immune influx and spatial heterogeneity of patient samples. Astir is freely available as an open source Python package at https://www.github.com/camlab-bioml/astir.

INTRODUCTION

Recent advances in high-throughput sequencing and imaging technologies have enabled high-dimensional profiling of gene expression, protein abundance, and phosphorylation status in both human tissues and model organisms. Much of the attention in computational methods development for these technologies has focused on single-cell RNA-sequencing (scRNA-seq), which has led to improvements in our understanding of organism development [1], drug resistance in cancer treatment [2], as well as the role of the tumour microenvironment in cancer initiation and progression [3, 4]. However, despite its ability to capture an unbiased representation of the whole-transcriptome at single-cell resolution, scRNA-seq cannot directly measure protein abundance or phosphorylation status and requires special modifications to operate on archival tissue [5], with spatial transcriptomics technologies such as HDST [6] suffering from similar drawbacks.

To address these issues, a range of technologies have been invented to measure the expression of multiple proteins while retaining the spatial location of origin in the tissue section, collectively termed "highly multiplexed imaging". These include Imaging Mass Cytometry (IMC) [7] and Multiplexed Ion Beam Imaging (MIBI) [8], which stain tissues with antibodies conjugated with heavy metals followed by laser ablation and mass

spectrometry, allowing for the unbiased expression quantification of around 40 features [9] while retaining the spatial location of origin of each cell. Additional related technologies such as MxIF [10] and CyCIF [11] have also been developed allowing for the multiplexed detection of proteins using microscopy and immunofluorescence techniques , thereby retaining spatial information while only requiring standard microscopy equipment.

A major analytical challenge when interpreting the large quantities of single-cell data generated by these technologies is to relate each high-dimensional cell phenotype to *a priori* known cell types, which is crucial for a number of applications such as linking tumour microenvironment composition to patient outcomes [12]. For single-cell protein and gene expression assays, several methods have been developed to assign cells to cell types that rely on clustering cells followed by the manual annotation of clusters [13–17] (Figure 1A). For lower-dimensional platforms such as flow cytometry, manual gating strategies are traditionally employed on expression values of marker genes. However, these methods require prior knowledge of marker gene expression thresholds, lack reproducibility [18], and cells that fall outside the thresholded values will not be assigned/be mis-assigned rather than being assigned probabilities that distribute across the most likely cell types. While these issues have recently been addressed in the field of scRNA-seq with the advent of computational tools such as CellAssign that use probabilistic modelling to identify cell types based on marker gene expression [12, 19], such tools have yet to be created for single-cell highly multiplexed imaging and proteomic technologies. An additional challenge for cell type assignment in these data modalities is the set of proteins measured is typically experiment-specific meaning supervised learning approaches to align cells to reference populations that have become popular in scRNA-seq analysis [13, 20] cannot be used.

To address these challenges, we introduce Astir (ASsignmenT of sIngle-cell pROteomics), a scalable probabilistic model to robustly assign cell types to data generated from a range of single-cell highly-multiplexed imaging and proteomic technologies. Expanding on our previous work [12], Astir takes both measured expression data and an *a priori* specified set of marker proteins for all expected cell types as input and employs a novel statistical machine learning model to assign cell type probabilities to each cell (Figure 1B). We apply Astir to three IMC datasets and one suspension mass cytometry dataset comprising a total of 552 patients and 930 cores from breast cancer tumours representing all clinical subtypes and grades of pathology, as well as normal tissue from the same patients. In addition, we include a CyCIF multiplexed microscopy dataset of formalin-fixed paraffin-embedded tissue comprising normal and cancer tissue from 40 different patients. We compare the Astir cell type assignment to multiple alternative workflows and show that Astir outperforms existing methods on average when assigning cells to the set of expected cell types present in the data and is uniquely able to assign cells to an *Other* type when cell type marker proteins are missing or unknown. We further demonstrate that Astir is more robust to variability in cell segmentation – a key upstream processing step in the analysis of highly multiplexed imaging data – and to imbalance in the underlying cell type proportions. We apply Astir to two IMC datasets of breast cancers and use the cell type assignments coupled with the spatial information to derive a novel interpretable spatial heterogeneity score that enables us to reproducibly track phenotypic changes with spatial tumour architecture across cohorts. Finally, we investigate the expression of marker proteins in *Unknown* and *Other* cells and show that this assignment is driven by the simultaneous expression of implausible marker proteins in part caused by non-specific staining of the individual cores.

RESULTS

AUTOMATED ASSIGNMENT OF CELL TYPE FOR SINGLE-CELL MULTIPLEXED IMAGING AND PROTEIN EXPRESSION DATA

The Astir framework implements a statistical model that assumes a cell type represents a phenotypically static expression profile. Astir takes a cell-by-protein normalized expression matrix as input along with a `yaml` file specifying marker proteins for each expected cell type within the dataset, which is closely guided by the initial design of the protein measurement (antibody) panel. The Astir cell type model builds upon our previous work [12], assuming cell types express their marker proteins at relatively higher levels than other cell types with a modified likelihood function allowing for continuous intensity values as typically measured by such experimental technologies and a novel protein-protein covariance structure that requires marker protein expression to be correlated within each cell type (see *Methods*). It further allows for the

2

existence of an *Other* cell type that does not express the specified marker proteins in a recognizable manner and may represent *de novo* cell types within the data. Further variation due to batch or patient specific effects may be incorporated in the model which supports the inclusion of arbitrary experimental design matrices.

To perform cell type assignment, Astir calculates the probability that each cell is of a given specified cell type under a Bayesian framework (see *Methods*). Notably, we employ deep neural networks [21] that map from the observed data to the cell assignment probabilities, allowing for fast mini-batch inference scaling to millions of cells that is both constant in memory and enables immediate prediction of cell types for new, unseen data (Figure 1C). To critique the resulting cell type assignments, Astir includes a set of post-fit diagnostics to ensure all cell types express their marker proteins at significantly higher levels than other cell types and flags all marker and cell type combinations for which this is violated. The Astir software is implemented in Pytorch [22], with a full set of documentation and tutorials online and is publicly installable from pypi[1].

To demonstrate the benefits of Astir, we assigned cells to cell types across a number of highly-multiplexed imaging and suspension mass cytometry datasets. We collated (i) two cohorts of IMC data ("Basel" and "Zurich") from a recent publication [23] of breast cancer tumour and normal samples comprising 307 patients and 735 cores with a total of 1.2 million cells, (ii) a suspension mass cytometry breast cancer dataset ("Wagner") comprising 144 patients with one sample per patient and a total of 643,898 cells [24], (iii) a further IMC breast cancer dataset ("Schapiro") with 51 patients and cores comprising a total of 45,585 cells [25], and (iv) a CyCIF multiplexed microscopy dataset from 13 different anatomical sites made up of cancer and normal tissue ("Lin") comprised of 40 patients and a total of 179,558 cells [11]. Overall we assigned cell types to over 2.1 million cells using the marker proteins listed in Figure S1. Cells that co-expressed improbable marker combinations are automatically assigned as *Other* by Astir, while cells with a high probability of belonging to more than one cell type are assigned as *Unknown* (Figures S2 to S15).

Using Astir we assigned over 92% of cells to a cell type on average across all cohorts (Figure 1D), with the remainder assigned to *Unknown* and *Other* types. A certain fraction of cells being non-identifiable is expected given the limited set of proteins measured results in a fundamental inability to quantify all possible cell types that would require proteome- or transcriptome-wide measurement technologies. t-distributed stochastic neighbour embedding (t-SNE, [26]) reduced dimensionality representations of the cellular expression profiles based on overlapping proteins between the Basel, Zurich, and Wagner cohorts highlights a clear separation between the assigned cell types (Figure 1E) in the reduced dimensional space. It further shows separation between the Wagner and Basel/Zurich cohorts (Figure 1F) due to them originating not only from distinct experiments and antibody panels but from fundamentally different experimental technologies. This demonstrates a marked strength of Astir in its ability to automatically assign cells to common cell types across distinct experimental modalities.

Using the Astir cell type assignments we computed the proportion of each core containing a given cell type and compared to pathologist estimates of epithelial, stromal, and inflammatory compartments in cores from the Basel cohort. There was significant positive correlation in each case (Figure 1G), though with expected variation given the fundamental differences in measurement technologies. These largely arise due to pathologist estimates relying on cellular morphology with difficulty resolving small lymphocytes compared to the Astir assignments based on expression of marker proteins alone. We conclude that Astir can be used to robustly assign large numbers of cells to interpretable pre-defined cell types.

ASTIR OUTPERFORMS EXISTING CLUSTER-AND-INTERPRET METHODS AND IS ROBUST TO INPUT MARKER MIS-SPECIFICATION

We next compared Astir to existing workflows that assign cell types based on clustering followed by cluster interpretation via either inspection of marker proteins or gene set enrichment analysis. Using a set of commonly employed clustering methods for similar data types (PhenoGraph [15], FlowSOM [16] and ClusterX [17]), we clustered cells from all cohorts considered in an unsupervised manner across a range of plausible parameter values (*Methods*) and compared the identified clusters to the Astir assignments (Figures S16 to S45). Using the subsequently identified clusters, we sought to quantify how well each cluster corresponded
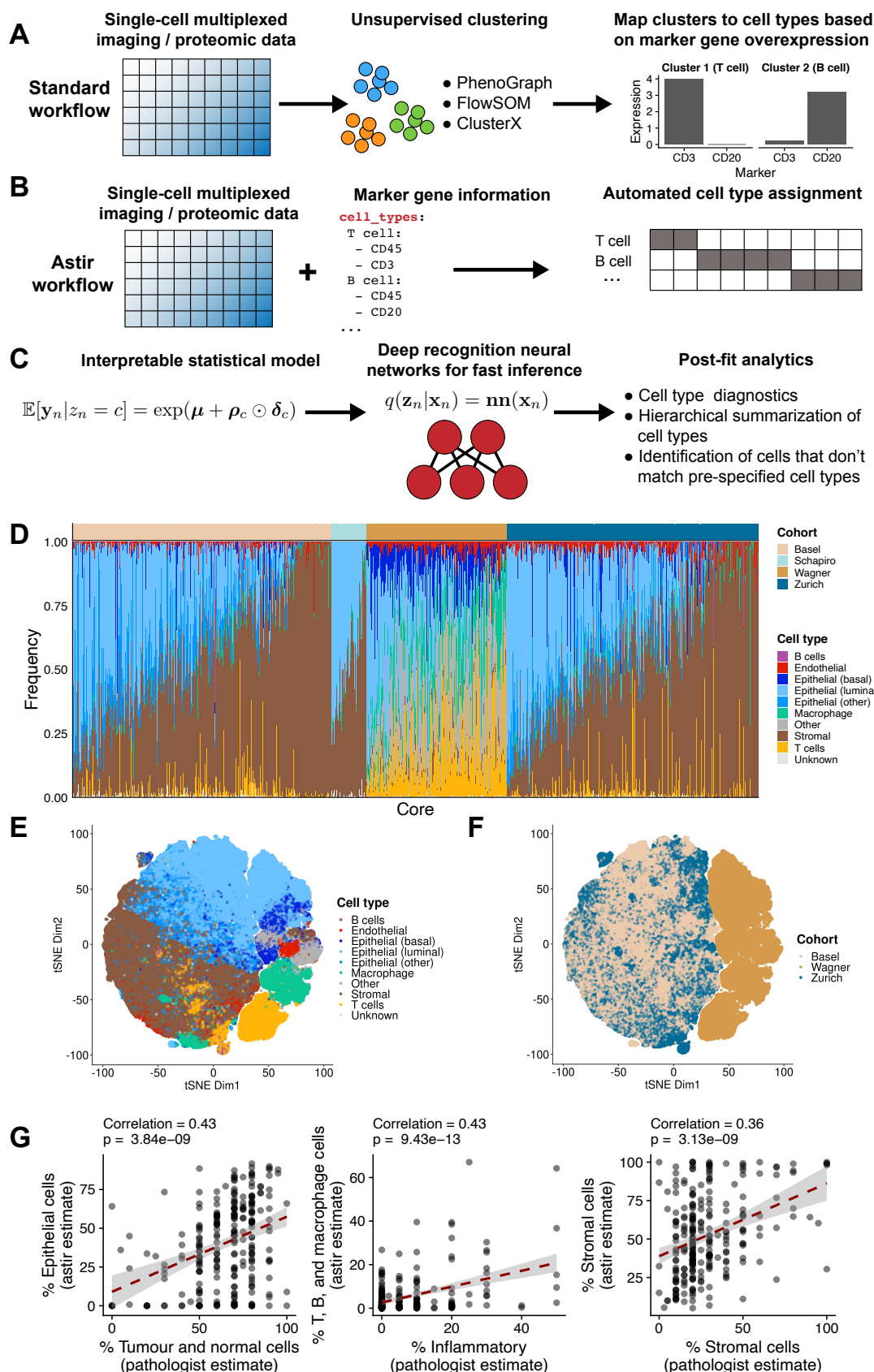
---

[1] https://pypi.org/project/astir/

Figure 1: Astir automates cell type assignment for highly multiplexed imaging and single-cell proteomic data.

Figure 1: A Existing cell type annotation methods cluster expression data and subsequently leverage marker over-expression to annotate clusters. **B** Astir combines input single-cell multiplexed imaging or proteomic expression data with a list of known marker proteins and probabilistically assigns each cell to a cell type. **C** Astir is based on an interpretable statistical model that assumes that marker proteins are more highly expressed in their respective cell types and uses deep recognition neural networks for fast Bayesian inference. Post-fit diagnostics then allow for quality assessment of the assigned cell identities. **D** Astir identifies inter-patient tumour cellular composition heterogeneity across breast cancer samples from the Basel, Zurich, Schapiro, and Wagner cohorts. **E-F** tSNE plots for the Basel, Zurich and Wagner cohorts coloured by cell type and cohort demonstrates both cohort-specific variation and Astir's ability to confidently assign cell types across cohorts. **G** A comparison of Astir cell type proportions to pathologist estimates (to the nearest 5%) based on immunohistochemistry for the Basel cohort demonstrates significant association across epithelial, inflammatory, and stromal compartments.

to a desired cell type we wished to identify. Using gene set enrichment analysis [27], we matched each cluster to its most likely cell type and derived a score to quantify how well a given clustering can identify a cell type without further post-processing (*Methods*). Briefly, this awards each method a $+1$ if a given cell type within a dataset is resolved to exactly one cluster and -1 if a given cell type is not identified or identified multiple times (i.e. corresponding to multiple clusters), with the sum of the individual scores representing the overall method score for a given dataset. We found that Astir outperformed all existing workflows across all cohorts (Figure 2A). This is largely driven by the fact that many methods will over or under cluster the given cells, even when the correct number of output cell types is specified (such as for FlowSOM k8 in the Basel cohort, see Figures S16 to S45 for representative examples).

Next, we quantified the robustness of Astir and alternative methods to imbalance in cell type proportions in the input data. Taking 45,303 cells from the Basel cohort assigned to the luminal epithelial, stromal, macrophage, and T cell components with $> 95\%$ probability, we created new datasets by mixing the cells of different types with a pre-defined portion of luminal epithelial cells (25%, 30%, 40%,...,80%, 99%) with the other cells types being equally represented in the dataset. We then re-ran Astir and the alternative methods as above, and examined the impact of the underlying cell type proportion on both the percent of cells assigned to the luminal epithelial cell type as well as the number of clusters corresponding to luminal epithelial cells. We found Astir to be the only approach whose assignment of luminal epithelial cells matched the ground truth proportions (Figure 2B, top), with the existing approaches inferring incorrect proportions as the underlying composition became highly skewed towards an abundance of luminal epithelial cells. Similarly, we found Astir to be the only method where the number of luminal epithelial clusters inferred was independent of the underlying sample composition (Figure 2B, bottom). These results have important implications for the clinical utility of such methods given the importance of accurate cell type abundance identification to derive effective oncological biomarkers.

We next examined the robustness of each method to variability in cell segmentation of highly multiplexed imaging data. Cell segmentation is the process of taking the raw, pixel level data and computing the boundary of each cell via machine learning methods applied to membrane and nuclear markers [7] before summarizing to single-cell expression data. We leveraged existing data of three regions of interest (ROIs) segmented by three different users from a previous publication [25] to quantify the consistency of recovered cell type proportions from each method. We retrieved the segmentation mask for each user and ROI and summarized the cell specific expression as the mean expression of a given protein inside the cellular mask, and applied both Astir and the alternative methods to assign cell types. For each core, segmentation, and method, we computed the frequency of each cell type assigned and used the standard deviation in frequency across each segmentation as a metric, with a lower value implying more consistent assignment across segmentations.

Upon ranking each method we found Astir had the top overall ranked consistency between segmentations (Figure 2C), with each workflow based on clustering and post-hoc interpretation exhibiting large variations due with parameter configurations. An exception was the `Cy1x6_33` ROI where Astir ranked poorly in consistency of assignment across segmentations. Upon closer inspection, we found this to be caused by variation in assignment between `Epithelial (luminal)` and `Epithelial (other)` cell types (Figure S46), demonstrating

Figure 2: Comparison of Astir to existing workflows for cell type assignment and robustness to mis-specification of cell type marker proteins.
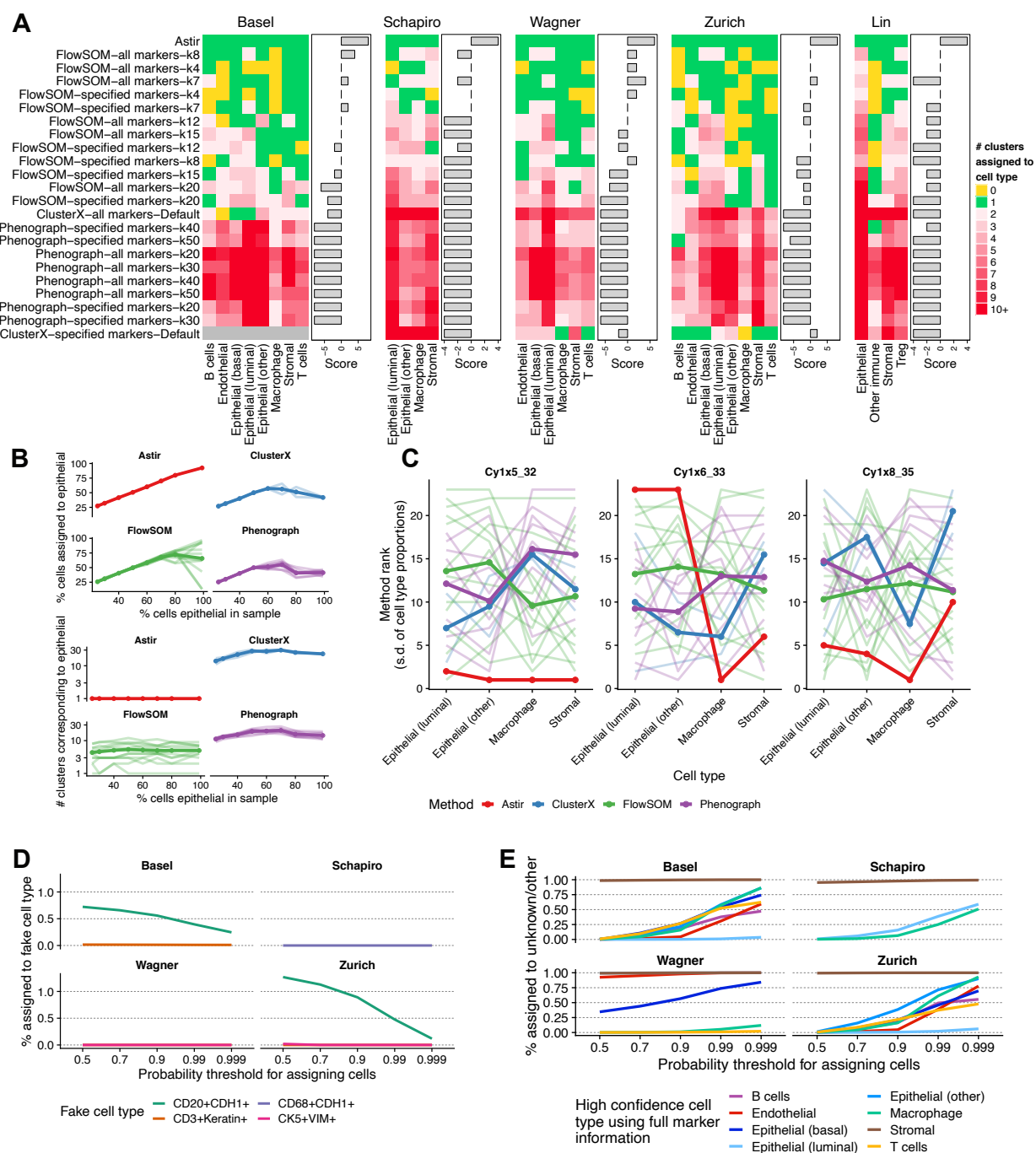
Figure 2: A For each possible workflow (row), cell type (column) and dataset (individual heatmap), we quantified the number of clusters assigned to a given cell type. An ideal algorithm for cell type identification without further post-processing will assign a single cluster to each cell type (green), and not miss a cell type (yellow) or call multiple clusters per cell type (reds). Quantifying this as a "score" by adding a point if exactly one cluster is assigned per cell type and subtracting a point otherwise demonstrates Astir is the only method to assign cell types exactly. Only a single cluster was identified for ClusterX using markers specified in the Basel cohort resulting in the lack of an enrichment score. **B** Consistency between ground truth proportion of luminal epithelial cells in a simulated sample compared to inferred proportion (top) and number of clusters identified (bottom) across a range of methods. Each line represents a possible parameter configuration as per (**A**) with darker line representing the mean trend. **C** Astir has the highest overall consistency of cell type abundance estimates between different segmentations of IMC data across three regions of interest. The y-axis depicts the average rank of each method in terms of lowest standard deviation of cell type frequencies between segmentations. Each line represents possible parameter configurations as per (**B**). **D** Percentage of cells that get assigned to a fictional cell type by Astir defined by combinations of unlikely marker combinations as a function of the probability threshold used to assign cells. **E** Percentage of cells assigned as *Other* or *Unknown* upon the removal of markers for each cell type as a function of the probability threshold used to assign cells.

variability in Astir's cell type assignment in response to segmentation occurs between highly similar cell types. Overall, these results show Astir's assignments are the most robust to upstream processing of highly multiplexed imaging data with important implications for a range of clinical applications.

We next benchmarked the robustness of Astir to mis-specification of the input cell type marker lists. We first created an input marker set containing biologically implausible (or "fictional") cell types, including (i) co-expression of B-cell and epithelial markers (CD20+CDH1+), (ii) co-expression of T-cell and epithelial markers (CD3+Keratin+), (iii) co-expression of macrophage and epithelial markers (CD68+CDH1+, CD68+KRT6+), and (iv) co-expression of basal epithelial and fibroblast markers (KRT5+VIM+). We then re-assigned cell types using Astir including these fictional cell types in the input marker matrix for the four cohorts for which these markers were measured (Basel, Schapiro, Wagner, Zurich), at examined the proportion of cells erroneously assigned to these cell type as a function of the probability threshold for calling a given cell type (Figure 2D). We found that while the mis-classification rate is variable across both cohorts and cell types, Astir exhibits excellent specificity with no fictional cell type being assigned $> 1\%$ of cells when the probability threshold was moderately stringent ($p \geq 0.9$).

Finally, we tested the robustness of Astir to the opposite scenario in which cell types present in the data are not specified in the input marker list. This is a fundamentally harder problem due to the correlated nature of expression data across cell types. As an example, consider the specification of a macrophage, typically defined by co-expression of CD45 and CD68. However, if we remove the canonical macrophage marker CD68 as input, cells instead simply appear as CD45-high, given CD45 is quantified for other cell types positive (such as T cells). Combined with non-specific staining of other markers, we therefore expect macrophage cells will be given some probability of being assigned as a T cell.

To quantitatively assess this, for every cohort we removed each cell type from the input set of cell types to which cells should be assigned and re-fit the Astir assignments. We then evaluated the proportion of cells newly assigned to *Unknown* or *Other* that were originally assigned to the removed cell type with high confidence ($p > 0.95$). The results show significant variability across both cohorts and cell types (Figure 2E). For example, for both the Wagner and Zurich cohorts, Astir demonstrates high sensitivity assigning almost 100% of cells originally assigned to Stromal as *Unknown*/*Other* when Stromal is removed from the input marker lists, presumably due to its phenotypic distinctiveness from other cell types present. In contrast, other cell types such as epithelial and T cells are erroneously assigned to alternative cell types, likely due to marker gene overlap with alternative epithelial and immune cell subsets.

We note that while Astir exhibits variability in this sense, the majority of existing methods do not have the ability to assign cells to *Unknown* and *Other* types. To quantitatively demonstrate this, we iteratively removed

cell type markers starting with stromal cells, stromal and macrophages and finally, stromal, macrophage and endothelial cells prior to running the gene set variation analysis. This led to the misclassification of all stromal, macrophage and endothelial cells upon removal of their respective markers for all methods, parameters and cohort combinations (Figures S47 to S53). While it is theoretically possible to come up with a threshold for the GSVA score at which point a cluster is determined to be of an *Unknown* cell type, this is hard to implement in practice given that this threshold is not immediately obvious as the GSVA score does not represent a probability [27]. Thus, we conclude that Astir is uniquely capable of correctly assigning cells even when appropriate cell identity markers are unavailable, missing, or mis-specified.

INTERPRETABLE STRATIFICATION OF PATIENTS BASED ON CELL TYPE SPATIAL HETEROGENEITY SIGNATURES

We next sought to demonstrate that the retained spatial locations of cells coupled with the Astir cell type assignments could enable automated single-cell pathology and stratification of patients based on higher-order spatial features including immune cell influx or lack thereof. We calculated pairwise distances between all cell types in a given core and derived a core-specific spatial heterogeneity signature as the first principal component of the core-by-cell-type-interaction matrix (*Methods*). This creates a spatial heterogeneity score with higher values corresponding to smaller distances between different cell types (Figure 3A), enabling discrimination between mixed and compartmentalized spatial architectures, which have previously been associated with patient outcomes in triple negative breast cancer [28]. Indeed, inspection of cell types cores with the highest and lowest spatial heterogeneity scores identifies cases with extreme stromal-epithelial-immune mixing (case SP41_83_X14Y2, Basel cohort, Figure 3B left) and extreme compartmentalization between epithelial and T cells (SP43_114_X15Y1, Basel cohort, Figure 3B right).

Given the clinical importance of this spatial heterogeneity across a range of cancers, we examined whether the the impact of spatial tumour architecture on cellular phenotypes could be inferred in a manner reproducible between cohorts. To assess this, for each cell type we regressed the mean expression (per-core) of each protein on the spatial heterogeneity measure assigned to that core and compared the resulting effect sizes between each cohort (Basel and Zurich). The results were reproducible between cohorts, with 7/8 cell types displaying a positive correlation between cohorts in differential gene regulation with 5/8 significantly positively correlated at $p < 0.05$ (Figure 3C).

This analysis is only possible due to the systematic assignment of cell types by Astir that is reproducible between cohorts, enabling the comparison of a phenotypic profile of a given cell type between perturbative conditions such as tumour spatial architecture. This is in contrast to existing approaches whose clusters may conflate the notion of a cell type and phenotypic state, thus prohibiting the standardized measurement of distances between cell types without further post processing. Given the prognostic value of immune influx into the tumour mass [28], computational methods such as Astir that systematically assign cell types without further tuning will form the basis for standardized inference of complex tumour-immune architectures that can act as biomarkers in immuno-oncology.

IDENTIFICATION OF *Unknown* AND *Other* CELL TYPES

Finally, we investigated the expression profiles of cells assigned as *Unknown* or *Other*. Given that the focus of highly multiplexed imaging experiments is often a heterogeneous population of cells and it is impossible to realistically design an antibody panel for all possible cell types, we expect any assignment cells to include those assigned to both *Unknown* and *Other*. However, we wished to ensure that the cell types classified as *Unknown*/*Other* do not over-express a particular combination of measured proteins, implying either we have mis-specified the input marker list or the Astir cell type assignments are incorrect.

Astir classified 0.7%, 0.8%, 25.4%, 0.4% and 3.4% of cells as *Unknown* or *Other* for the Basel, Schapiro, Wagner, Zurich and Lin cohorts respectively using the default setting that if a cell has $> 50\%$ probability of assignment it is assigned to a cell type. We conducted a differential expression analysis of all available measured proteins for each cohort, contrasting *Other* and *Unknown* cells against all assigned cells (Figure 4A-E). We found that across all cohorts, no proteins were expressed higher in *Unknown* and *Other* cells than in assigned cells with the exception of the stromal marker vimentin in the Wagner and Lin cohorts. Concerned that this may imply a mis-classification of stromal cells as *Unknown*/*Other*, we next contrasted the expression of all *Unknown*/*Other* cells against cells assigned as stromal, and found significantly decreased stromal marker expression and up-regulation of epithelial cell markers such as pan-Cytokeratin (Figure 4A-E).
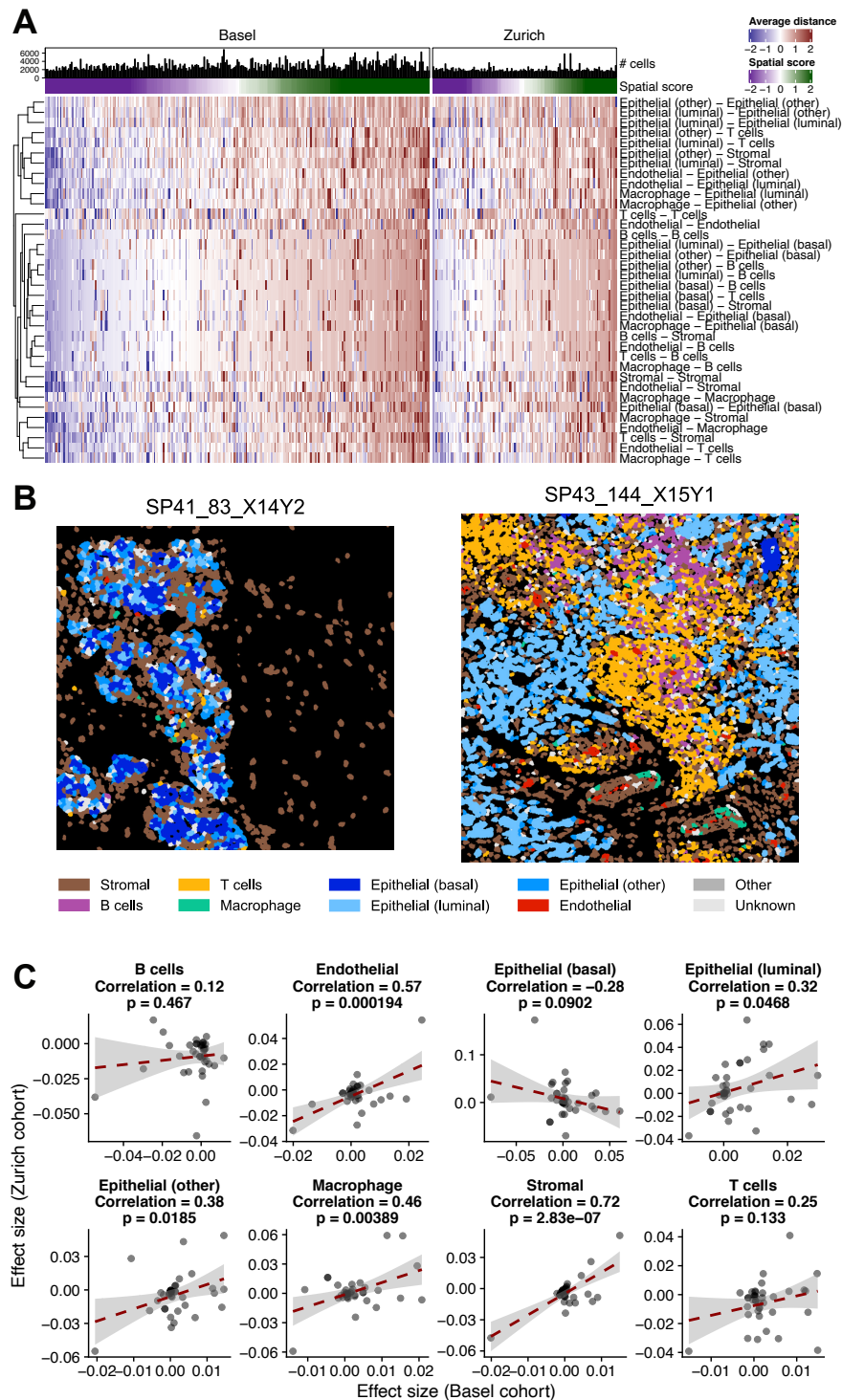
Figure 3: Automated quantification of immune influx and spatial heterogeneity from IMC identifies reproducible phenotypic perturbations across cohorts. **A** The average distance between each cell type pair per core tracks with the inferred spatial heterogeneity score. **B** Two representative cores with a low spatial heterogeneity score (left) indicating high levels of inter-cell type mixing, and a high spatial score (right) indicating a highly compartmentalized phenotype, coloured by Astir cell type assignment. **C** By regressing the per-core mean expression of each protein on the core's spatial heterogeneity score in a cell type specific manner, we can compare the effect sizes between cohorts. This leads to a significantly correlated effect sizes between cohorts (Basel and Zurich), implying reproducible phenotypic changes associated with spatial tumour architecture.
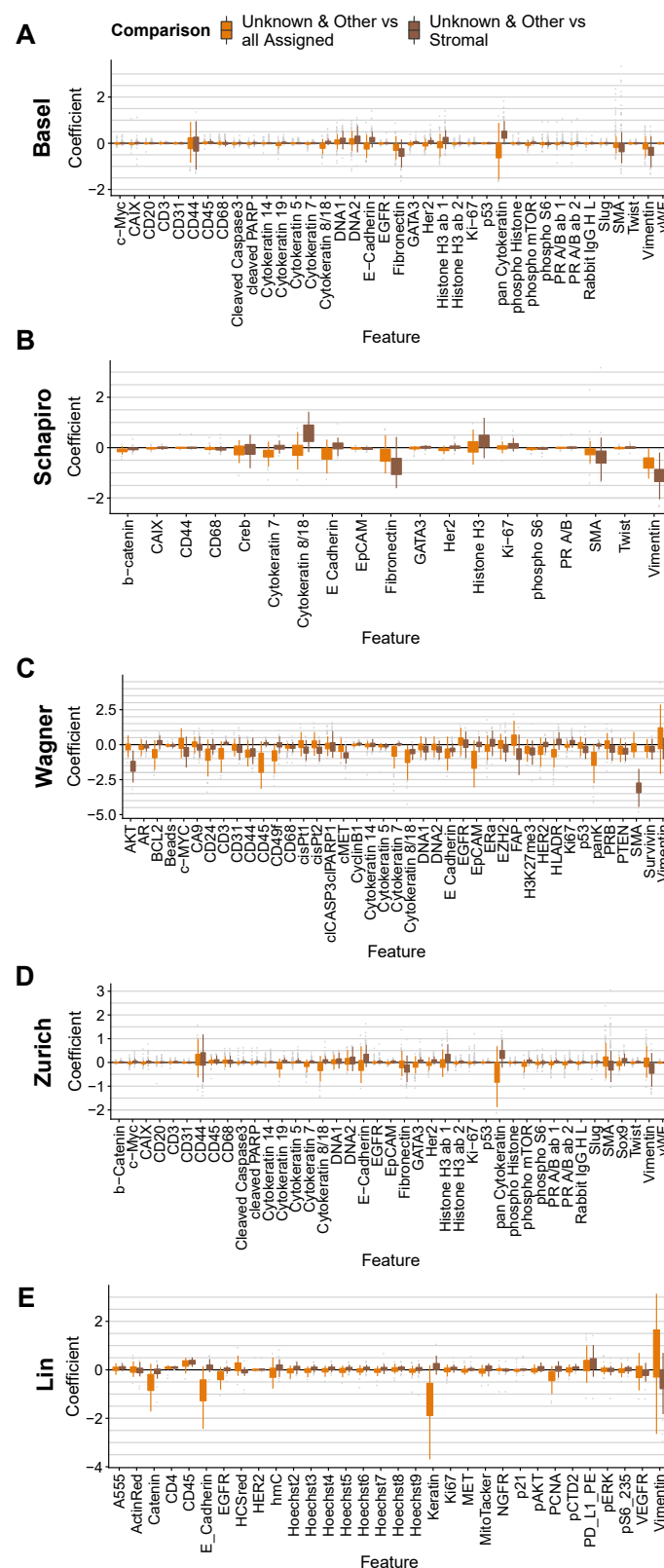
9

Figure 4: Analysis of *Unknown* and *Other* cells.

Figure 4: Boxplots showing the distribution of the estimate of a linear model contrasting the expression of each protein in *Unknown* and *Other* cells compared to all assigned cells, and all *Unknown* and *Other* cells compared to stromal cells for the Basel (**A**), Schapiro (**B**), Wagner (**C**), Zurich (**D**), and Lin (**E**) cohorts.

Together, this analysis suggests those cells classified as *Unknown* /*Other* are not a distinct cell type identifiable from over-expression of a particular marker combination, but could instead represent an intermediate stromal-epithelial state, or is due to either mis-segmentation of neighbouring cells or non-specific antibody staining. To investigate whether the latter possibility was responsible for the assignment of *Unknown* and *Other* cells, we randomly selected a set of 20 cores from the Basel and Zurich cohort and coloured each image by the maximum assignment probability (Figure S54). Visual inspection indeed demonstrates general lower assignment probabilities at the tissue edge which could be related to reduced antibody staining quality and which we consequently conclude is one factor influencing the assignment of cells to *Unknown* /*Other* types by Astir.

DISCUSSION

Here we have introduced Astir, a novel probabilistic method for highly multiplexed imaging and single-cell proteomics designed to classify single cells into distinct cell types based on prior information of relevant marker proteins. While we have applied Astir to three IMC, one suspension CyTOF, and one CyCIF datasets, Astir may equally be applied to other expression assays including MIBI [8] and highly-multiplexed FISH. To our knowledge, Astir is the first dedicated tool for systematic and scalable assignment of cell types in these data modalities, where the use of classical referenced-based supervised learning approaches is not possible due to experiment-specific antibody panels.

As we have shown, an additional advantage of Astir is that the user can specify the probability threshold at which a cell is considered *Unknown* in type, thereby allowing for flexibility in the level of certainty required for the analysis at hand. Such an approach is not currently possible with any other method, which typically make a "hard" assignment of all given cells to a cell type regardless of accuracy. Similarly, while performing gene set enrichment analysis for the interpretation of clusters, it is theoretically possible to determine an enrichment score threshold at which a cluster is assigned to an *Unknown* type, identifying a relevant threshold that can be interpreted as a probability is challenging.

Although Astir is primarily designed for highly multiplexed imaging technologies, we do not currently make use of the spatial locations of cells for the purposes of cell type assignment. Future work may incorporate custom spatially-informed priors that have recently been incorporated into spatial clonal inference [29]. For example, in the context of breast cancers a cell adjacent to a large empty area interior to a duct (the lumen) is an epithelial cell. However, the spatial mapping of such anatomical structures is exceptionally challenging and we leave the incorporation of such anatomically-informed priors as future work.

As the volume of single-cell multiplexed imaging and proteomic data continues to grow, further phenotypic states will become apparent as both the number of cells and markers profiled increases. Consequently, we anticipate Astir becoming an additional and highly-used method in the interpretation of such data modalities for the discovery of novel cellular mechanisms, clinical biomarker studies, and personalized medicine.

METHODS

ASTIR CELL TYPE MODEL

The Astir cell type model requires two inputs: a protein expression matrix and a list of marker genes for the cell types of interest. The expression matrix is an N by G matrix of protein expression $\mathbf{Y}$ for N single cells and G proteins. These data could be generated by single-cell assays that measuring protein abundance, such as mass cytometry and imaging mass cytometry (post segmentation).

The list of marker genes for all cell types of interest is required as a *yaml* file but is re-coded into a binary matrix $\boldsymbol{\rho}$ where $\rho_{gc} = 1$ if protein g is a marker for cell type c and 0 otherwise. Next, the conditional probability that a cell is of a certain type is specified based on the assumption that cell type marker proteins are more highly expressed in their respective cell types than other cells. Thus the expected expression of protein g

in cell n given that that cell z_n is of cell type c can be formulated as

$$\mathbb{E}[\ \overbrace{y_{ng}}^{\substack{\text{Expression of}\\ \text{protein } g\\ \text{in cell } n}}\ |\ \underbrace{z_n = c}_{\substack{\text{Given we}\\ \text{assign cell } n\\ \text{to cell type } c}}\ ] := \mu_{ngc} = \exp\left(\underbrace{\alpha_g}_{\substack{\text{Base expression}\\ \text{protein } g}} + \overbrace{\rho_{gc} \times \delta_{gc}}^{\substack{\text{If } g \text{ marker for } c\\ \text{add } \delta_{gc} > 0}} + \underbrace{\sum_{h=1}^{H} \beta_{hg} x_{nh}}_{\substack{\text{Contribution from additional}\\ \text{covariates (e.g. patient)}}}\right)$$

where α_g is the base mean of protein g across all cell types and $\delta_{gc} > 0$ is the overexpression of protein g in cell type c. Therefore, when protein g is a marker for cell type c, the conditional expectation is modelled as being $\exp(\rho_{gc} \times \delta_{gc}) = \exp(\delta_{gc})$ times higher than in a cell type for which g is not a marker protein (i.e. $\rho_{gc} = 0$). Note that the overexpression parameter δ is cell type specific, allowing certain cell types to overexpress a given marker more than others (e.g. multiple types of macrophage overexpress *CD68* to different degrees). Finally, additional cell-specific covariates (such as cell cycle stage or patient of origin) may be incorporated through H covariates via the design matrix $\mathbf{X}$.

Rather than assume the expression of marker genes are conditionally independent in a given cell type as in previous work [12], we introduce a novel co-variance structure that requires the expression of proteins to be correlated within a given cell type. Let $\boldsymbol{\Sigma}^{(c)}$ be the $G \times G$ matrix of pairwise covariances between proteins in cell type c which we specify as

$$\boldsymbol{\Sigma}^{(c)}_{g,g'} := \sigma_g \sigma_{g'} \left(r_c^2 \rho_{gc} \rho_{g'c} + \delta_{g,g'}(1 - r_c^2 \rho_{gc} \rho_{g'c}) \right)$$

where $\delta_{g,g'} = 1$ if $g = g'$ and 0 otherwise.

This ensures the covariance has three desirable properties:

1. If g and g' are both markers for cell type c and $g \neq g'$ then the product $\rho_{gc}\rho_{g'c} = 1$ and the covariance between g and g' is $\sigma_g \sigma_{g'} r_c^2$, where r is a cell-type specific gene-gene correlation. Note we constrain $r \in [0, 1]$ since if two genes are a marker for a cell type they should not be anti-correlated.

2. If either g or g' aren't markers for c then and $g \neq g'$ then $\rho_{gc}\rho_{g'c} = 0$ and the covariance is 0.

3. If $g = g'$ then the covariance becomes the variance σ_g^2.

If we additionally define $\boldsymbol{\mu}_{nc} = [\mu_{n1c}, \ldots, \mu_{nGc}]$ then we define the likelihood for a given cell n expression vector $\mathbf{y}_n$ as

$$p(\mathbf{y}_n | z_n = c, \boldsymbol{\sigma}, \mathbf{r}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\delta}, \boldsymbol{\beta}) = \mathcal{N}\left(\mathbf{y}_n | \boldsymbol{\mu}_{nc}, \boldsymbol{\Sigma}^{(c)}\right)$$

where $\mathcal{N}$ is the multivariate normal density. Then assuming the conditional likelihoods for each cell are independent we can write the likelihood for the entire expression matrix as

$$p(\mathbf{Y} | \mathbf{z}, \boldsymbol{\sigma}, \mathbf{r}, \boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\delta}, \boldsymbol{\beta}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{y}_n | \boldsymbol{\mu}_{nz_n}, \boldsymbol{\Sigma}^{(z_n)}\right).$$

We append an additional column to ρ filled with constant 0, to signify the existence of an *Other* cell type that does not overexpress any marker and thus acts to capture all cells that do not fit a pre-defined cell type. Finally, we specify a prior for cell type assignment $\gamma_c = p(z_n = c)$ on the simplex such that $\sum_{c=1}^{C} \gamma_c = 1$ with a hierarchical prior specification $\gamma_1, \ldots, \gamma_C \sim \text{Dirichlet}(C, \ldots, C)$.

ASTIR CELL TYPE INFERENCE

The overall object of inference is the posterior distribution of cell type assignments $p(\mathbf{z}|\mathbf{Y})$. Since this is impossible to calculate analytically, we resort to variational inference introducing variational distributions to minimize

$$\min_{\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\sigma},\boldsymbol{\delta},\mathbf{r},\boldsymbol{\gamma}} \text{KL} \left(\underbrace{q(\mathbf{z}|\boldsymbol{\lambda})}_{\text{Variational distribution}} || \underbrace{p(\mathbf{z}|\mathbf{Y},\boldsymbol{\alpha},\boldsymbol{\sigma},\boldsymbol{\delta},\mathbf{r},\boldsymbol{\gamma})}_{\text{Posterior distribution}} \right)$$

where KL is the Kullback-Leibler divergence. Rather than explicitly instantiate variational distributions $q(z_n = c) = \psi_{nc}$ for each parameter, we use recent results from amortized inference [21] to propose a variational distribution of the form

$$\overbrace{p(z_n = c|\mathbf{Y})}^{\substack{\text{Probability cell } n \text{ of} \\ \text{type } c, \text{ given data } \mathbf{Y}}} \approx \overbrace{q(z_n = c|\boldsymbol{\lambda})}^{\text{Variational approximation}} = \text{nn}(\mathbf{y}_n; \boldsymbol{\lambda})$$

where nn is a neural network that maps the input data $\mathbf{y}_n$ for cell n to a simplex of probabilities over cell types, where λ are the (variational) parameters of the recognition neural network.

ASTIR CELL TYPE IMPLEMENTATION

Astir is implemented as a Python package using Pytorch for inference. Given the non-convexity of the minimization problem, the model is trained using five random initializations for a small number of epochs, following which the instantiation with the lowest ELBO is selected for continued training.

The cell type model parameters are initialized as follows:

- $\boldsymbol{\rho}$ is a matrix of size $G \times (C+1)$. Each entry $\rho_{gc} = 1$ if protein g is a marker for cell type c, and $\rho_{gc} = 0$ otherwise.

- $\boldsymbol{\alpha}$ is a vector of size G. To initialize $\boldsymbol{\alpha}$, the following procedure is used: (i) a copy of the input expression matrix is created standardized to have empirical mean 0 and standard deviation 1, (ii) for each specified cell type c, a set number N_L of "putative cells" of type c is inferred as the N_L cells with the highest standardized expression of the markers of cell type c, with $N_L = 10$ by default, (iii) for each protein g, the set of all cell types for which g is not a marker is collated, and the expression of g in the "putative cells" belonging to these cell types is averaged, with the log value of this average forming the initial value of α_g.

- $\boldsymbol{\sigma}$ is a vector of size G. Element at σ_g is initialized to be the standard deviation of expression value of protein p across all cells.

- $\boldsymbol{\delta}$ is a matrix of size $G \times (C+1)$. Each element at δ_{gc} is the overexpression of protein g in cell type c. Its initial log value is sampled from normal distribution with mean 0 and variance 0.1

- $\boldsymbol{r}$ is a matrix of size $G \times (C+1)$. It is a cell-type specific gene-gene correlation which is constrained within $[0, 1]$ and is initialized to be 0.5.

- The set of all parameters used in the recognition network including weights and bias are randomly initialized by PyTorch's nn module. By default, the recognition network has two hidden layers of size 20 with a leaky ReLU activation [30].

DIAGNOSTICS

The Astir package comes with a set of functions to perform post-fit diagnostics on the outputs of the model. This diagnostic ensures that all cell types express their marker proteins at significantly higher levels than other cell types, and flags all marker-cell types combinations for which this is violated. To identify cell types that violate this a t-test is run on the expression levels of a marker g between cells for which g is a marker and cells for which g is not a marker, and an error logged of the marker is not expressed significantly higher (at $\alpha = 0.05$ significance) in the cell type for which it is a marker.

IMC DATA PROCESSING & ASTIR CELL TYPE ASSIGNMENT

For the Basel and Zurich IMC cohorts, processed and spillover-corrected data were already available and normalized via an arcsinh transformation [31, 32] with a co-factor of 5 and winsorized to [0%, 99.9%]. For the Schapiro IMC cohort, the raw image files and masks were used to summarize to single-cell expression, where the expression of each protein in each cell was computed as the mean expression of that protein

across all pixels belonging to that cell (code at https://github.com/camlab-bioml/schapiro-2017). The resulting data was normalized as per Basel and Zurich cohorts. Expression data from the Wagner cohort was similarly normalized with an arcsinh transformation and cofactor of 5. For the Lin cohort an arcsihn transformation with a cofactor of 100 was employed, with expression values winsorized to [1%, 99%]. The markers used by Astir to assign cell types for all cohorts are listed in Figure S1.

BENCHMARKING AGAINST EXISTING CELL-TYPE IDENTIFICATION APPROACHES

We compared the performance of Astir to that of FlowSOM [16], ClusterX [17] and Phenograph [15] using cytofkit [17], all of which are existing methods commonly used for the identification of cell types for flow and mass cytometry. For FlowSOM, cluster sizes were chosen to match the expected number of cell types (e.g. 8 for the Basel cohort) and varied such that potential cell states or Other cells could also be captured. For FlowSOM we used cluster sizes k = 4, 7, 8, 12, 15, and 20, for Phenograph we used nearest neighbour values 20, 30, 40 and 50, and for ClusterX we used default values.

Upon having assigned each cell to a cluster we used a gene set variation analysis [27] using the same cell type markers listed in Figure S1 to assign each cluster to a cell type. We did this by calculating the average enrichment for each cluster and cell type. For instance, for cluster one we calculated the average enrichment for T cells by computing the mean gsva enrichment across all cells in cluster one. We then repeated this for all target cell types and assigned the cell type that had the highest average enrichment to cluster one. We repeated this process for all clusters.

We compared the performance of the methods tested using an interpretability and accuracy score. Given that the true clusters are Unknown and that the goal of any cell type assignment is to generate a single interpretable assignment for each cell type, we created a score to measure the number of clusters per cell type. This score penalizes a method for assigning multiple cell types per cell type or assigning no clusters to a cell type by -1, and rewards a method for assigning exactly one cluster to each cell type with a score of +1.

SPATIAL SIGNATURE ANALYSIS

To identify novel spatial signatures enabled by Astir's cell type assignment, we computed the mean distance from every cell type to every other cell type on the Basel and Zurich cohorts [23] by computing the pairwise distance from the centre of each cell for all possible cell type pairings followed by taking the mean. To enable distance comparisons across cell type pairs, the distances were standardized (mean 0 variance 1) in a cell-type-pair specific manner. Only cores with at least 1,000 cells present were retained to prevent low numbers of cells skewing distance calculations. Any missing distance values due to missing cell types from cores were imputed to the median of observed values for given cell-type-pairs.

IDENTIFICATION OF *Unknown* CELL TYPES

To confirm the status of cells assigned as *Other* and *Unknown*, we conducted a differential expression analysis for the entire protein panel available for each cohort. We did this by fitting a linear model using the `lm` function in `R` to the expression of each protein and regressed on whether a cell was assigned as *Other* /*Unknown* or assigned as one of the specified cell types. We repeated this, splitting the data into *Other* /*Unknown* and Stromal cells and regressed on these categories as described above.

SOFTWARE PACKAGES USED

With the exception of the single-cell spatial image plots which were created in Python using matplotlib [33] and seaborn [34] all plots were made using custom R scripts. tSNE plots were created with FasttSNE [35] and visualized using ggplot [36] and the wesanderson package for the cohort colour palette. All heatmaps were created using the ComplexHeatmap R package [37], alluvial plots were made with ggalluvial [38] and correlations were calculated with the Hmisc package. All scripts to recreate the analysis are available online at https://www.github.com/camlab-bioml/astir-manuscript.

REFERENCES

1. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357,** 661–667 (2017).

2. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546,** 431–435 (2017).

3. Azizi, E. *et al.* Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174,** 1293–1308 (2018).

4. Bernard, V. *et al.* Single-cell transcriptomics of pancreatic cancer precursors demonstrates epithelial and microenvironmental heterogeneity as an early event in neoplastic progression. *Clinical Cancer Research* **25,** 2194–2205 (2019).

5. Foley, J. W. *et al.* Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. *Genome research* **29,** 1816–1825 (2019).

6. Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nature methods* **16,** 987–990 (2019).

7. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* **11,** 417–422 (2014).

8. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nature medicine* **20,** 436–442 (2014).

9. Spitzer, M. H. & Nolan, G. P. Mass cytometry: single cells, many features. *Cell* **165,** 780–791 (2016).

10. Gerdes, M. J. *et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences* **110,** 11982–11987 (2013).

11. Lin, J.-R. *et al.* Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *Elife* **7,** e31657 (2018).

12. Zhang, A. W. *et al.* Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature methods* **16,** 1007–1015 (2019).

13. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nature methods* **15,** 359–362 (2018).

14. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature methods* **14,** 1083–1086 (2017).

15. Levine, J. H. *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162,** 184–197 (2015).

16. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A* **87,** 636–645 (2015).

17. Chen, H. *et al.* Cytofkit: a bioconductor package for an integrated mass cytometry data analysis pipeline. *PLoS computational biology* **12,** e1005112 (2016).

18. Finak, G. *et al.* Standardizing flow cytometry immunophenotyping analysis from the Human ImmunoPhenotyping Consortium. *Scientific reports* **6,** 1–11 (2016).

19. Zhang, A. W. & Campbell, K. R. Computational modelling in single-cell cancer genomics: methods and future directions. *arXiv preprint arXiv:2005.01549* (2020).

20. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. en. *Nat. Immunol.* **20,** 163–172 (Feb. 2019).

21. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

22. Paszke, A. *et al. Automatic differentiation in PyTorch* in *NIPS-W* (2017).

23. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578,** 615–620 (2020).

24. Wagner, J. *et al.* A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* **177,** 1330–1345 (2019).

25. Schapiro, D. *et al.* histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods* **14,** 873 (2017).

26. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9,** 2579–2605 (2008).

27. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics* **14,** 7 (2013).

28. Keren, L. *et al.* A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **174,** 1373–1387 (2018).

29. Elyanow, R., Zeira, R., Land, M. & Raphael, B. STARCH: Copy number and clone inference from spatial transcriptomics data. *Physical Biology* (2020).

30. Xu, B., Wang, N., Chen, T. & Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015).

31. Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332,** 687–696 (2011).

32. Diggins, K. E., Ferrell Jr, P. B. & Irish, J. M. Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. *Methods* **82,** 55–63 (2015).

33. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in science & engineering* **9,** 90–95 (2007).

34. Waskom, M. *et al.* seaborn: v0. 5.0 (November 2014). *Zenodo, doi* **10** (2014).

35. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature methods* **16,** 243–245 (2019).

36. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-3-319-24277-4. https://ggplot2.tidyverse.org (Springer-Verlag New York, 2016).

37. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).

38. Brunson, J. C. ggalluvial: Layered Grammar for Alluvial Plots. *Journal of Open Source Software* **5,** 2017 (2020).

DATA AVAILABILITY

The data for the Basel and Zurich cohorts used in this study are available through Zenodo at https://zenodo.org/record/3518284. The data for the Schapiro cohort are available at http://www.bodenmillerlab.org/research-2/histoCAT/. Masks corresponding to the different segmentatations are available upon request from the authors. The data for the Wagner cohort are available through Mendeley Data at https://data.mendeley.com/datasets/gb83sywsjc/1. The data for the lin-cycif cohort are available through the HMS LINCS publication page https://lincs.hms.harvard.edu/lin-elife-2018/.

AUTHOR CONTRIBUTIONS

K.R.C. conceived the study and machine learning models. M.J.G. conceived the analysis plan, performed data analysis and created figures. J.H. and S.L. coded astir python package and performed additional analysis. M.J.G., H.W.J., S.L, J.H, and K.R.C. wrote the manuscript.

ACKNOWLEDGEMENTS

SUPPLEMENTAL FIGURES