Modelling cell-cell interactions from spatial molecular data with spatial variance component analysis

Damien Arnol[1], Denis Schapiro[4,5], Bernd Bodenmiller[4], Julio Saez-Rodriguez[1,2,#], Oliver Stegle[1,3,#]

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD, Hinxton, Cambridge, UK
2. Joint Research Center for Computational Biomedicine, RWTH Aachen University, Faculty of Medicine, Pauwelsstrasse 19, D-52074 Aachen, Germany
3. European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany
4. Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
5. Life Science Zurich Graduate School, ETH Zurich and University of Zurich, Zurich, Switzerland

Corresponding authors

Correspondence should be addressed to: Julio Saez-Rodriguez (saezrodriguez@gmail.com) or Oliver Stegle (stegle@ebi.ac.uk)

Method: SVCA (spatial variance component analysis)

Abstract

Technological advances allow for assaying multiplexed spatially resolved RNA and protein expression profiling of individual cells, thereby capturing physiological tissue contexts of single cell variation. While methods for the high-throughput generation of these spatial expression profiles are increasingly accessible, computational methods for studying the relevance of the spatial organization of tissues on cell-cell heterogeneity are not established. Here, we present *spatial variance component analysis (SVCA),* a computational framework for the analysis of spatial molecular data. SVCA enables quantifying the effect of cell-cell interactions, as well as environmental and intrinsic cell features on the expression levels of individual genes or proteins. In application to a breast cancer Imaging Mass Cytometry dataset, our model allows for robustly estimating spatial variance signatures, identifying cell-cell interactions as a major driver of expression heterogeneity. Finally, we apply SVCA to high-dimensional imaging-derived RNA data, where we identify molecular pathways that are linked to cell-cell interactions.

Introduction

Experimental advances have enabled assaying RNA and protein abundances of single cells in spatial contexts, thereby allowing to study single cell variation in tissues. Already, these technologies have delivered new insights into tissue systems and the sources of transcriptional variation (Battich, Stoeger, & Pelkmans, 2013; Bodenmiller, 2016), with potential use as biomarkers for human health (Bodenmiller, 2016). Spatial expression variation can reflect interactions between adjacent cells, or can be caused by cells that migrate to specific locations in a tissue to perform their functions (e.g. immune cells).

Different technologies allow for generating spatially resolved expression profiles. Imaging Mass Cytometry (IMC) (Giesen et al., 2014) and Multiplexed Ion Beam Imaging (MIBI) (Angelo et al., 2014) rely on protein labeling with antibodies coupled with metal isotopes of specific masses followed by high-resolution tissue ablation and ionisation. IMC enables profiling of up to 27 targeted proteins with subcellular resolution. Other methods such as MxIF and CycIF use immunofluorescence for protein quantification of tens of markers at a time (Gerdes et al., 2013; Lin, Fallahi-Sichani, & Sorger, 2015). Increasingly, there also exist optical imaging-based assays to measure single cell RNA levels. Mer-FISH and seq-FISH use a combinatorial approach of fluorescence-labeled small RNA probes to identify and localise single RNA molecules (Chen, Boettiger, Moffitt, Wang, & Zhuang, 2015; Gerdes et al., 2013; Lin et al., 2015; Shah, Lubeck, Zhou, & Cai, 2017), which allows for a larger numbers of readouts (currently between 140 and 250). Even higher-dimensional expression profiles can be obtained from spatial expression profiling techniques such as Spatial Transcriptomics (Ståhl et al., 2016), which currently however do not offer single cell resolution and are therefore not adequate to study cell-to-cell variation.

The availability of spatially resolved expression data represents an unprecedented opportunity to disentangle largely unexplored sources of single cell variations: i) intrinsic sources of variation due to difference in cell types or states (e.g. cell-cycle state), ii) environmental effects from the cell microenvironment and iii) interactions between adjacent cells. Yet, although experimentally spatial omics profiles can be generated with high throughput, the required computational strategies for interpreting the resulting data are only beginning to emerge. On the one hand, there exist methods to cluster the cells based on their expression profiles (Achim et al., 2015), and there also exist statistical tests to assess

the overall relevance of the spatial topology on gene expression (Svensson, Teichmann, & Stegle, n.d.). However, these methods do not allow for directly assessing cell-cell interactions. On the other hand, there are methods that explain dependencies between cells based on cell-type assignments and predefined cellular neighbourhoods (Schapiro et al., 2017; Schulz et al., 2018). While these approaches provide qualitative insights into interactions between cell types, they do not allow for quantifying the impact of spatial effects on individual genes or proteins. Finally, there are regression-based models to assess interaction effects based on predefined features of cell neighbourhood (Battich, Stoeger, & Pelkmans, 2015; Goltsev et al., 2018), which however require discretization steps (see **Methods** for a detailed comparison).

Here, we present *Spatial variance component analysis (SVCA),* a computational framework to model spatial sources of variation of individual genes or proteins. SVCA allows for decomposing the sources of variation into intrinsic effects, environmental effects and cell-cell interactions. The model is parameterized based on continuous expression profiles of individual cells, and in particular avoids the need to define discrete cell types and microenvironmental variables. We illustrate SVCA using data from different technologies and biological domains, including IMC proteomics profiles data from human cancer tissue and spatial single-cell RNA profiles from the mouse hippocampus generated using seqFISH. Across these applications, we find that cell-cell interactions substantially contribute to gene expression variability, and we identify biologically relevant genes and pathways that participate in these interactions.

Results and Discussion

SVCA considers random effect components that capture i) *intrinsic* sources of variation due to differences in cell types or states, ii) *environmental* sources of variation due to local extracellular factors iii) source of variations due to *cell-cell interactions* (**Fig. 1a**). The contribution of these factors on expression variability is modelled using additive random effects (**Fig. 1b**), where we borrow covariance models that were originally designed for social genetic effects (Baud et al., 2017) to capture cell-cell interactions. Importantly, SVCA does not require assigning cells to discrete types, but instead is based on a continuous measure of cell-cell similarities that are directly estimated from cell expression profiles (**Fig. S1**). The model also circumvents the need to define discrete neighbourhoods but instead weights interactions between pairs of cells as a function of their distance (**Fig. S1**). Once

fitted, SVA allows for decomposing the sources of variation of individual genes and proteins (**Fig. 1c**) and the model can be used to assess the significance of these respective components (**Methods**).

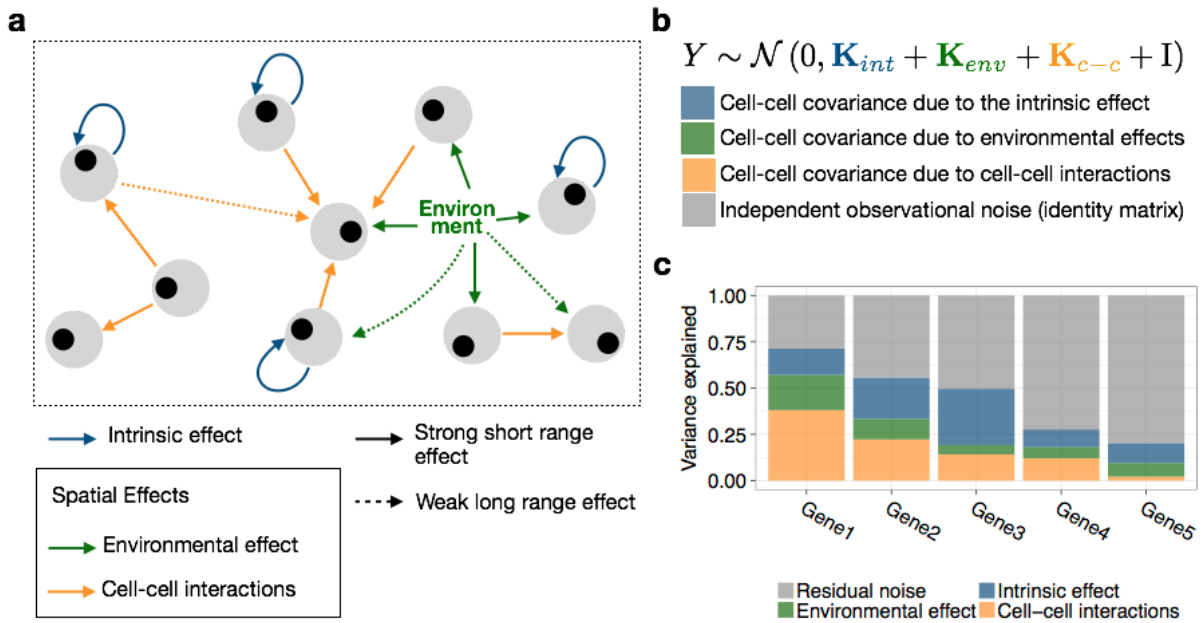

Figure 1 | Spatial variance component analysis (SVCA): approach and overview. (**a**) SVCA decomposes the variability of individual genes and proteins into i) cell intrinsic effects (due to differences in intrinsic cell type or state, blue), ii) general environmental effects that capture expression differences due to local extracellular factors (green) and iii) a cell-cell interaction component that captures differences in expression level attributable to different cellular composition of a cell's neighbourhood (yellow). (**b**) SVCA builds on a random effect framework to model additive contributions of these components**.** See **Fig. S1** and **Methods** for details on the definition of the corresponding random effects. (**c**) SVCA output: gene-level break down of the proportion of variance attributable to different components.

Initially, we used simulated data to assess the statistical calibration and the accuracy of SVCA variance estimates. Briefly, we generated expression profiles by sampling from the model, using empirical parameters derived from 11 real datasets (**Methods**), including the empirical position of cells, cell state covariance and ranges of fractions of variance explained by different components that mimic those observed in real data. First, we simulated expression profiles with no interaction effects to assess the calibration of the corresponding test, finding that the model yields conservative estimates (**Fig. 2a**). We also assessed the detection power of the test when simulating cell-cell interactions that explain increasing fractions of the gene expression variance (**Fig. 2c**), and when varying the number of cells in the dataset (**Fig. 2d**). We also compared the estimated variance components for cell-cell

interactions with the simulated variance components, again observing that the model estimates are conservative (**Fig. 2b**). To illustrate the interpretation of the environmental effect as modelling non-specific spatial sources of variation, and to better understand the empirical identifiability of cell-cell interactions versus environmental effects, we also compared the estimates of the full model with a a reduced model that does not account specifically for cell-cell interactions. These results suggest that estimates of cell-cell interactions are not confounded by other spatial sources of variation and are therefore conservative. Taken together, these results demonstrate that SVCA can be used to estimate and test for spatial drivers of single cell variability, in particular cell-cell interactions.

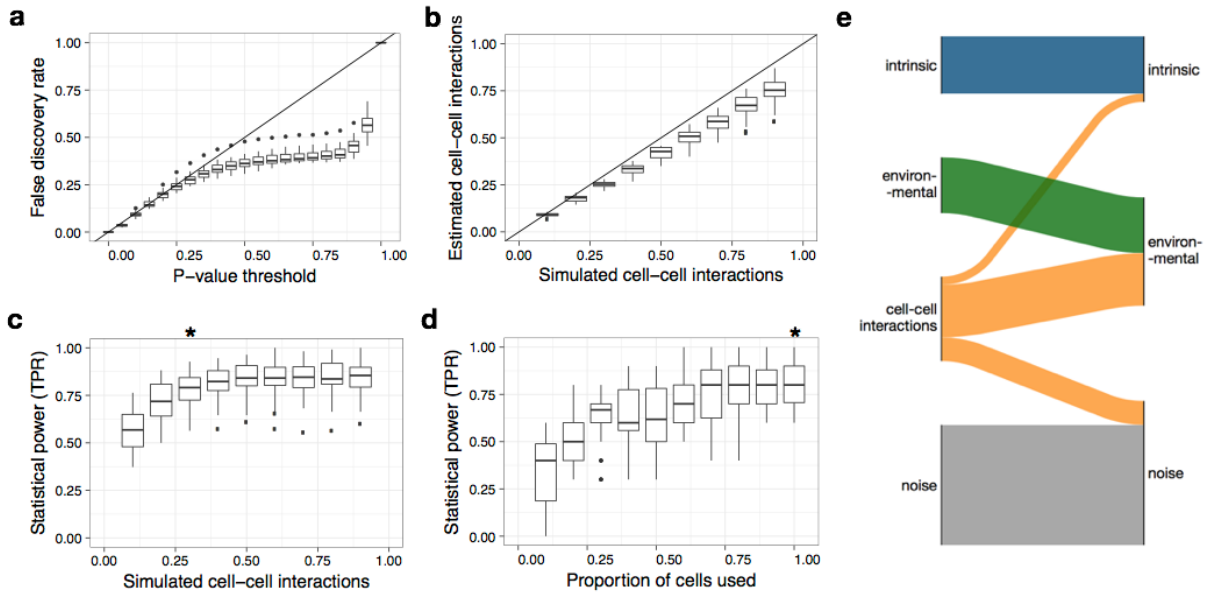

Figure 2 | Validation of cell-cell interaction tests and variance components using simulated data.
(a) Empirical false discovery rate for the cell-cell interaction test using data simulated from the null, without cell-cell interactions. Shown is the empirical false discovery rate (FDR) as a function of the applied P value threshold. **(b)** Fraction of variance due to cell-cell interactions estimated by SVCA when varying the simulated fraction of variance explained by cell-cell interactions (**Methods**). **(c,d)** Statistical power of the test for cell-cell interactions (at family-wise error rate <1%), when varying the simulated fraction of variance explained by cell-cell interactions (**c**) and when considering different subsets of cells from the full dataset for model fitting (**d**). The full dataset with all cells was used for model fitting in **c** (indicated using the asterix symbol). In each panel, boxplots display the distribution of results across 26 proteins. Rates in panel **a**, **c**, **d** (True Positive Rate - TPR and False Discovery Rate - FDR) are computed for each protein, aggregated across 110 simulations (11 images times 10 repeat experiments). Similarly, panel **b** depicts average variance estimates across the same 110 simulations for each protein. (**e**) Sankey plot displaying how the variance explained by

cell-cell interactions is captured by the other terms when omitting interaction effects in the model. Bar height denotes estimated variance fractions (lef: SVCA; right: reduced model). Linking edges indicate the redistribution of variance estimates between the full and the reduced model. Both models were fitted on the same simulated data for 11 images and 26 proteins with cell-cell interactions explaining 30% of the variance of simulated expression levels. (**Methods**).

Application of SVCA to spatial proteomics data of breast cancer tissues

Next, we applied SVCA to a recent Imaging Mass Cytometry (IMC) dataset from human breast cancer, consisting of 52 breast biopsies from 27 breast cancer patients with variable disease grade and from different cancer subtypes, sampled from different tumour locations (Schapiro et al., 2017). SVCA revealed substantial differences of the importance of cell-cell interaction components across proteins, explaining up to half of the expression variability (**Fig 3a**). Immune cell markers in particular were enriched among the set of proteins with the strongest cell-cell interaction effects: CD44, CD20, CD3 and CD68 had significant cell-cell interaction components in 35, 34, 34 and 38 out of the 52 images respectively (FDR<1%, Benjamini-Hochberg adjusted, **Fig. 3a**). We used bootstrapping to confirm the robustness of the variance estimates (**Fig. S4**), and we observed that a variance component model that accounts for cell-cell interactions was able to more accurately impute gene expression profiles than models that ignore such effects (**Fig. 3b, Fig. S2)**.

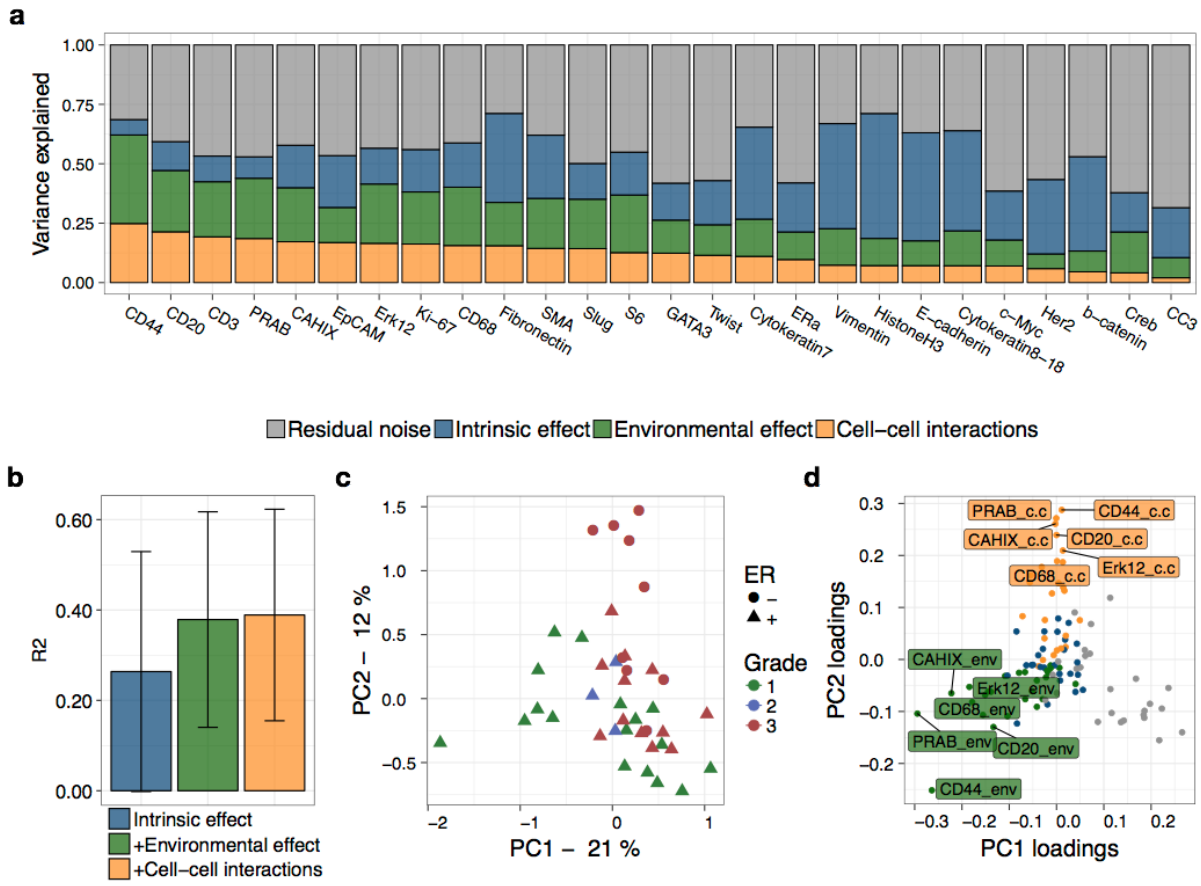

Figure 3 | Application of SVCA to 52 breast cancer samples profiled using IMC. (**a**) Bottom panel: SVCA signatures for 26 proteins. Shown are averages of the fraction of variance explained by intrinsic effects, environmental effects and cell-cell interactions, across 52 images. Proteins are ordered by the magnitude of the cell-cell interaction component. Top panel: number of images with significant cell-cell interaction effects for the corresponding proteins (FDR<1%, **Methods**, **Fig S5**). (**b**) Accuracy of SVCA and models of lower complexity to predict gene expression markers for hold out cells (5 fold cross validation, masking cells). Shown are average coefficients of determination (r^2) between predicted and observed gene expression profiles, averaged across proteins and images. Error bars correspond to plus and minus one standard deviation across images and proteins. (**c**) First two principal components for 42 clinically annotated images, calculated based on the spatial variance signature, with individual images coloured by tumour grade. (**d**) Loadings for the first principal components as in **c**, depicting the relevance of individual proteins and types of variance components.

We also observed substantial differences of the estimated spatial variance signatures between individual images (**Fig. S3**), motivating investigating the relationship between these variations and clinical covariates, including tumour grade, ER status, PR status, HER2 status. Using principal component analysis (PCA), we identified tumour grade as the most important explanatory variable for differences in spatial variance components (**Fig 3c**), followed by ER status and PR status (**Fig. S6**). Inspection of the PCA loadings (**Fig. 3d**)

identified the cell-cell interaction component and the environmental component for a subset of proteins (including PR and CD44) as the most informative SVCA features affected by tumour grade. This was also evident when comparing environmental and cell-cell interaction effects across tumours of different grades (**Fig. S7**), showing a relative increase in the cell-cell interaction component in higher grade tumours. This effect could partly be linked to the underlying disorganized and irregular cellular architecture characterizing these high grade tumors, which is associated with larger cells, increased proliferation and thus potentially higher cell density in comparison to healthy breast tissue (Elston & Ellis, 1991).

RNA datasets based on imaging technologies

SVCA can be used for the analysis of data from a broad range of spatially resolved technology, including optical imaging-based assays. To explore this, we considered a mouse hippocampus datasets, profiling 240 RNA expression levels in 21 distinct brain regions of a single animal, using seqFISH (Shah et al., 2017). Leveraging the higher dimensionality of these data, we sought to identify individual genes and pathways that are most linked to cell-cell interactions.

Similarly to the IMC dataset, SVCA signatures were robust and could accurately predict gene expression levels out of sample (**Fig S13**, **S14**, **S15**). We considered the average cell-cell interaction components across images and tested for enriched pathways using Reactome (Croft et al., 2014), thereby identifying biological processes that contain genes with the largest cell-cell interaction components (**Methods**). This identified five pathways with significant enrichment (FDR<5%, considering 80 pathways with three or more genes contained in our dataset, **Methods**, **Fig. 4b, Supp. table 1**), with *transmission across chemical synapses, neurotransmitter release cycle* and *neuronal system* among the ones with highest significance. For comparison, we also considered gene set enrichment based on the average gene expression levels across images, which identified distinct GO categories with significant enrichment (*transmembrane transport* and *SLC-mediated transmembrane transport*). This confirms that spatial and abundance-based signatures yield complementary information for studying biological activity in tissues.

Similarly to results obtained on the IMC datasets, we observed differences in the spatial variance signatures across different images, which were sampled from functionally distinct regions of the hippocampus (Shah et al., 2017). Principal components of the spatial variance

signature for the dorsal region clustered together, irrespective of their CA1/CA3 location. Similarly, images from the Dentate Gyrus (DG) also clustered together, and there was some proximity between signatures from the ventral region, although with more variation between them (**Fig. 4d,e**). This is consistent with Shah et al's observation that the ventral and dorsal regions of the CA1 and the CA3 mirrored each other with respect to their cellular compositions, and that ventral regions are more heterogeneous in their cellular composition. Spatial variance signatures for intermediate regions, however, did not show much resemblance (**Fig. 4d,f**).

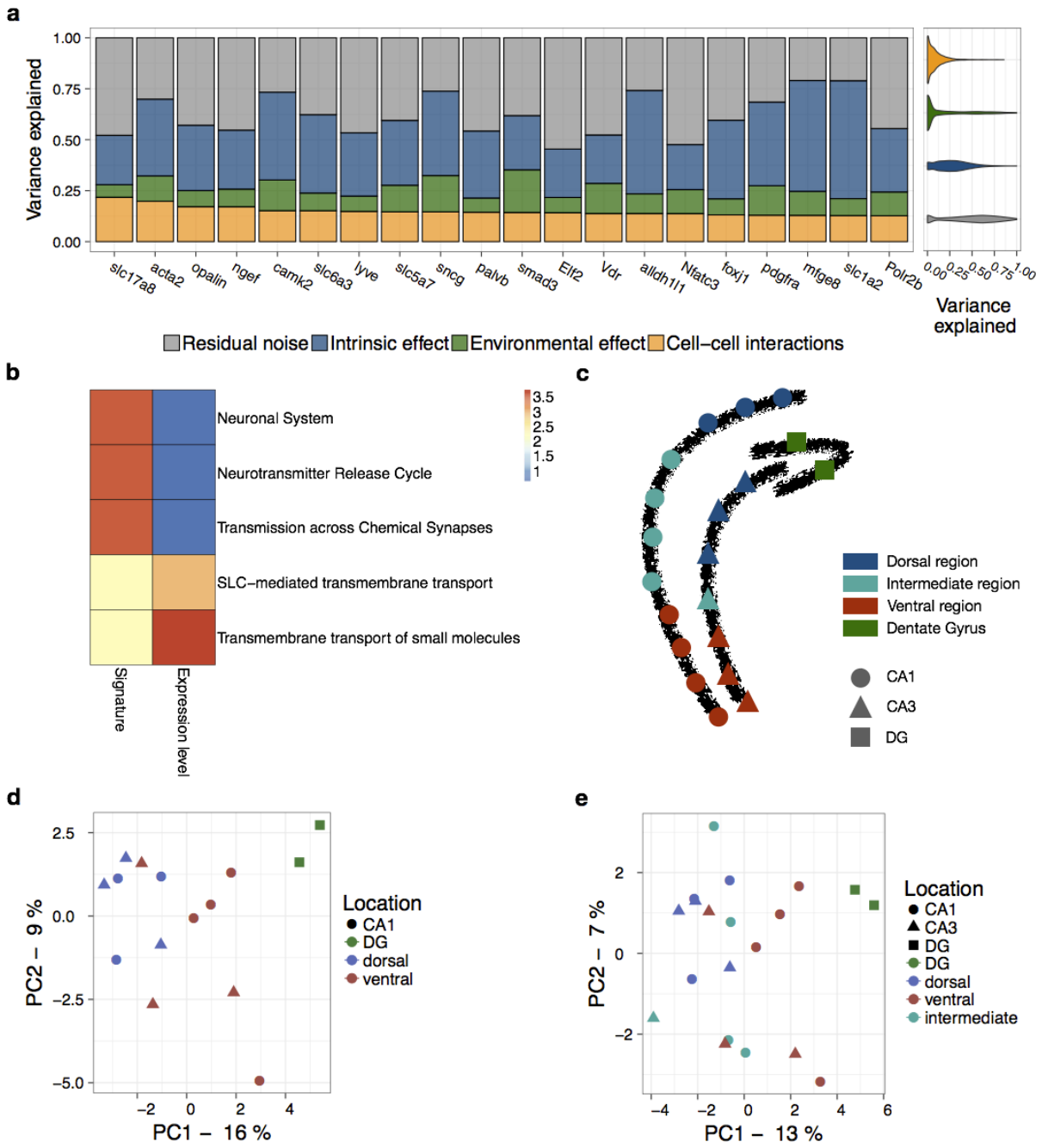

Figure 4: Application of SVCA to 20 images profiled using seqFISH.
(a) Left: SVCA signatures for the top 20 proteins with highest cell-cell interaction effects. Shown are averages of the fraction of variance explained by intrinsic effects, environmental effects and cell-cell interactions, across 20 images. Proteins are ordered by the magnitude of the cell-cell interaction component. Right: Variance estimates distribution across images and proteins for all genes (violin plots). **(b)** Top 5 enriched pathways out of 80 Reactome pathways considered using a rank-based enrichment test (based on a consensus ranking across images, **Methods**), either considering the fraction of variance explained by cell-cell interactions (left), or the expression level of individual genes (right). Colours denote statistical significance (negative log Benjamini-Hochberg adjusted P values). **(c)** Spatial organisation of the mouse hippocampus with dots corresponding to individual images. Colours and shapes denote regions using the classification as in (Shah et al., 2017). **(d)** First two principal components of the spatial variance signatures for individual images from the DG, the dorsal region and the ventral region. The colours and shapes represent the location of the biopsy in the hippocampus. **(e)** First two principal components of the spatial variance signatures for all 20 images.

Finally, we applied SVCA to a breast cancer cell culture dataset profiled using the mer-FISH technique (140 RNA expression levels) (Moffitt et al., 2016). The large spatial domain and the total number of cells imaged (more than 50,000) from a biological homogenous cell culture system enabled defining different field of views as biological replicates (N=20, 2600 cells on average). Consistent with this, the variance components of SVCA were consistent across images (**Fig. S8**, **Fig. S9**, **Fig. S10**, **Fig. S12**), with average coefficients of variation for cell-cell interactions of 20-40%, compared to typically 75-150% (**Fig. S11**) for the results obtained from the IMC and seq-fish data (**Fig. S3**, **Fig. S16**). These results further support the general applicability of the model to data from different technologies and the robustness of SVCA variance signatures.

Conclusion

We presented Spatial Variance Component Analysis (SVCA), a regression-based framework for the analysis of spatially resolved molecular expression data. Our model computes a spatial variance signature for individual mRNA or protein levels, decomposing their sources of variation into spatial and non-spatial components. Most prominently, SVCA provides a quantitative assessment of the effect of cell-cell interactions on the expression profile of individual molecules. The model avoids the explicit definition of cell types and neighbourhoods, instead using continuous measure of cell state and euclidean distances between cells.

We have applied SVCA to multiple datasets that were generated using alternative technologies, probing either RNA transcripts or proteins, demonstrating the broad applicability of the model. Across these applications, we observed that cell-cell interactions can substantially contribute to gene expression variation, which is consistent with previous reports (Battich et al., 2015; Goltsev et al., 2018; Kamińska et al., 2015; Nasra Naeim Ayuob and Soad Shaker Ali, 2012) and supports the concept that studying single cell expression in the native context is important for understanding the sources of these variations. We also showed that the variability of spatial variance signatures across samples of the same biological system could be linked to different clinical contexts or to the internal structure of a given organ, which underlines the biological relevance of these signatures. Finally, using higher dimensionality data obtained from optical technologies, we observed that genes with largest cell-cell interaction components were enriched in specific pathways, suggesting that spatial variance signatures could help us understanding biological activity in tissue. These results suggest that analysing single cell data in their native context, with SVCA or related methods, will help us understand how the spatial organisation of tissues impact single cell biology.

Although we have tested the calibration and robustness of SVCA, the model is not free of limitations. At present, the model does not account for technology-specific noise and instead assumes Gaussian distributed residuals, thus requiring suitable processing of the raw data such that these assumptions are sufficiently met (See **Methods**). Further development could consider a generalized random effects model, for example to couple the random effect component with a negative-binomial likelihood. A second limitation of SVCA is that the model is univariate, which means that individual genes or proteins are modelled independently from each other. Multivariate extensions could account for relationships between genes involved in the same pathways, either in an unsupervised manner or using prior knowledge (Buettner, Pratanwanich, McCarthy, Marioni, & Stegle, 2017). Such approaches could give a more comprehensive understanding of how biological processes are affected by tissue structure.

There is a growing appreciation of the role of spatial distribution of proteins, transcripts and other molecules in determining tissues functioning and its deregulation in disease, with potential value as predictors of clinical outcomes (Bodenmiller, 2016). This is largely driven by vigorous development of novel technologies that enable us to capture such data (Aichler

& Walch, 2015; Bodenmiller, 2016; Goltsev et al., 2018; Lin et al., 2017; Schulz et al., 2018). We believe that the SVCA framework and extensions thereof will be of broad use to analyze this burgeoning spatially-resolved molecular data to advance our understanding of the pathophysiology of multiple diseases.

Methods & software

A full derivation of SVCA and details on the analysis are provided as Supplementary Methods. An open source implementation of SVCA is available at https://github.com/damienArnol/svca.git, which builds on the limix package (Lippert, Casale, Rakitsch, & Stegle, 2014).

Author contributions

Author contributions: DA, OS developed the statistical method. DA implemented the model and analysed all the data. DS, BB contributed to the interpretation of the data and the design of the method. DA, DS, OS, JSR wrote the manuscript with input from all authors. JSR and OS conceived the project and supervised the work.

Acknowledgments

DA, JSR, OS acknowledge EMBL core funding. DS was supported by the Forschungskredit of the University of Zurich, grant FK-74419-01-01, and the BioEntrepreneur-Fellowship of the University of Zurich, reference no. BIOEF-17-001. B.B.'s research is funded by an SNSF R'Equip grant, an SNSF Assistant Professorship grant, the SystemsX Transfer Project "Friends and Foes", the SystemsX MetastasiX and PhosphoNetX grant, NIH grant (UC4 DK108132), and the European Research Council (ERC) under the European Union's Seventh Framework Program (FP/2007-2013)/ERC grant agreement no. 336921.

We thank R. Argelaguet, V. Svensson, R. Vento, H. Jackson, A. Baud, N. Cai, F.P. Casale and D. Horta for discussions on data processing, model design and implementation and results visualisation.

Competing financial interest statement

The authors declare no competing financial interests.

References

Achim, K., Pettit, J.-B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., & Marioni, J. C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*, *33*(5), 503–509. https://doi.org/10.1038/nbt.3209

Aichler, M., & Walch, A. (2015). MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, *95*(4), 422. https://doi.org/10.1038/labinvest.2014.156

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., … Nolan, G. P. (2014). Multiplexed ion beam imaging of human breast tumors. *Nature Medicine*, *20*(4), 436–442. https://doi.org/10.1038/nm.3488

Battich, N., Stoeger, T., & Pelkmans, L. (2013). Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods*, *10*(11), 1127–1133. https://doi.org/10.1038/nmeth.2657

Battich, N., Stoeger, T., & Pelkmans, L. (2015). Control of Transcript Variability in Single Mammalian Cells. *Cell*, *163*(7), 1596–1610. https://doi.org/10.1016/j.cell.2015.11.018

Baud, A., Mulligan, M. K., Casale, F. P., Ingels, J. F., Bohl, C. J., Callebert, J., … Stegle, O. (2017). Genetic Variation in the Social Environment Contributes to Health and Disease. *PLoS Genetics*, *13*(1), e1006498. https://doi.org/10.1371/journal.pgen.1006498

Bodenmiller, B. (2016). Multiplexed Epitope-Based Tissue Imaging for Discovery and Healthcare Applications. *Cell Systems*, *2*(4), 225–238. https://doi.org/10.1016/j.cels.2016.03.008

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., & Stegle, O. (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*, *18*(1), 212. https://doi.org/10.1186/s13059-017-1334-8

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, *348*(6233), aaa6090. https://doi.org/10.1126/science.aaa6090

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., … D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *42*(Database issue), D472–D477. https://doi.org/10.1093/nar/gkt1102

Elston, C. W., & Ellis, I. O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, *19*(5), 403–410. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1757079

Gerdes, M. J., Sevinsky, C. J., Sood, A., Adak, S., Bello, M. O., Bordwell, A., … Ginty, F. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11982–11987. https://doi.org/10.1073/pnas.1300136110

Giesen, C., Wang, H. A. O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., … Bodenmiller, B. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, *11*(4), 417–422. https://doi.org/10.1038/nmeth.2869

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vasquez, G., … Nolan, G. (2018, February 5). *Deep profiling of mouse splenic architecture with CODEX multiplexed imaging*. *bioRxiv*. https://doi.org/10.1101/203166

Kamińska, K., Szczylik, C., Bielecka, Z. F., Bartnik, E., Porta, C., Lian, F., & Czarnecka, A. M. (2015). The role of the cell-cell interactions in cancer progression. *Journal of Cellular and Molecular Medicine*, *19*(2), 283–296. https://doi.org/10.1111/jcmm.12408

Lin, J.-R., Fallahi-Sichani, M., & Sorger, P. K. (2015). Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature*

Communications, *6*, 8390. https://doi.org/10.1038/ncomms9390

Lin, J.-R., Izar, B., Mei, S., Wang, S., Shah, P., & Sorger, P. (2017, June 19). *A simple open-source method for highly multiplexed imaging of single cells in tissues and tumours*. *bioRxiv*. https://doi.org/10.1101/151738

Lippert, C., Casale, F. P., Rakitsch, B., & Stegle, O. (2014, May 21). *LIMIX: genetic analysis of multiple traits*. *bioRxiv*. https://doi.org/10.1101/003905

Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(39), 11046–11051. https://doi.org/10.1073/pnas.1612826113

Nasra Naeim Ayuob and Soad Shaker Ali. (2012). Cell-Cell Interactions and Cross Talk Described in Normal and Disease Conditions: Morphological Approach. In S. Gowder (Ed.), *Cell Interaction*. InTech.

Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R. T., Schulz, D., … Bodenmiller, B. (2017). histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods*, *14*(9), 873–876. https://doi.org/10.1038/nmeth.4391

Schulz, D., Zanotelli, V. R. T., Fischer, J. R., Schapiro, D., Engler, S., Lun, X.-K., … Bodenmiller, B. (2018). Simultaneous Multiplexed Imaging of mRNA and Proteins with Subcellular Resolution in Breast Cancer Tissue Samples by Mass Cytometry. *Cell Systems*, *6*(1), 25–36.e5. https://doi.org/10.1016/j.cels.2017.12.001

Shah, S., Lubeck, E., Zhou, W., & Cai, L. (2017). seqFISH Accurately Detects Transcripts in Single Cells and Reveals Robust Spatial Organization in the Hippocampus. *Neuron*, *94*(4), 752–758.e1. https://doi.org/10.1016/j.neuron.2017.05.008

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., … Frisén,

J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, *353*(6294), 78–82. https://doi.org/10.1126/science.aaf2403

Svensson, V., Teichmann, S. A., & Stegle, O. (n.d.). SpatialDE - Identification of spatially variable genes. https://doi.org/10.1101/143321

Supplementary Figures

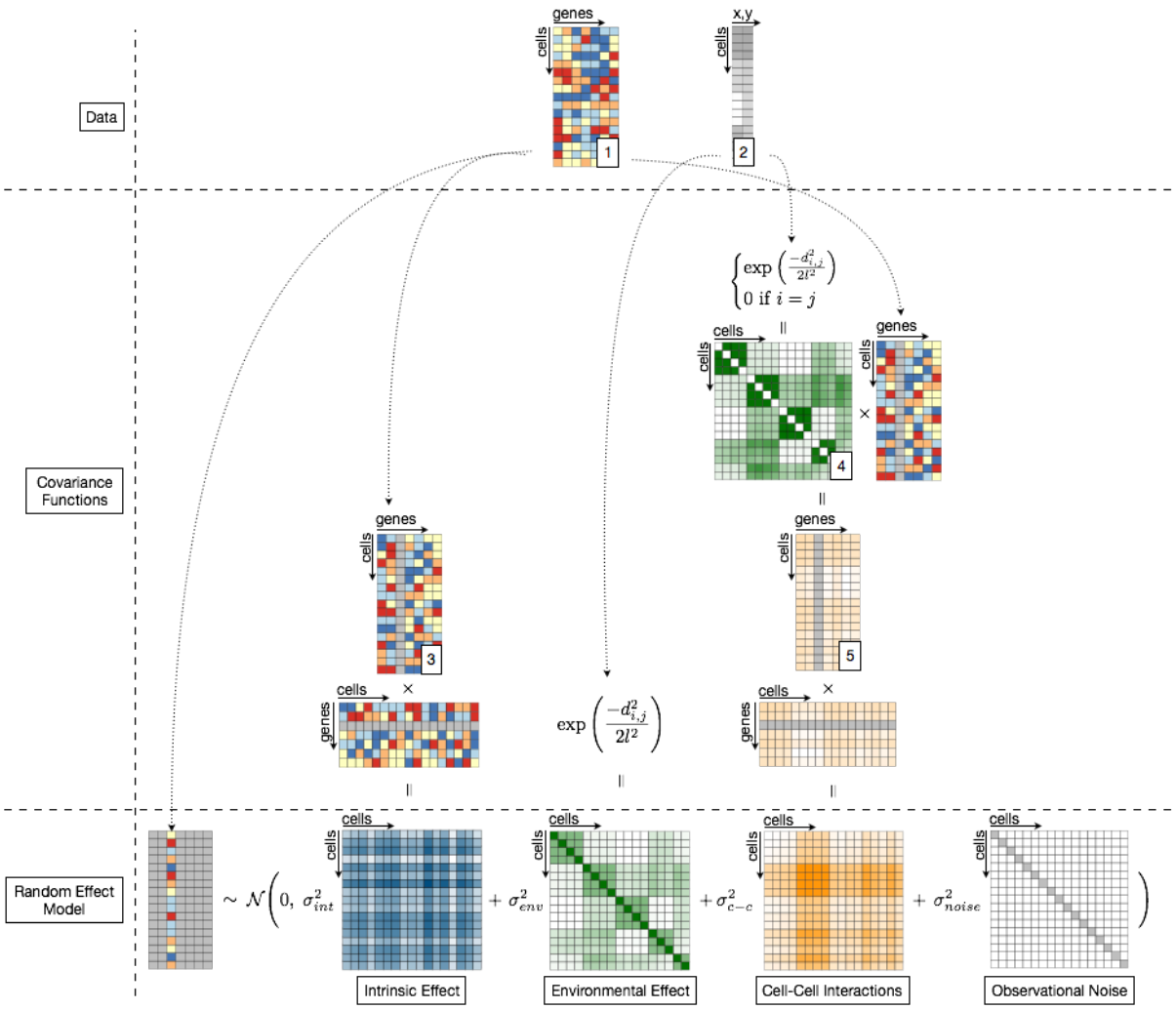

Figure S1 | SVCA model definition.
SVCA takes as an input single cell expression data as a cell times gene/protein matrix (1) and the spatial location of the cells as a cell times (x,y) coordinates matrix (2). Individual genes are modelled as multivariate normally distributed, with additive covariance components that account for different effects modelled by SVCA (intrinsic, environmental and cell-cell interactions, **Fig.1**). The covariance for intrinsic effects is computed as the empirical covariance of the expression profiles between cells, where the modelled gene has been removed from the expression matrix (3). Environmental effects are modelled using a Squared Exponential covariance defined on the relative distance between cells. Cell-cell interactions are modelled using a cellular neighbourhood matrix (5) which aggregates, for each cell, the molecular composition of the neighbouring cells. This is achieved by weighting the molecular profiles of all neighbouring cells with a squared exponential function of the distance. This can be written as a product between a squared exponential covariance matrix whose diagonal elements were removed (4) and the expression matrix (3). The final cell-cell

interaction covariance is computed as the empirical covariance of the cellular environment matrix between cells. SVCA's training then consists in learning the scale of every covariance term.

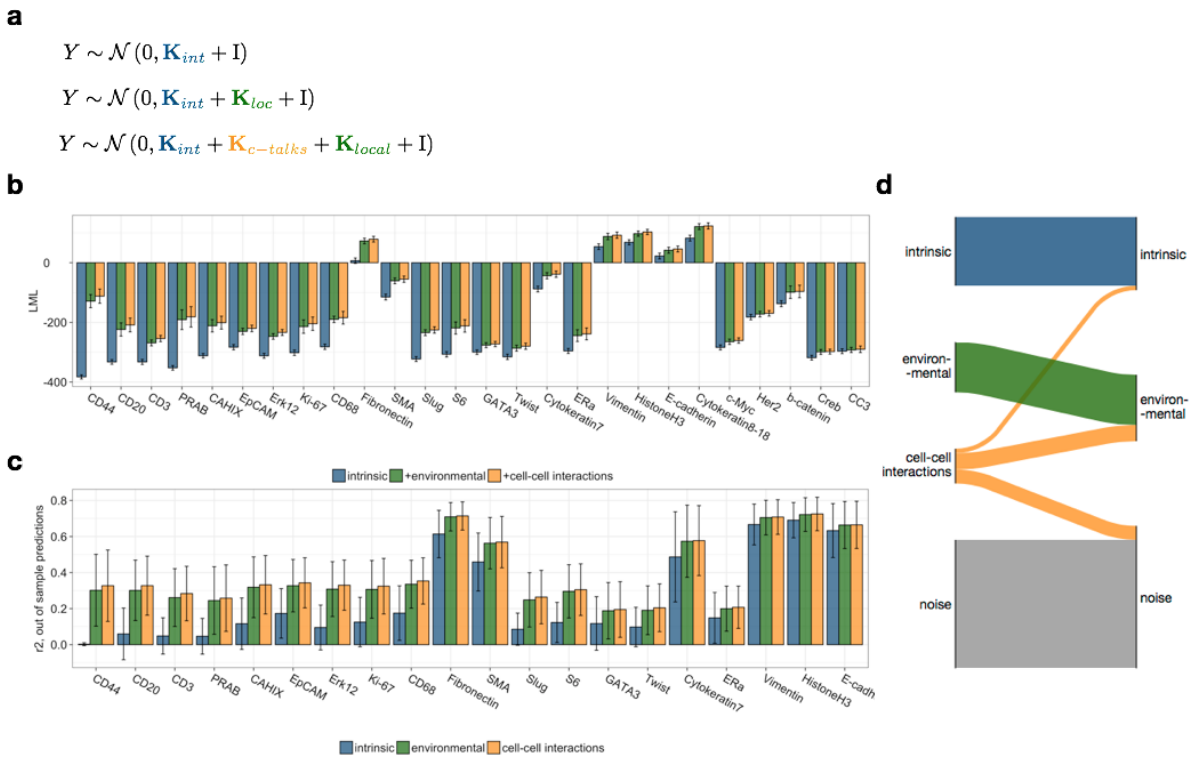

Figure S2 | Assessment of the predictive performance of SVCA and reduced models. **(a)** Comparison of three models in increasing order of complexity. The first model accounts only for the intrinsic component, the second model also includes the environmental effect. The third model includes all three components of SVCA **(b)** Log marginal likelihood of the three models of increasing complexity **(c)** Out of sample prediction performance (r2) for the three models of increasing complexity, in a 5-fold cross validation setting (see **Methods**). The results show that the increase in complexity also improves the generalisation performance to a validation set. Proteins are ordered by the size of the cell-cell interaction component. **(d)** Variance redistribution from the SVCA model to a reduced model including intrinsic and environmental effects only, averaged across all proteins (see **Fig. 2**).

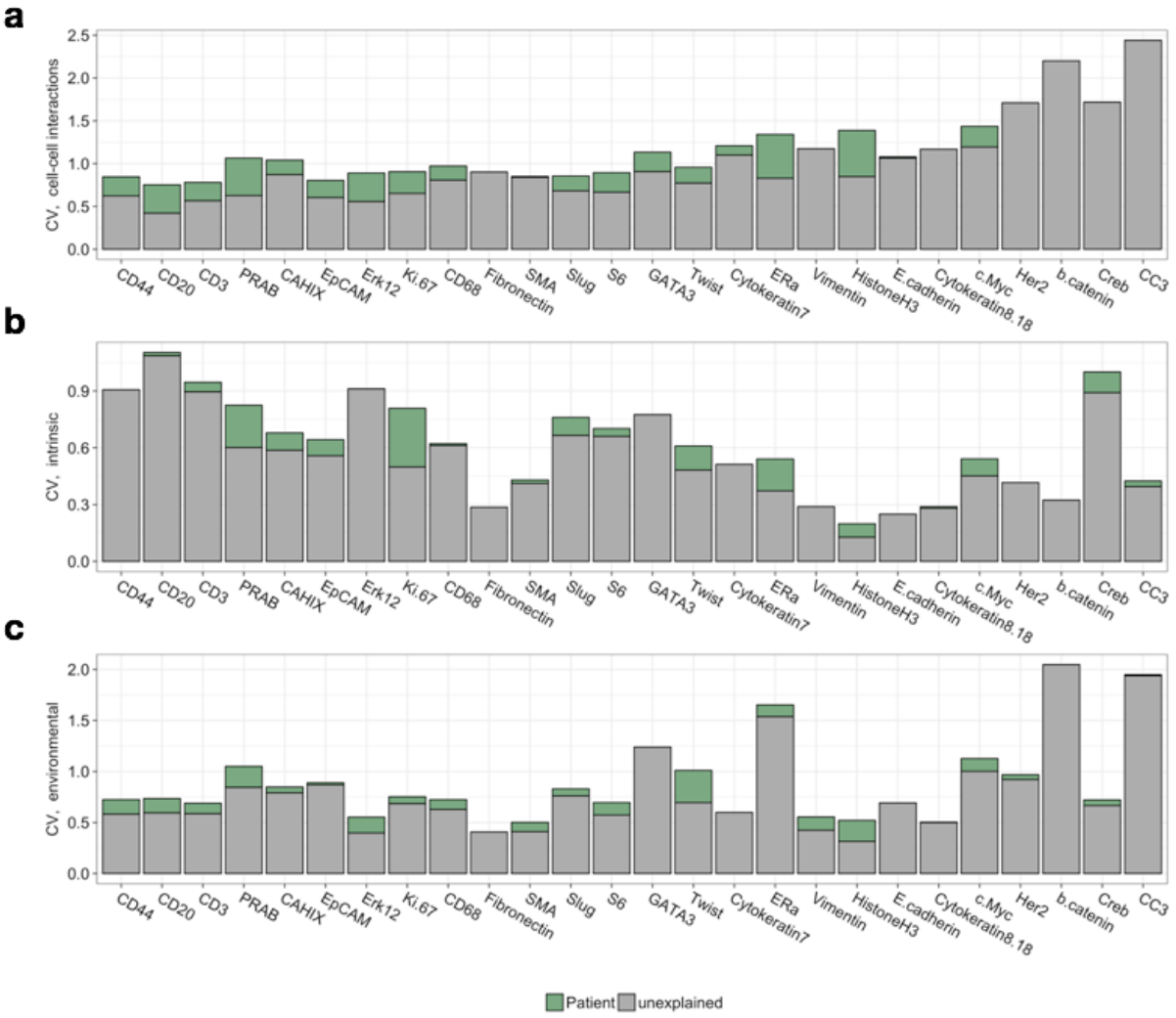

Figure S3 | Coefficient of variation of SVCA variance components across samples
Coefficient of variation of the different variance components. Shown in green is the proportion that is attributable to patient, estimated using a random effect model (see **Methods**) **(a)** Cell-cell interaction component **(b)** Intrinsic component **(c)** Environmental component.

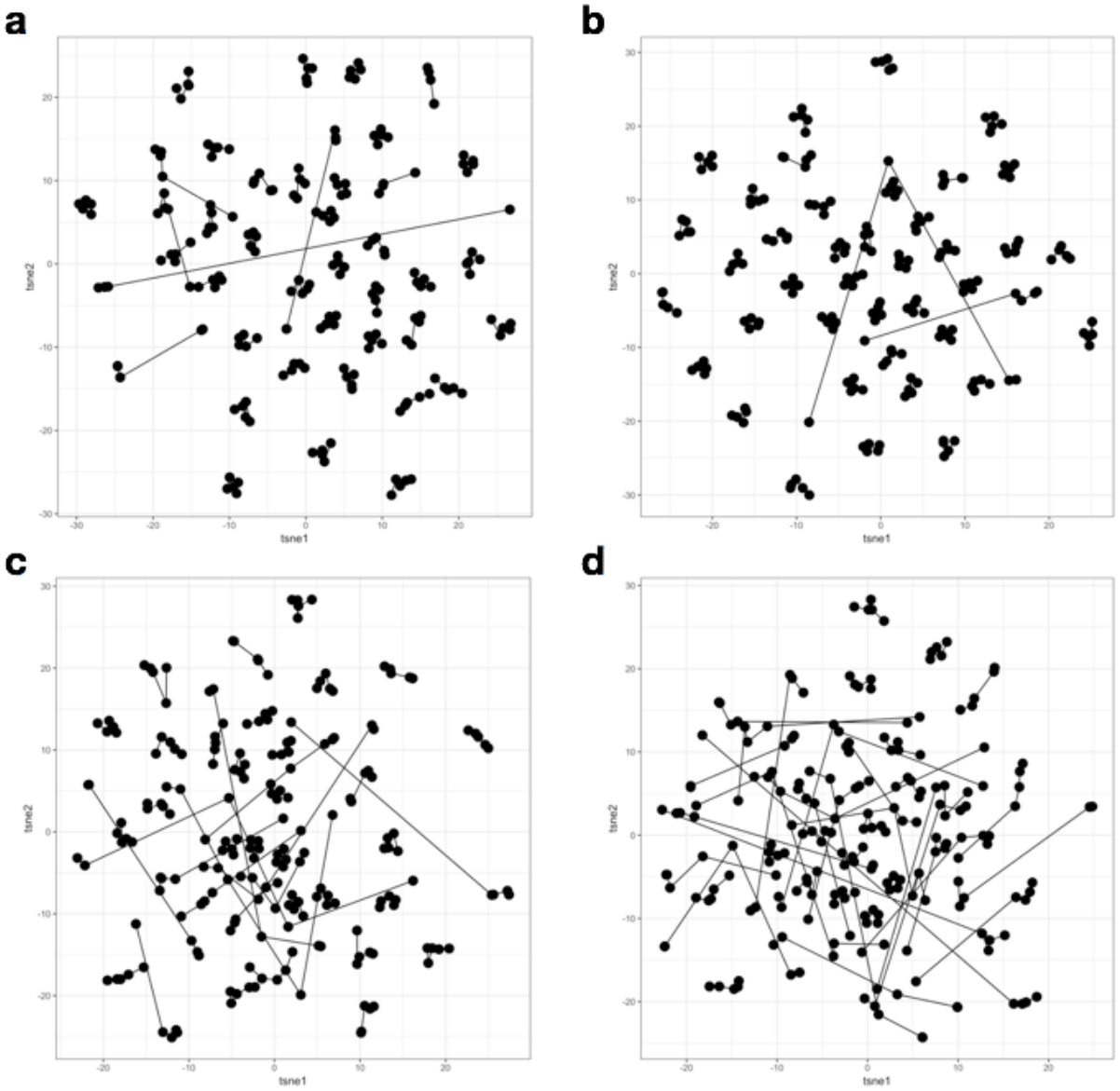

Figure S4 | Spatial variance signature robustness using a bootstrapping strategy (see Methods)

t-SNE representation of the spatial variance signature components including **(a)** the entire SVCA signature, **(b)** only the intrinsic component, **(c)** only the cell-cell interaction component **(d)** only the environmental component. Results from multiple bootstraps for the same image (linked with a line) cluster together, confirming the robustness of the spatial variance signatures

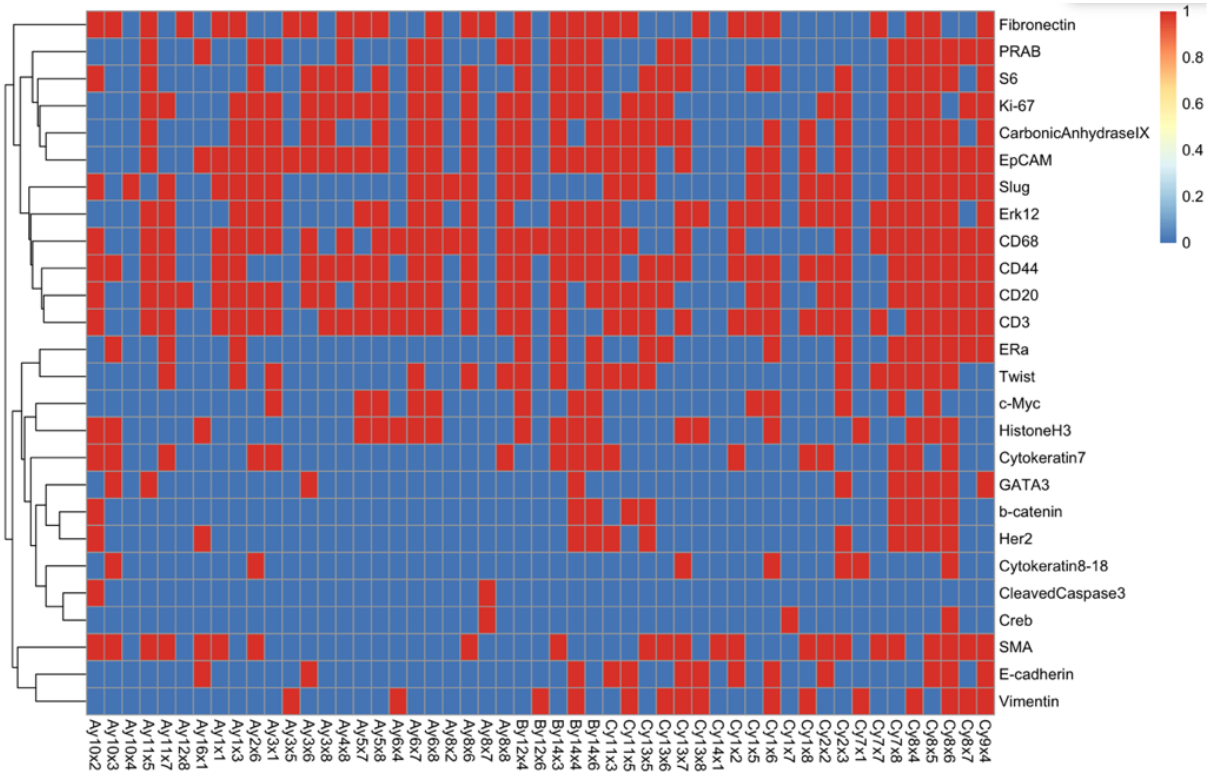

Figure S5 | Significance of the cell-cell interaction components per protein and image
In red, proteins for which the cell-cell interaction component is significant, for a FDR threshold of 1% (Benjamini-Hochberg corrected, see **Methods**)

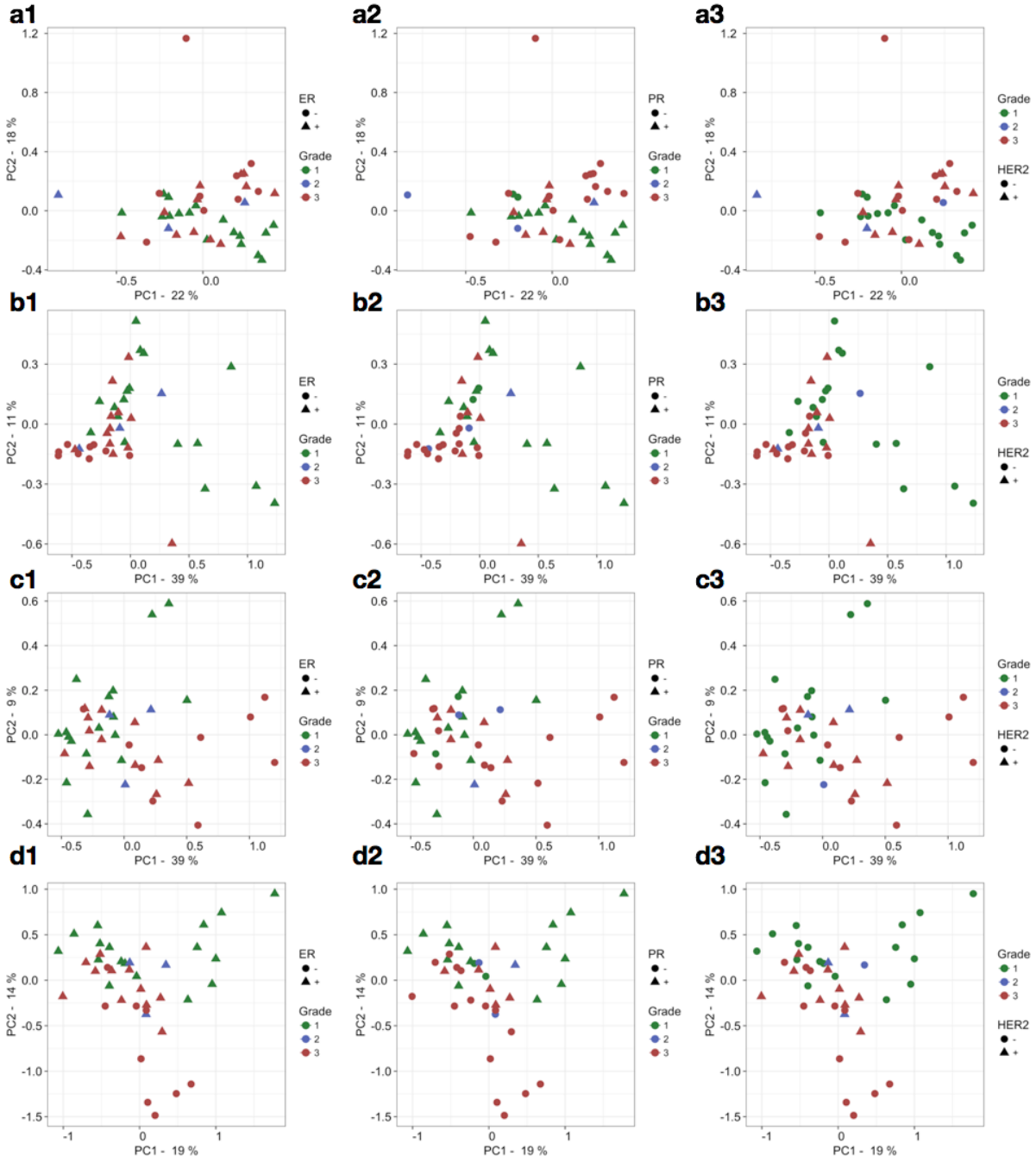

Fig S6 | Variance signatures PCA and link to the clinical data

(a) First two components for the intrinsic variance term for all proteins. The colour corresponds to the grade and the shapes correspond to **(a1)** the ER status **(a2)** the PR status and **(a3)** the Her2 status of the samples respectively. **(b)** is for the environmental variance component, **(c)** for the cell-cell interaction component and **(d)** for the entire Spatial Variance signature. Spatial variance signatures are mostly affected by tumour grades.

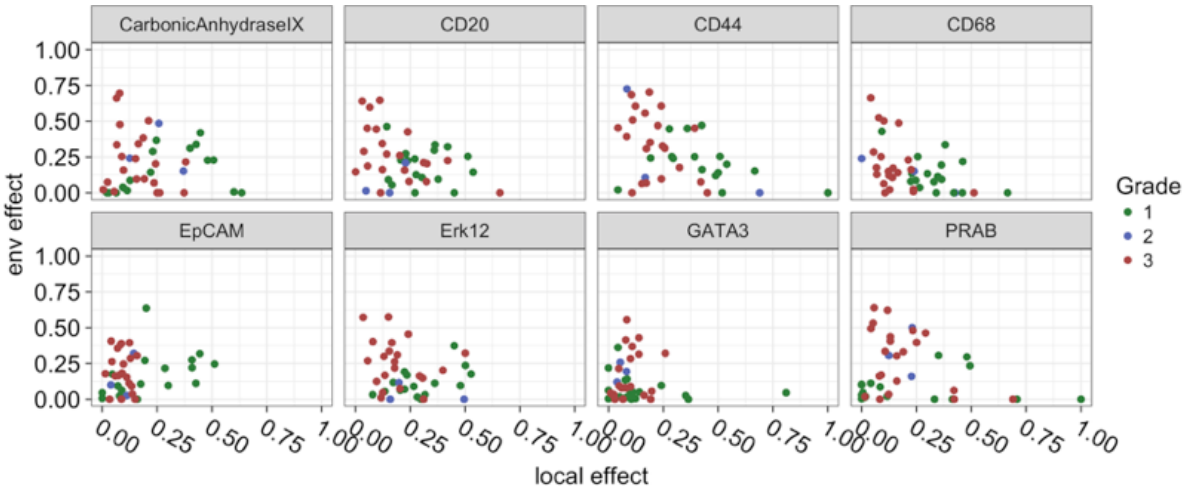

Fig S7 | Ratio between cell-cell interaction effects and environmental effect
Proteins are the relevant proteins from the PCA loadings inspection. x axis: environmental effect, y axis: cell-cell interaction effect. Each point corresponds to an image and the colour corresponds to the cancer grade

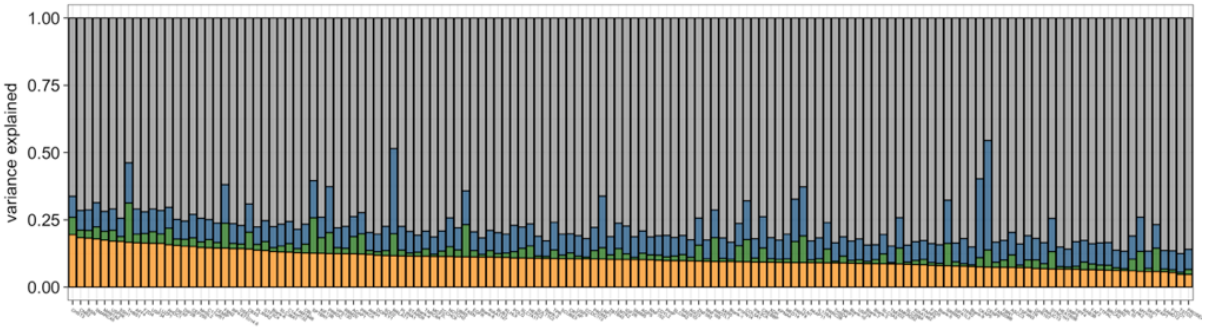

Fig S8 | Application to mer-FISH data
Spatial variance signature for all genes.

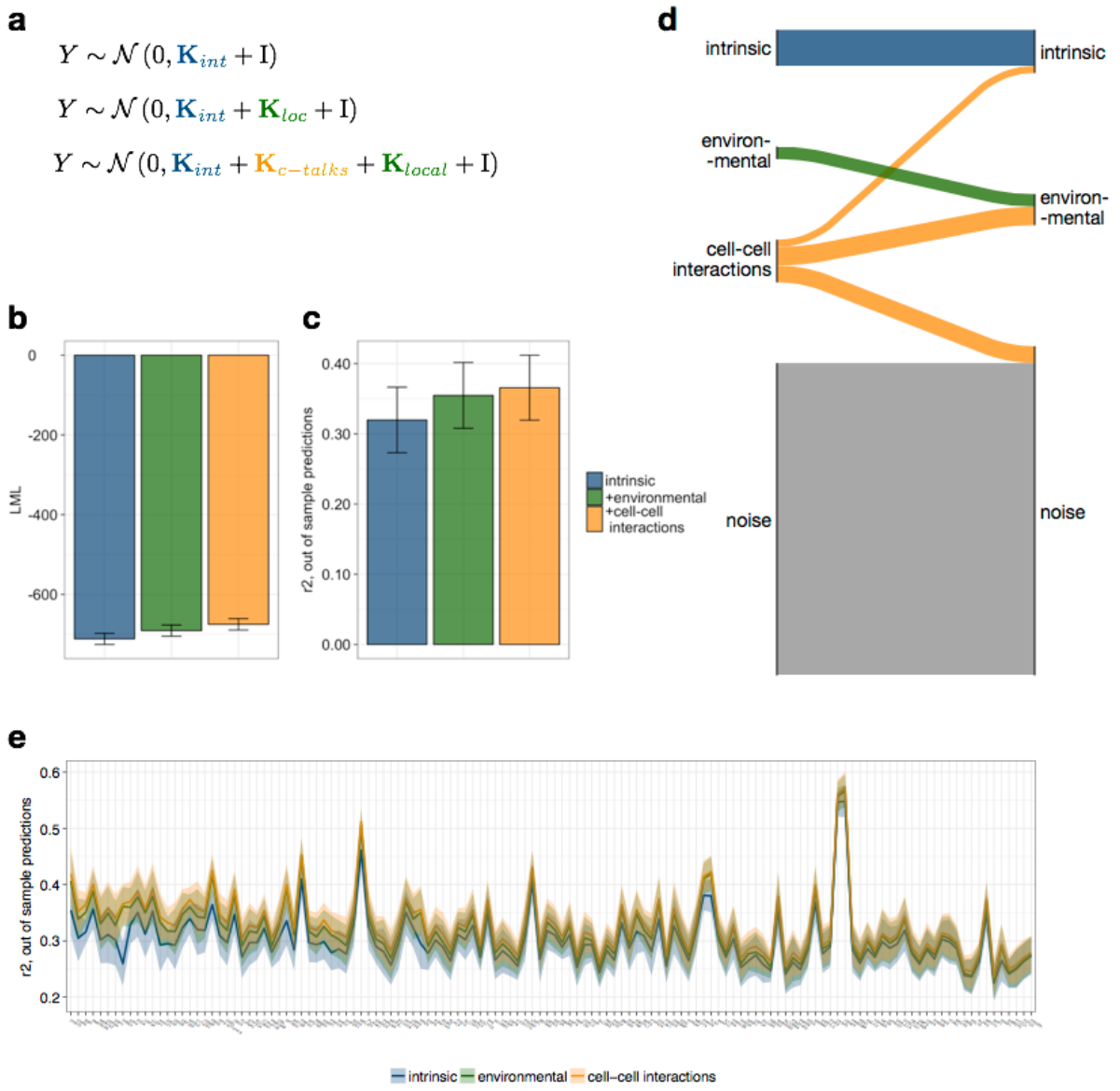

Fig S9 | Model comparison mer-FISH data

(a) Models compared in order of increasing complexity **(b)** Log marginal likelihood of the three models in increasing order of complexity. Results averaged across all genes. **(c)** Out of sample prediction: coefficient of determination averaged across all proteins for the models in increasing order of complexity. **(d)** Variance redistribution from the SVCA model to a reduced model including intrinsic and environmental effects only, averaged across all proteins (see **Fig. 2**). **(e)** Out of sample prediction for the three models, all proteins. Shaded areas correspond to plus and minus one standard deviation.

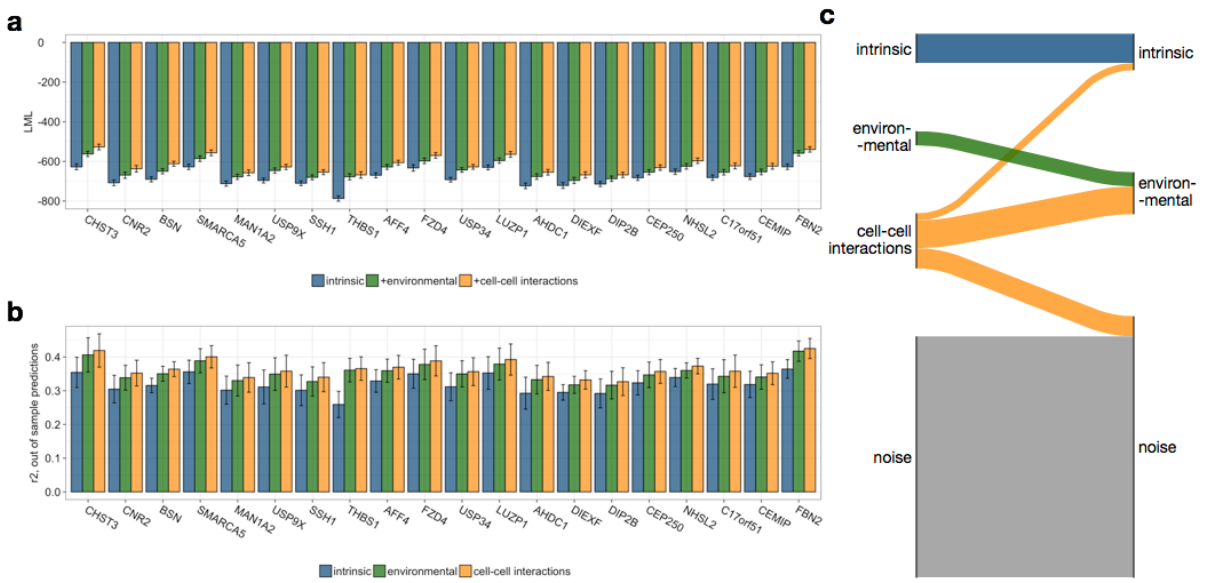

Fig S10 | Model comparison mer-FISH data, top 20 cross talk genes

(a) Log Marginal Likelihood **(b)** Out of sample prediction r2 **(c)** Variance redistribution from the SVCA model to the reduced model omitting cell-cell interactions (see **Fig. 2**).

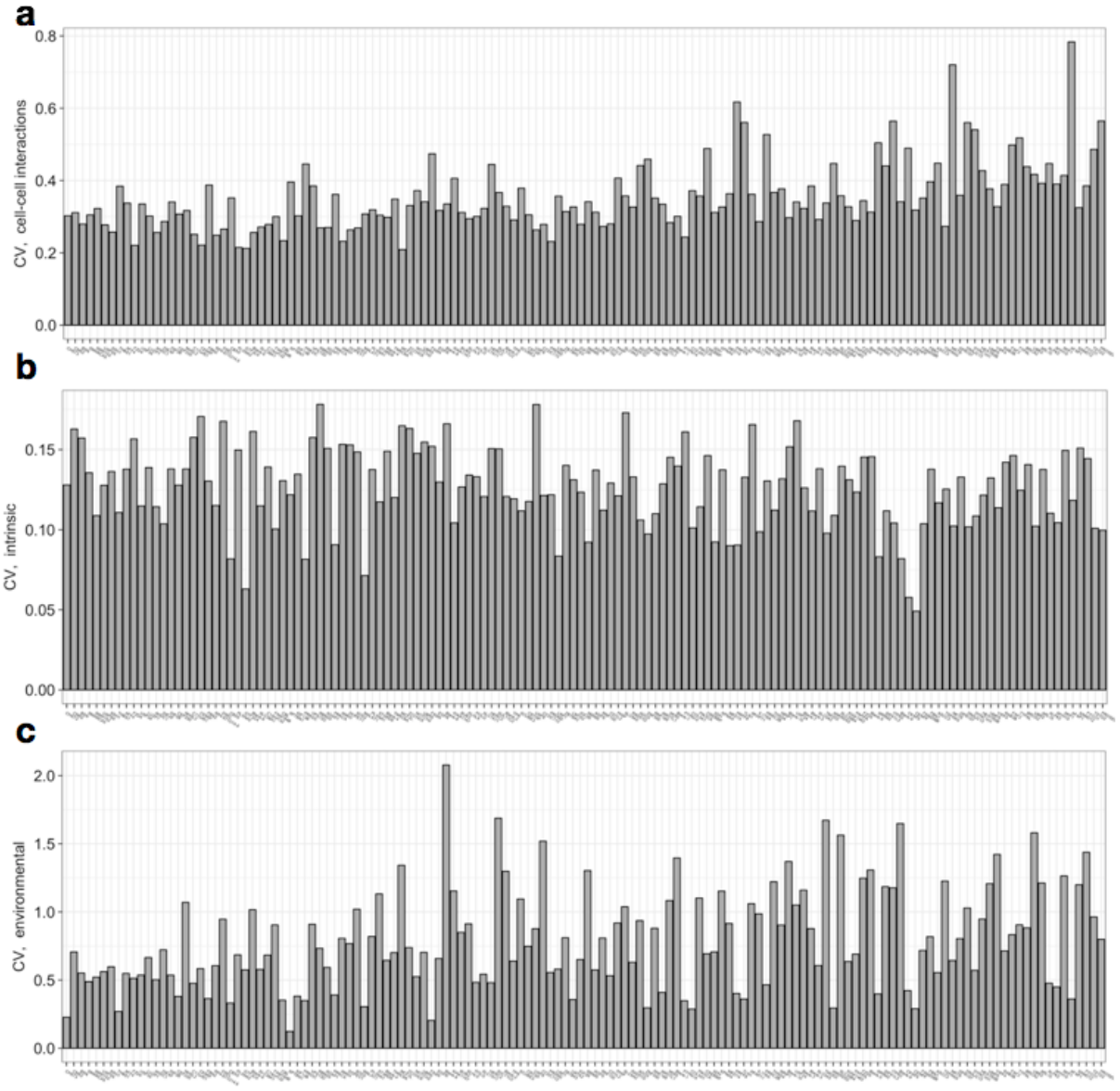

Fig S11 | Coefficient of variation of variance components (mer-FISH data)
(a) Cell-cell interaction term **(b)** Intrinsic term **(c)** Environmental term

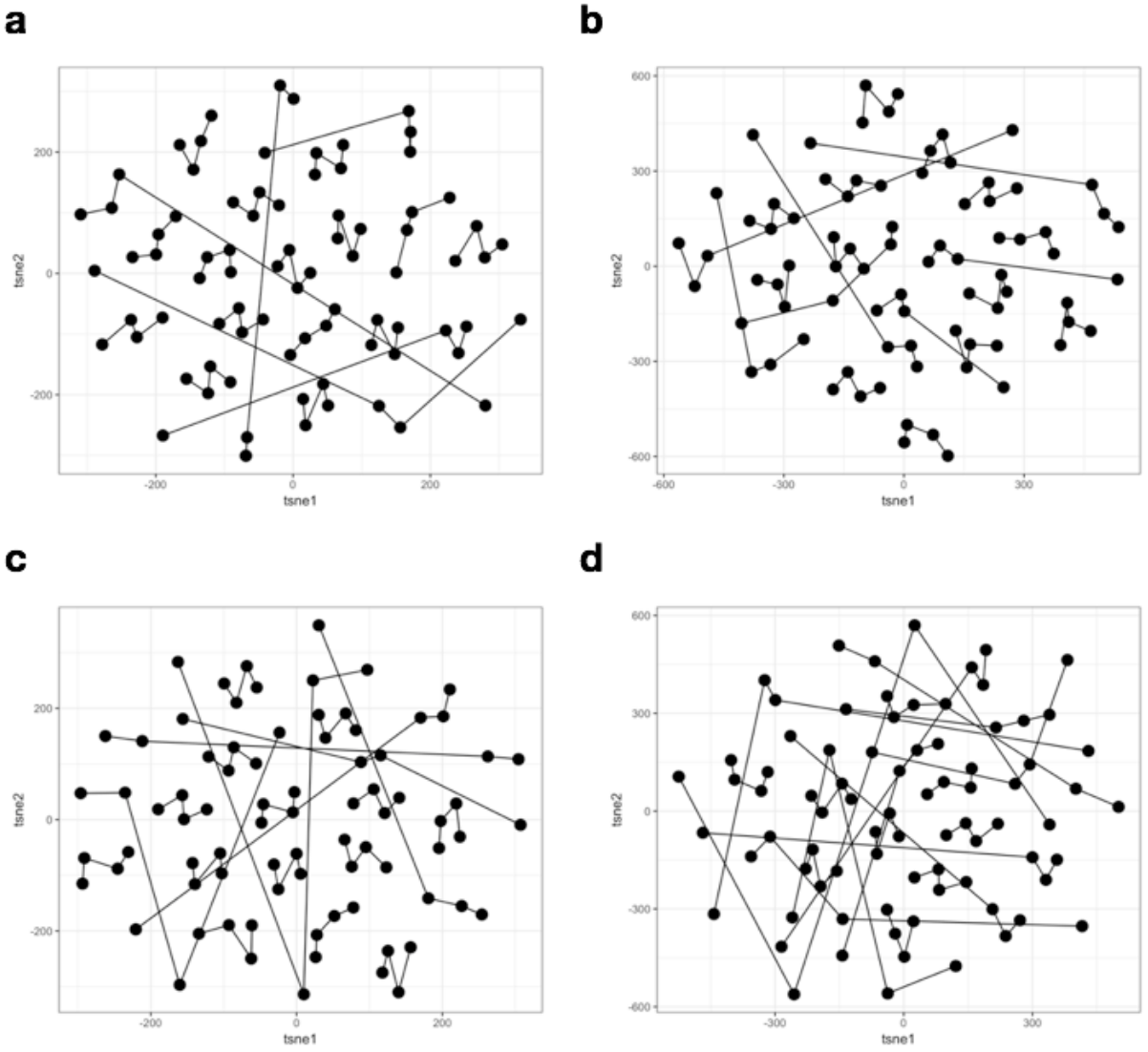

Fig S12 | Signature robustness visualised with t-SNE (mer-FISH data)
t-SNE representation of the spatial variance signature components including **(a)** the entire SVCA signature, **(b)** only the intrinsic component, **(c)** only the cell-cell interaction component **(d)** only the environmental component. Results from multiple bootstraps for the same image (linked with a line) cluster together, confirming the robustness of the spatial variance signatures.

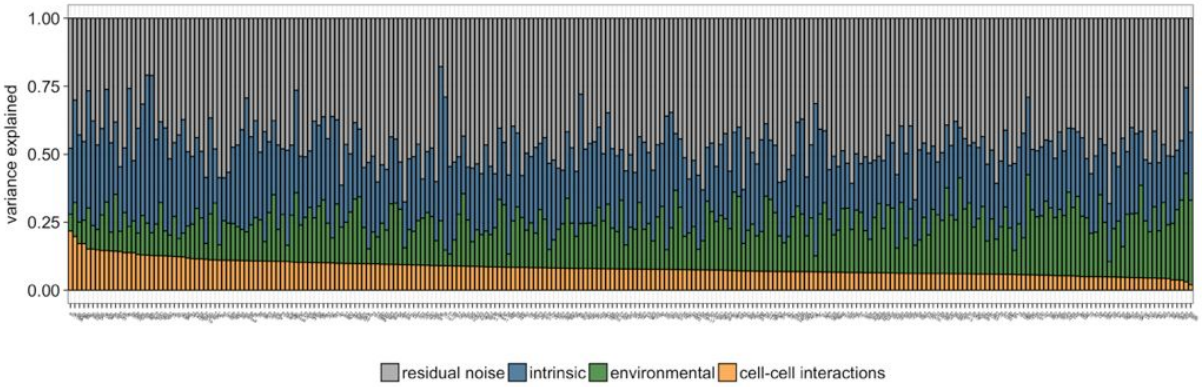

Fig S16 | Application to seq-FISH data
Spatial variance signature for all genes

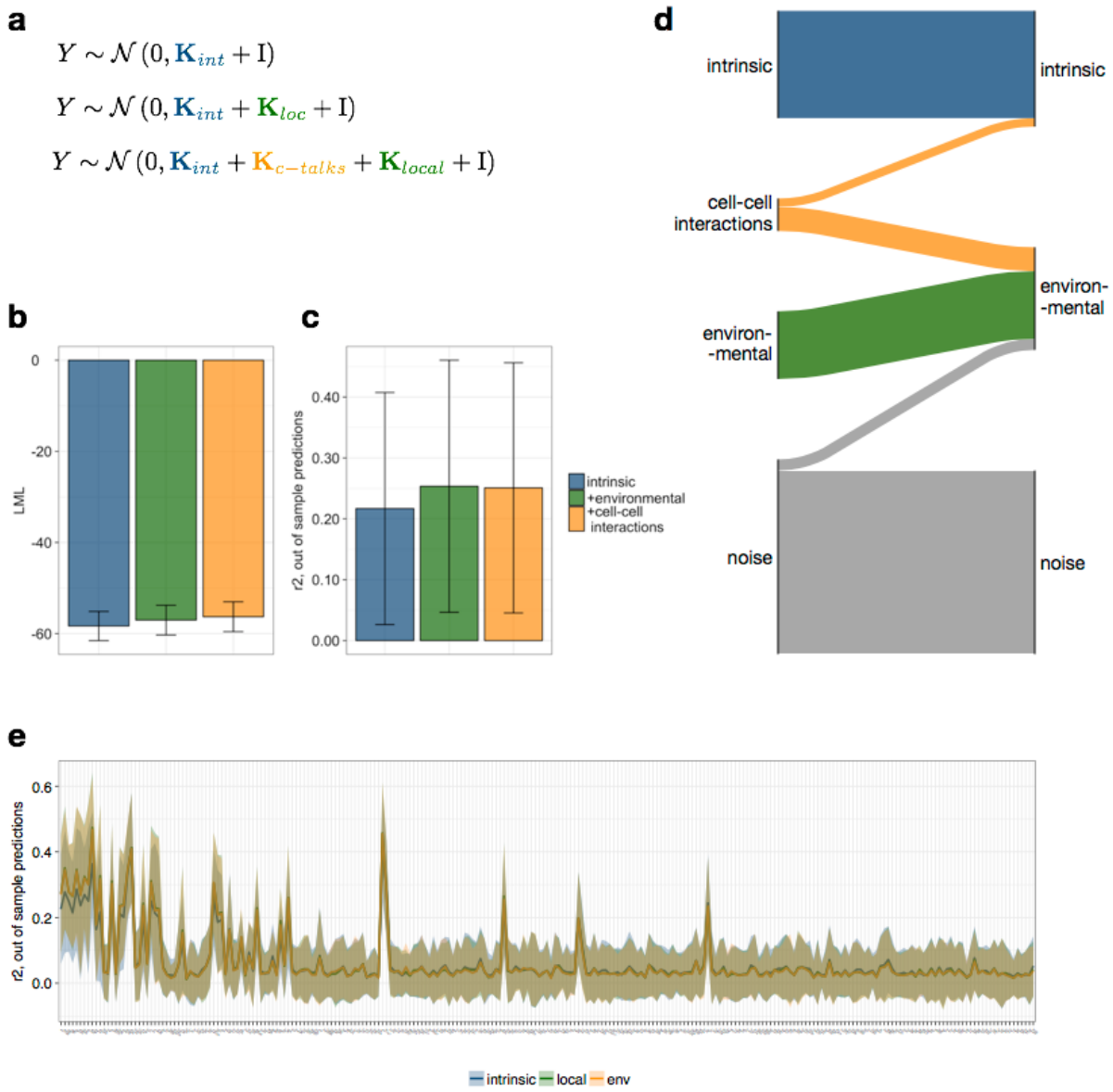

Fig S13 | Model comparison seq-FISH data

(a) Models compared in order of increasing complexity **(b)** Log marginal likelihood of the three models in increasing order of complexity. Results averaged across all genes. **(c)** Out of sample prediction: coefficient of determination averaged across all proteins for the models in increasing order of complexity. **(d)** Variance redistribution from the SVCA model to a reduced model including intrinsic and environmental effects only, averaged across all proteins (see **Fig. 2**). **(e)** Out of sample prediction for the three models, all proteins. Shaded areas correspond to plus and minus one standard deviation.

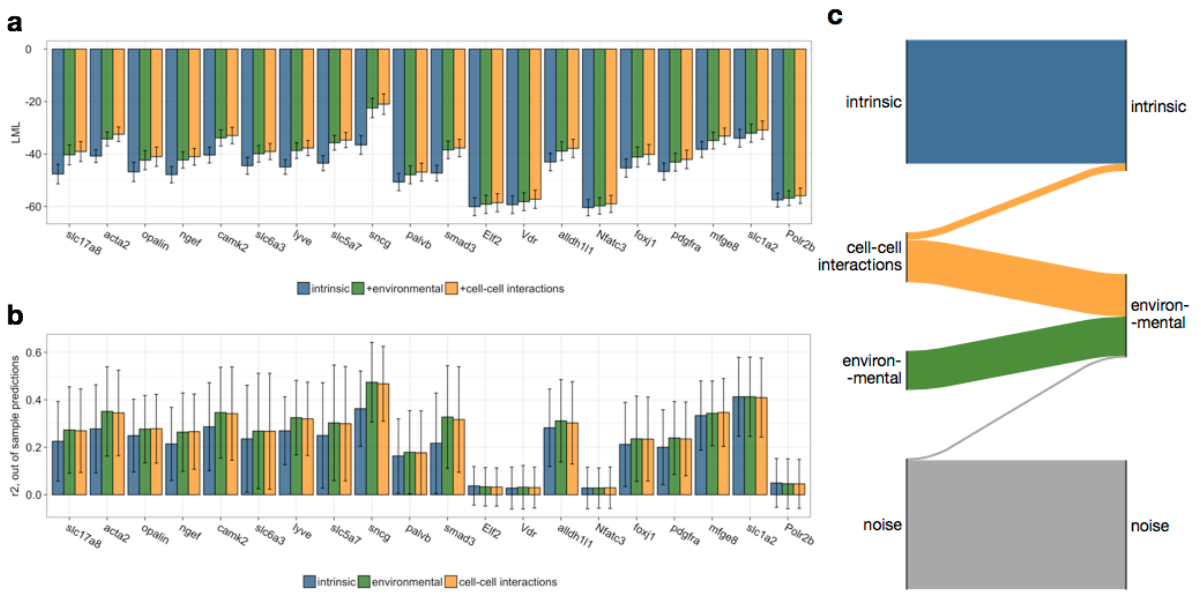

Fig S14 | Model comparison seq-FISH data, top 20 cross talk genes
(a) Log Marginal Likelihood **(b)** Out of sample prediction r2 **(c)** Variance redistribution from the SVCA model to the reduced model omitting cell-cell interactions (see **Fig. 2**).

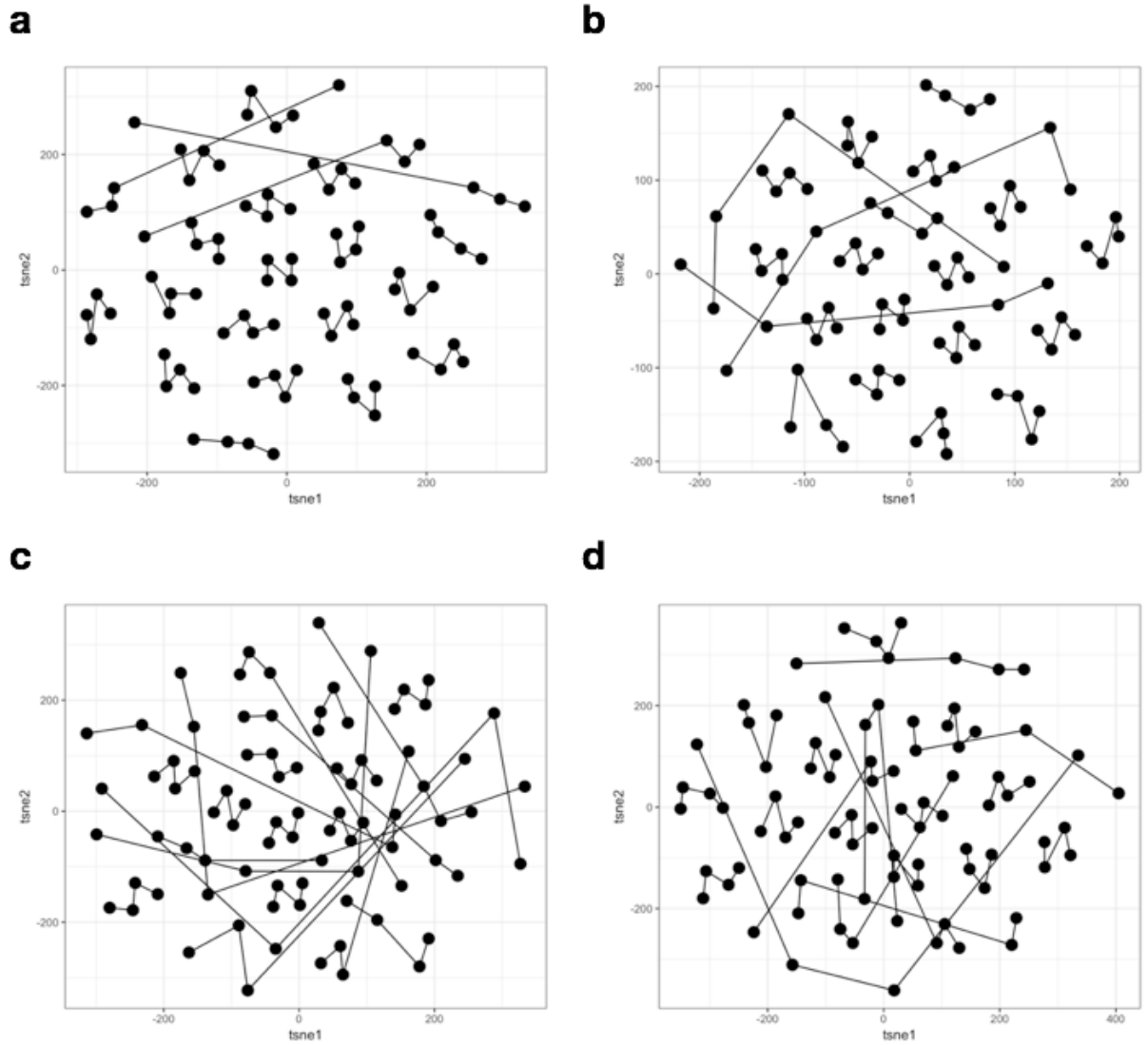

Fig S15 | Signature robustness visualised with t-SNE seq-FISH data)
t-SNE representation of the spatial variance signature components including **(a)** the entire SVCA signature, **(b)** only the intrinsic component, **(c)** only the cell-cell interaction component **(d)** only the environmental component. Results from multiple bootstraps for the same image (linked with a line) cluster together, confirming the robustness of the spatial variance signatures.

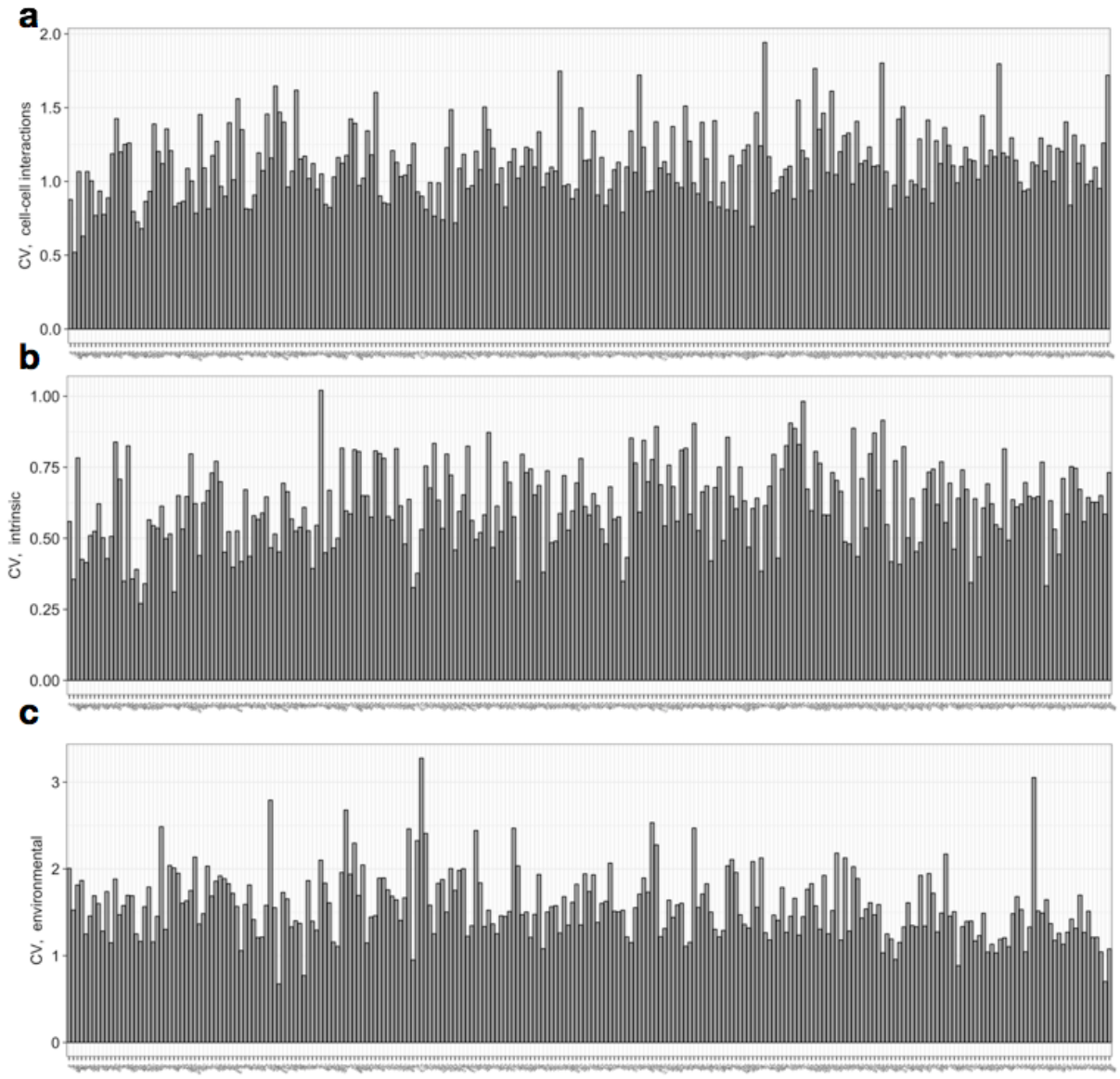

Fig S16 | Coefficient of variation of variance components (seq-FISH data)
(a) Cell-cell interaction term **(b)** Intrinsic term **(c)** Environmental term

Supplementary Methods: Modelling cell-cell interactions in spatial omics and spatial variance component analysis

Damien Arnol, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, Oliver Stegle

Contents

1 SVCA Model **2**
1.1 Model overview . . . 2
1.2 Definition of covariance terms . . . 3
1.3 Parameters' inference . . . 4
1.4 Estimates of variance components . . . 4
1.5 Significance of variance components . . . 5
1.5.1 General Approach . . . 5
1.5.2 Estimation of the null LLRs distributions . . . 5
1.5.3 Multiple testing correction and False Discovery Rate . . . 5
1.6 Comparison to related models . . . 6
1.6.1 Schapiro et al 2017 - HistoCAT . . . 6
1.6.2 Battich et al 2015 . . . 6
1.6.3 Goltsev et al 2018 . . . 6

2 Model Validation **7**
2.1 Simulations of the cell-cell interaction component . . . 7
2.2 P-value calibration for the cell-cell interaction component . . . 7
2.3 Results robustness . . . 8
2.4 Out of sample prediction . . . 8
2.5 Identifiability of cell-cell interactions versus environmental effects . . . 8

3 Downstream Analysis **9**
3.1 Variance across images of the IMC dataset and patient effect . . . 9
3.2 Signature PCA . . . 9
3.3 Gene Set Enrichment Analysis . . . 9

4 Dataset **10**
4.1 Dataset description . . . 10
4.1.1 Imaging Mass Cytometry (IMC) data . . . 10
4.1.2 mer-FISH and seq-FISH data . . . 10
4.2 Data processing . . . 11

1 SVCA Model

1.1 Model overview

The expression level of a single gene across cells is modelled using a random effect model implemented in a Gaussian Process (GP) framework [16, 13] with an additive covariance. The covariance is composed of 4 terms, modelling our assumption that the variance across cells of the gene expression level is due to three additive effects: an intrinsic effect due to the cell state, a cell-cell interaction effect, due to the state of the neighbouring cells, and an environmental effect due to unobserved factors in the cell micro environment, such as local access to oxygen, nutriments etc.

In the following, we will make explicit the different terms of the covariance, show how they are parametrised and how these parameters are optimised. We will then explain how this model can be used to assess the proportion of variance explained by each effect, as well as the statistical significance of each effect. See also **Fig. S1**.

We will rely on the following nomenclature and notations:

Nomenclature

Molecule of interest: Individual molecule, typically a gene or protein, on which SVCA is fitted.

Cell state: Intrinsic characteristic of a cell. In this paper, we take the overall expression profile excluding the gene or protein of interest as a multidimensional and continuous measure of cell state. Other possibilities include classifying cells into cell types.

cellular neighbourhood matrix: Continuous measure of the molecular composition of a cell's neighbouring cells, summarised by weighting the molecular profiles of all neighbouring cells with a squared exponential function of their distance to the focal cell.

Intrinsic effect: Effect of the cell state on the expression level of the molecule of interest.

Cell-cell interaction effect: Effect of cell-cell interactions on the expression level of the molecule of interest. These interactions may harvest signaling between cells but also cell-types cooccurrences for example.

Environmental effect: Effect of the cell's position on the expression level of the molecule of interest. This effect harvests unmeasured variables from the microenvironment with an effect on gene expressions, such as local glucose or oxygen access.

Spatial variance signature: Concatenation of all variance estimates (intrinsic effect, environmental effect, cell-cell interaction effect and residual noise) across all molecule for a given image.

Notations

- N - number of cells in a given image
- D - number of molecules (eg genes or proteins) in a given image
- Y - Expression level of the molecule of interest in all cells Dimensions: $N \times 1$
- X - Cell state matrix made of the entire expression profile of each cell minus the molecule of interest. Dimensions: $N \times (D-1)$. The molecule of interest is removed from the cell state matrix to prevent any cell-cell interaction false positive due to signal spillover between cells, as well as trivial intrinsic effect.
- $d_{i,j}$ - euclidean distance between cell i and j
- K_{int} - cell-cell covariance for the intrinsic effect. Dimensions: $N \times N$
- K_{c-c} - cell-cell covariance for the cell-cell interaction effect. Dimensions: $N \times N$
- K_{env} - cell-cell covariance for the environmental effect. Dimensions: $N \times N$

With these notations, SVCA models the expression level Y across cells as shown on Equation 1

$$Y = \mathcal{N}\left(0, \mathrm{K_{int}} + \mathrm{K_{c-c}} + \mathrm{K_{env}} + \sigma_\varepsilon^2 \mathrm{I}_n\right) \tag{1}$$

1.2 Definition of covariance terms

Intrinsic effect: The intrinsic effect is the effect of the cell state on the expression level of the molecule of interest. In our framework, it is modelled with the linear covariance term shown on Equation 2 and it resembles the Kinship matrix in GWAS studies.

The matrix X describes the cell states. If cell types are available, each row of X $X_{n,:}$ for cell n may be an indicator variable for the cell type of cell n. However, it is often preferable to not rely on uncertain and binary cell types and use a richer measure of cell states. In this paper, we use the expression profile of the cell, excluding the molecule of interest. This enables to account for any intrinsic correlation between genes.

$$\mathrm{K_{int}} = \sigma_I^2 \mathrm{X}\mathrm{X}^T \tag{2}$$

Cell-cell interaction effect: The cell-cell interaction effect harvests the effects of the types or states of all neighbouring cells on the expression level of the molecule of interest. In the GP framework, it is modelled with the covariance term shown on Equation 3.

$$\mathrm{K_{c-c}} = \sigma_C^2 \mathrm{Z}\mathrm{X}\mathrm{X}^T\mathrm{Z}^T \tag{3}$$

where

$$\mathrm{Z}_{i,j} = f(d_{i,j}) = \begin{cases} \exp\left(\frac{-d_{i,j}^2}{2l^2}\right) \text{ if } i \neq j \\ 0 \text{ if } i = j \end{cases} \tag{4}$$

This covariance term is equivalent to a random effect model where, gene expression profiles of all neighbouring cells are used as predictor variables and the effect of a cell i on a cell j is weighted by a function of the distance between them: $Z_{i,j}$, shown on equation 4.

Environmental effect: The environmental effect aims at harvesting other local sources of variation in the cell micro-environment which are not measured in the data and have an effect of the expression level of the modelled gene. To model this unobserved source of variation, we use a Squared Exponential Kernel (Eq 5) for its non-linearity and its successful application to spatial expression data. [23]

$$\mathrm{K}_{\mathrm{env}} = \sigma_E^2 \exp\left(\frac{-d_{i,j}^2}{2l^2}\right) \tag{5}$$

1.3 Parameters' inference

The variance parameters of the SVCA model are optimised by maximising the log likelihood in Equation 6 [16, 13]. The scales of the covariance terms, $\sigma_I, \sigma_C, \sigma_E$ and σ_ε are optimised with gradient descent using a lbfgs optimiser, which updates the parameters iteratively through small steps along the gradient of the likelihood until it reaches a local optimum (null gradient) The length scale of the environmental and the local terms is optimised with a grid search strategy.

$$\begin{aligned} \mathscr{L} &= \log P(y|\sigma_I, \sigma_C, \sigma_E, l) \\ &= -\frac{1}{2} Y^T \cdot \mathrm{K}(\sigma_I, \sigma_C, \sigma_E, l) \cdot Y - \frac{1}{2} \log |K(\sigma_I, \sigma_C, \sigma_E, l)| - \frac{N}{2} \log 2\pi \end{aligned} \tag{6}$$

with

$$\mathrm{K}(\sigma_I, \sigma_C, \sigma_E, l) = \mathrm{K}_{\mathrm{int}}(\sigma_I) + \mathrm{K}_{\mathrm{c-c}}(\sigma_C, l) + \mathrm{K}_{\mathrm{env}}(\sigma_E, l) + \sigma_\varepsilon^2 \mathrm{I}_n \tag{7}$$

1.4 Estimates of variance components

Variance components for each effect are estimated using Gower factors[19, 12] (Equation 8).

$$\mathrm{Gower}(K_{\mathrm{effect}}) = \frac{\mathrm{tr}(\mathrm{PK_{effect}P})}{n-1}, \text{ with } P = \mathrm{I}_n - 1_n \tag{8}$$

The Gower Factor of a covariance term computes the expected variance of a random variable which is normally distributed with the considered covariance (Equation 9). In other words, the Gower factor of each covariance term of the SVCA model computes the amount of gene/protein variance across cells which is explained by the corresponding effect.

$$\text{For } Y \sim \mathcal{N}(0, \mathrm{K}),\ \text{Gower}(K) = \mathbb{E}(Var(Y)) \tag{9}$$

To compute the fraction of variance explained by each effect modelled in SVCA (intrinsic, environmental, cell-cell interactions and noise), we normalise Gower factors as shown on Equation 10

$$\mathrm{Var}_{\mathrm{eff}} = \frac{\mathrm{Gower}(\mathrm{K}_{\mathrm{eff}})}{\sum_{\text{other effects}} \mathrm{Gower}(\mathrm{K}_{\mathrm{eff}}) + \mathrm{Gower}(\sigma_\varepsilon^2 \mathrm{I}_n)} \tag{10}$$

This procedure enables us to break down the variance of every protein, across cells in the three effects of interest plus the noise.

1.5 Significance of variance components

In this section, we explain how we assessed the significance of the cell-cell interaction component, as this is the variance component of main interest for our study. However, the significance of other variance components can be assessed in the same fashion.

1.5.1 General Approach

We assessed the significance of the cell-cell interaction component using a log likelihood ratio (LLR) between the SVCA model and a reduced model omitting the cell-cell interaction component (Equation 11). Given that the reduced model is nested in the full SVCA model, we relied on Wilks' theorem [24], which states that if the null hypothesis is true (no cell-cell interactions), the LLR statistics should follow a χ^2 distribution. We estimated the parameter of this χ^2 distribution using simulated LLRs in the null hypothesis [5, 7]. LLRs from the true data were then compared to this null distribution.

$$Y = \mathcal{N}\left(0, \mathrm{K}_{\mathrm{int}} + \mathrm{K}_{\mathrm{env}} + \sigma_\varepsilon^2 \mathrm{I}_n\right) \tag{11}$$

1.5.2 Estimation of the null LLRs distributions

The simulation procedure is as follows. For all proteins of all images, we fitted the null model (Equation 11), and simulated data from the fitted normal distribution. We simulated 100 data points for each protein and each image and then fitted a χ^2 distribution to those using an off-the-shelf non linear optimisation method [11]. We then compared the LLR obtained, for each protein and each image, on the real data to the corresponding fitted χ^2 distribution and inferred a quantile-based p-value from this comparison.

1.5.3 Multiple testing correction and False Discovery Rate

For every 52 images and every 26 proteins, we computed a cell-cell interaction p value using the method described before and used the Benjamini-Hochberg procedure [4] to adjust p values for multiple testing. For each protein, we then counted the number of images in which the cell-cell interaction component was significant for a FDR threshold of 1%.

1.6 Comparison to related models

In this section, we compare SVCA to other published modelling framework to quantify cell-cell interactions in spatially resolved single cell molecular data.

1.6.1 Schapiro et al 2017 - HistoCAT

HistoCAT [17] aims at measuring spatial cooccurrence of different cell types. Briefly, cells of one or multiple images are classified into discrete cell-types based on their expression profile using a clustering algorithm. For every cell, a neighbourhood is defined as containing all cells within a fixed distance threshold (measured from membrane to membrane). Using this fixed neighbourhood definition, histoCAT counts the number of occurrences of a given pair of cell types, in the same neighbourhood. This number is then compared to a null distribution obtained from permuting the cells' positions, which gives a p-value for positive and negative cell types interactions.

Unlike SVCA, histoCAT does not measure the effect of these interactions on individual expression levels.

1.6.2 Battich et al 2015

Battich et al [3] uses a regression approach to measure the effect of the cell microenvironment on individual expression levels. Briefly, 183 features are collected, quantifying intrinsic cell properties and microenvironmental properties. Microenvironmental features namely account for local cell crowding, number of adjacent neighbours, intercellular space around the cell, as well as the molecular profile of the neighbours, based on a fixed distance threshold. The dimensionality of this feature set is then reduced using principal component analysis (PCA), and single cell expression profiles are modelled with a fixed effect linear model with the first 20 PCs as covariates. The PCs are then a posteriori linked to the microenvironmental features of interest. Biological replicates are used to quantify the amount of variance explained by each covariate using out of sample prediction.

This method therefore quantifies directly the effect of microenvironmental features including cell-cell interactions. Unlike SVCA however, it relies on a definition of discrete microenvironmental features and the definition of fixed parameters such as a distance threshold to define a cell's neighbourhood.

1.6.3 Goltsev et al 2018

Goltsev et al's [1] approach also relies on the definition of discrete microenvironmental variables, used in a fixed effect linear model to predict the expression level of individual markers out of sample. In contrast to Battich et al [3], microenvironmental variables are not defined directly based on the molecular profile of neighbouring cells, but based on the cell-type composition of the neighbourhood. The different neighbourhood cell-type compositions are clustered into discrete *i-niches*, used as a discrete input for the linear model.

This method therefore enables to quantify directly the effect of cell-cell interactions on individual molecular profiles of single cells. However, it again relies on a priori definition of microenvironmental variables, this time based on discrete cell-type assignments.

2 Model Validation

2.1 Simulations of the cell-cell interaction component

In order to be as realistic as possible, our simulations were based on real data from 11 images of the IMC dataset: real cell positions, cell states, intrinsic and environmental effects were used, and only the cell-cell interaction effect was rescaled for the purpose of the simulations.

Our workflow was as followed:

1. Fitting the SVCA model to the real dataset considered here (11 images and 26 proteins)
2. Simulating data from a multivariate normal distribution, with a covariance made of:
   - the intrinsic covariance from the fitted model
   - the environmental covariance from the fitted model
   - the noise covariance from the fitted model
   - a cell-cell interaction covariance which is a **rescaled version** of the one fitted to the data (Equation 12)
3. Reffiting SVCA to the simulated data, where the variance explained by cell-cell interactions is known from the rescaling step.
4. Comparing the variance estimates for cell-cell interactions with the ground truth.

Multiple experiments were performed, where the rescaling parameter in Equation 12 was set so that cell-cell interactions explain a proportion of variance ranging from 10% to 90% ($x \in [0.1; 0.9]$). Additionally, 10 repeat experiments were performed for each value of x, and for each image and each protein. The rescaling factor to use for a simulation where the environmental term should explain a proportion of variance $x \in [0.1; 0.9]$ is computed using Gower Factors (Equation 12).

$$\begin{aligned} Y &= \mathcal{N}\left(0, \hat{\mathrm{K}}_{\mathrm{int}} + k_{sim} \times \hat{\mathrm{K}}_{\mathrm{c-c}} + \hat{\mathrm{K}}_{\mathrm{env}} + \hat{\sigma}_{\varepsilon}^2 \mathrm{I}_n\right) \\ k_{sim} &= \frac{x}{1-x} \times \frac{\mathrm{Gower}(\hat{\mathrm{K}}_{\mathrm{int}} + \hat{\mathrm{K}}_{\mathrm{env}} + \hat{\sigma}_{\varepsilon}^2 \mathrm{I}_n)}{\mathrm{Gower}(\hat{\mathrm{K}}_{\mathrm{c-c}})} \end{aligned} \tag{12}$$

2.2 P-value calibration for the cell-cell interaction component

We used simulations to assess the p-value calibration for the cell-cell interaction component. For 11 random images and all 26 proteins of the IMC dataset, we fitted a null model (SVCA without the cell-cell interaction term) and simulated data from it. We then used the procedure described above to compute p-values for the null distribution, and computed the empirical false positive rate for multiple p-value thresholds (**Fig. 2**).

2.3 Results robustness

We used a bootstrapping strategy to quantify the uncertainty on the variance estimates. For a given image and a given protein, the model is fitted 5 times on a different subset of 80% of the cells. In order to visualise the robustness of SVCA signatures for all images at once, we then represented signatures computed from each bootstrap and each image in a low dimensional latent space using t-SNE in order to compare visually the distance between signatures coming from multiple bootstraps of the same image and signatures coming from different images (**Fig. S4-S12-S15**).

2.4 Out of sample prediction

We used a 5-fold cross-validation setting in order to perform predictions of the expression profiles of hold-out cells. In order to check whether every covariance term increased the predictive power of the model, we fitted in turn:

1. a model with only an intrinsic covariance to it.
2. a model with an intrinsic component and a local component.
3. the full model with all three terms

The predictions are then compared to the real expression profiles of the cells using a coefficient of determination r^2.

The results showed a consistent increase in predictive power for the more complex models for all proteins.

2.5 Identifiability of cell-cell interactions versus environmental effects

To better understand the identifiability of cell-cell interactions versus environmental effects, we compared the variance estimates of SVCA with the variance estimates of a reduced model which does not account for cell-cell interactions (Equation 11). Both models were fitted in the simulation setting described in the main text (26 proteins, 11 images and 10 repeat experiments). Variance estimates of SVCA and the reduced model were averaged across all proteins, images and experiments.

We then built a Sankey plot (**Fig. 2e**) to visualise what terms of the reduced model captured the residual variance from the lack of cell-cell interaction modelling. The width of the edges correspond to the increase in the variance estimates for the intrinsic effect, the environmental effect and the noise, from the SVCA model to the reduced model. This represents the redistribution of the cell-cell interaction component to other variance estimate from the SVCA model to the reduced model.

3 Downstream Analysis

3.1 Variance across images of the IMC dataset and patient effect

We also asked how the variability of the results across the IMC images related to the patient the biopsy came from (so called patient effect, **Fig. S3**). In order to do that, we fitted the following variance decomposition model for each protein and each effect from the spatial variance signature:

$$\text{eff}_{\text{prot}} \sim \mathcal{N}(0, \sigma^2_{\text{patient}} \text{K}_{\text{patient}} + \sigma^2_{\varepsilon} I_{n_{images}}) \tag{13}$$

where

$$\text{K}_{patient,(i,j)} = \begin{cases} 1 \text{ if images i and j come from the same patient} \\ 0 \text{ otherwise} \end{cases} \tag{14}$$

The percentage of variance explained by the patient effect was then computed with Gower factors.

3.2 Signature PCA

Unless otherwise specified, Principal Component Analysis plots are obtained from running PCA on the entire spatial variance signature for each sample. This signature is concatenated across all proteins and all effects. When specified, only specific effects are considered.

3.3 Gene Set Enrichment Analysis

When the number of genes allowed (seq-FISH), we ran a gene set enrichment analysis based on the ranking of the cell-cell interaction component of the spatial variance signature. The aim here was to understand the types of biological processes which are influenced by cell-cell interactions.

A simple procedure based on the idea of GSEA [22, 15] was implemented and used:

1. Genes are ranked based on the size of the cell-cell interaction component in the Spatial Variance signature
2. Gene sets are taken from the reactome database [8, 9] and are filtered for pathways including at least five genes of the gene set analysed in the seq-FISH data
3. A GSEA-like trace is computed for each pathway and the height of this trace is saved
4. Gene names are permuted 10,000 times and the same trace and trace height is computed on the permuted data in order to compute a null distribution of trace heights and thus a p-value for the pathway considered.
5. p-values are adjusted for multiple testing using a Benjamini-Hochberg procedure [4]

4 Dataset

4.1 Dataset description

4.1.1 Imaging Mass Cytometry (IMC) data

With IMC, the analysed tissue or cell culture is laser-disected into a sub cellular resolution grid of so-called voxels of dimension $1\mu m \times 1\mu m$. Every voxel of this grid is then analysed with cyTOF (anti-body based method), which results in protein counts of 26 proteins per voxel, which can be aggregated into single cell counts after cell segmentation [10, 21, 18, 6].

We analysed a dataset of 52 breast cancer biopsies imaged with Imaging Mass Cytometry coming from 27 patients [17]. 41 of these images are associated to clinical data:

- ER status
- PR status
- Her2 status
- Grade
- Biopsy location (periphery or centre)

These images contain between 267 and 1455 cells, with an average around 900 cells. 26 proteins counts are quantified at a sub cellular level (between 10 and 100 pixels/measurements per cell). These proteins are:

- CD44
- CD20
- CD3
- PRAB
- CAHIX
- EpCAM
- Erk12
- Ki-67
- CD68
- Fibronectin
- SMA
- Slug
- S6
- GATA3
- Twist
- Cytokeratin7
- ERa
- Vimentin
- Histone H3
- E-cadherine
- Cytokeratin-8
- cMyc
- Her2
- b-catenin
- Creb
- Cleaved Caspase 3

4.1.2 mer-FISH and seq-FISH data

Although mer-FISH and seq-FISH techniques differ slightly, the data produced and available online [14, 20] come in a similar format. Briefly, it comes as a list of detected individual RNA molecules, associated to a precise position on the tissue and the index of the cell each molecule belongs to (obtained with automatic cell segmentation). Summarising this data into a molecule count at the single cell level is therefore straightforward.

We analysed a mer-FISH dataset of 20 images taken on a single plate of breast cancer cell culture. Each image contained between 2500 and 2900 cells and 140 genes were measured. Additionally, we analysed

a seqFISH dataset consisting of 20 images of a single mouse hippocampus [20]. The images were taken in different regions of the hippocampus and 249 genes were measured.

4.2 Data processing

For the IMC dataset, single cell expression levels were computed by taking the median protein count across pixels. The median was chosen instead of the mean or the sum for its greater robustness against technical outliers produced by the experimental pipeline. For the mer-FISH and the seq-FISH datasets, single cell RNA counts were obtained directly from summarising the data described before.

In all cases, the data was then transformed with an Anscombe's transformation for variance stabilisation of Negative Binomial data [2]. The dispersion parameter ϕ in Equation 15 is optimised with gradient descent and the log transformation of Equation 16 is applied to the data.

$$\sigma^2 = \mu + \phi\mu^2 \tag{15}$$

$$y = \ln\left(x + \frac{1}{2\phi}\right) \tag{16}$$

The resulting signal is then normalised by regressing out the log of the total signal in the cell. This last step aims at taking into account local batch effects which would make some cells "brighter" overall.

Before fitting the SVCA model on a given protein or gene, the stabilised expression levels are subsequently raked standardised and transformed into normally distributed data using the probit function. This ensures a more robust fitting process due to a lesser sensitivity to outliers.

References

[1] [PDF]Deep profiling of mouse splenic architecture with CODEX multiplexed ..

[2] F J Anscombe. THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA. *Biometrika*, 35(3-4):246–254, December 1948.

[3] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of transcript variability in single mammalian cells. *Cell*, 163(7):1596–1610, December 2015.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, 1995.

[5] Petra Bůžková, Thomas Lumley, and Kenneth Rice. Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Ann. Hum. Genet.*, 75(1):36–45, January 2011.

[6] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, 7(10):R100, October 2006.

[7] Francesco Paolo Casale, Danilo Horta, Barbara Rakitsch, and Oliver Stegle. Joint genetic analysis using variant sets reveals polygenic gene-context interactions. *PLoS Genet.*, 13(4):e1006693, April 2017.

[8] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, Bijay Jassal, Steven Jupe, Lisa Matthews, Bruce May, Stanislav Palatnik, Karen Rothfels, Veronica Shamovsky, Heeyeon Song, Mark Williams, Ewan Birney, Henning Hermjakob, Lincoln Stein, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res.*, 42(Database issue):D472–7, January 2014.

[9] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res.*, 46(D1):D649–D655, January 2018.

[10] Charlotte Giesen, Hao A O Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods*, 11:417, March 2014.

[11] Steven G. Johnson. The nlopt nonlinear-optimization package. *?*, ?(?):?, ?

[12] Emrah Kostem and Eleazar Eskin. Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *Am. J. Hum. Genet.*, 92(4):558–564, April 2013.

[13] Christoph Lippert, Francesco Paolo Casale, Barbara Rakitsch, and Oliver Stegle. LIMIX: genetic analysis of multiple traits. May 2014.

[14] Jeffrey R Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U. S. A.*, 113(39):11046–11051, September 2016.

[15] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler, and Leif C Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, 34(3):267–273, July 2003.

[16] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Press, January 2006.

[17] Denis Schapiro, Hartland W Jackson, Swetha Raghuraman, Jana R Fischer, Vito R T Zanotelli, Daniel Schulz, Charlotte Giesen, Raúl Catena, Zsuzsanna Varga, and Bernd Bodenmiller. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods*, 14(9):873–876, September 2017.

[18] Peter J Schüffler, Denis Schapiro, Charlotte Giesen, Hao A O Wang, Bernd Bodenmiller, and Joachim M Buhmann. Automatic single cell segmentation on highly multiplexed tissue images. *Cytometry A*, 87(10):936–942, October 2015.

[19] Shayle R Searle. In *Matrix algebra useful for statistics (wiley series in probability and statistics)*, page 67. 1982.

[20] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron*, 94(4):752–758.e1, May 2017.

[21] C Sommer, C Straehle, U Köthe, and F A Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 230–233, March 2011.

[22] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, October 2005.

[23] Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. SpatialDE - identification of spatially variable genes.

[24] S S Wilks. The Large-Sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 9(1):60–62, March 1938.