Vesalius: high-resolution in silico anatomization of Spatial Transcriptomic data using Image Analysis

Patrick C.N. Martin[1,2], Cecilia Lövkvist[1,2], Byung-Woo Hong[3], Kyoung Jae Won[1,2,*]

[1]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark.

[2]Novo Nordisk Foundation Center for Stem Cell Biology, DanStem, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdams vej 3B, 2200 Copenhagen N, Denmark.

[3]Computer Science Department, Chung-Ang University, 84 Heukseok-ro Heukseok-dong Dongjak-gu, Seoul, Republic of Korea.

* To whom correspondence should be addressed to KJW (Tel: +45-3533-1419 ; Email: kyoung.won@bric.ku.dk)

Abstract: Characterization of tissue architecture is important to understand cell function and mechanisms *in vivo*. Spatial transcriptomics (ST), by recovering the transcriptome of a cell or tissue section while maintaining the two-dimensional location of the cell or tissue, can provide detailed insights into the cellular function in association with tissue architecture. Still, however, algorithmic development for ST has focused on identifying cells or spots with the same cell types located nearby and, therefore, cannot detect a tissue structure often composed of multiple cell types. Here, we present Vesalius to decipher tissue anatomy from ST data by converting transcriptomic information into a color code for image segmentation. Vesalius successfully detected tissue architecture in mouse embryo and brain from high resolution ST data by incorporating image processing algorithms. In contrast, previous spatial clustering approaches developed for low resolution ST data were unable to recover these structures. Intriguingly, Vesalius identifies genes linked to the morphology of tissue structures. In short, Vesalius is a tool to perform high-resolution *in silico* anatomization and molecular characterization from ST data.

Introduction

Organs and tissue are highly organized structures with numerous substructures and cell types. Cells function within anatomical structures or tissue organizations. In the context of development and cellular communication, understanding cell transcriptomic profiles in a given tissue or local environment will be highly useful.

One approach to investigate tissue and cellular organization is by using Spatial transcriptomics (ST)[1]. ST is a set of methods that recover gene expression while maintaining the spatial component intact. A number of ST methods have been developed in the last few years and fall into two categories: image-based approaches using fluorescence in situ hybridization (FISH) and sequencing based approaches using spatially resolved barcodes. Image-based ST approaches including seqFISH[2,3] or merFISH[4] provide cellular resolution ST. However, it is still challenging to capture a large number of mRNA species and genes of interest are usually pre-selected for image-based ST approaches. On the other hand, sequencing based ST approaches such as 10X Visium[5], high density spatial transcriptomics (HDST)[6] or Slide-seq[7,8] provide non-biased detection of mRNAs but with lower sensitivity compared with image based ST approaches. In this manuscript, we will focus on sequencing-based ST. The resolution of sequencing-based approaches relies on the density of spatial barcodes. For sequencing-based ST, therefore, a barcode is not guaranteed to contain mRNA transcripts from a single cell and may contain the transcriptome of multiple cells.

A number of tools including Seurat[9], BayesSpace[10] and Giotto[11] have been developed for the analysis of ST data. Seurat leverages refence single cell data sets to map cell identities to their respective location in ST data. While this approach demonstrates the cellular heterogeneity of tissues, the task of recovering and extracting anatomical regions is still challenging due to their cellular complexity. On the other hand, BayesSpace and Giotto provide distinct models both utilizing Hidden Markov Random Fields. Their respective methods attempt to cluster cells together by considering the transcriptome and local spatial neighborhood. Therefore, spots with a similar cell type spatially located nearby can be identified. They were successful in isolating rough territories from 10X Genomics Visium data[10,11]. However, these approaches are not designed to identify anatomical regions exhibiting heterogeneous cell type composition. Especially for the ST datasets with higher resolution[6–8], these local spatial clustering approaches recover countless patches composed of similar cell type. One solution to this challenge is to link anatomical territories from companion hematoxylin and eosin (H&E) staining images to their spatial transcriptomic

assay[12–14]. Still, the segmentation of anatomical territories remains challenging without manual annotation[15–17].

To address this limitation and perform *in silico* anatomization, we developed Vesalius that is designed to extract anatomical territories from sequencing based spatial transcriptomic data. The Vesalius algorithm converts ST assays into images by projecting transcriptomic data into the RGB color space. Image representation of ST assays enables the use of image processing techniques including dimensionality reduction and segmentation. Vesalius recovers known as well as previously uncharacterized anatomical structures in the mouse brain and embryo from SlideSeqV2 [8] and Visium data. Vesalius provides an effective analysis tool for the identification of sub-tissue architecture and the spatial gene expression in the identified tissue structure. Remarkably, Vesalius identifies the genes whose expression is linked to the morphology of the tissue structure and provides a new way to perform high-resolution anatomization from ST data.

Results

Vesalius embeds the transcriptome in the RGB color space

The core concept of the Vesalius algorithm is to represent the transcriptome of a barcode as a single color in the RGB color space (Fig.1a). First, Vesalius pre-processes sequencing based spatial transcriptomic data by log normalizing and scaling counts values, and extracting highly variable features. To convert the mRNA gene expression into a color code, Vesalius reduces the dimensionality of the data via principal component analysis (PCA) on the top variable features and converts loading values for each principal component (PC) into an RGB color channel (see Method for details). Alternatively, Vesalius can also embed 3D UMAP projections into the RGB color space. Vesalius handles uneven location of barcodes in the slide by expanding the color of a bead using Voronoi tessellation. The combination of tiles and color codes are embedded into image array structures (Fig 1a).

Next, Vesalius applies image analysis techniques in order to perform *in silico* anatomization (Fig. 1b). After balancing the color histogram and smoothing (see Methods), image segmentation based on k-means clustering is applied to produce color clusters that can be further subdivided into territories. In the case that barcodes belong to the same color cluster but are separated by too great of a distance in 2D, the Vesalius algorithm assigns them into a separate territory. The distance between territories is defined as a proportion of the maximum

possible distance between all barcodes. This approach ensures that both transcriptional similarity and location are considered in the determination of anatomical territories.

Uniquely to Vesalius, the isolation of anatomical territories and image representation of ST data enhances ST analysis by providing territory-based framework for analysis (Fig. 1c). Isolated territories can be further clustered to recover the finer details of cellular organization. Territories can be compared to investigate territory specific genes expression. Neighboring territories can be manipulated to recover tissue border gene expression and gene expression patterns arising within specific anatomical structures.

Vesalius recapitulates anatomical structures in mouse brain and embryo

To demonstrate Vesalius's ability to isolate anatomical structures, we used Slide-seq V2 that provides a high-resolution sequencing-based ST for mouse hippocampus and embryo[8]. Applying segmentation to the Slide-seq V2 mouse hippocampus data, Vesalius identified 59 anatomical structures in mouse hippocampus and the surrounding brain (Fig. 2a). Isolated territories are characterized by too little barcodes (<40) and too far away from another territory. Vesalius correctly isolated the Dentate gyrus, the CA pyramidal layer, and the Corpus callosum in the mouse hippocampus (Figure 2a-b - Image is from Allen Institute[18]). Applying Vesalius to Slide-seq V2 mouse embryo data (E12.5), we identified 17 regions including embryonic eye and liver (Fig. 2c).

Vesalius can uniquely detect tissue architecture from high resolution ST data

We assessed Vesalius together with Seurat[9] and BayesSpace[10] on Slide-seq V2 mouse hippocampus data (Fig 2a-c). Seurat does not provide a spatial clustering approach but rather maps cell identities to their respective location. On the other hand, BayesSpace explicitly acquires spatial clusters by integrating spatial neighbourhood into their model. While BayesSpace and Seurat display anatomical structures, these structures contain countless patches of homogenous cell populations. Structures containing heterogenous cell populations do not appear clearly (Fig 2d). For example, BayesSpace identified numerous sections in the inner CA field (as well as other regions). By contrast, Vesalius recovers this structure as a uniform territory (Fig 2a). These results are expected in that the algorithms developed so far were designed to detect spots with the same cell type. Spots with different cell type are usually separated and cannot be used as for tissue anatomy detection. These challenges are

unique to high resolution ST data and might not apply to lower resolution ST data such as Visium data. It is also of note that we were unable to run Giotto on Slide-seq V2 due to high computational time and memory faults.

Even though we observed that Vesalius can uniquely detect tissue architecture from Slide-seq V2 datasets, we additionally, compared the performance of Vesalius with Seurat[9], BayesSpace[10] and Giotto[11] using twelve 10X genomics Visium assays of human dorsolateral prefrontal cortex[20]. To compare the performance of each tool over the 12 samples, we assessed their ability to accurately recover manually annotated clusters using an Adjusted Rand Index (ARI) as well as their computational run time (Fig 2e-f). Vesalius outperforms both Seurat and Giotto and only slightly lags behind BayesSpace when assessed with Visium datasets (Fig 2e-f). It is of note that Vesalius is remarkably faster than BayesSpace and Giotto (Fig 2e, 30 to 50 times faster for a Visium dataset).

Tissue dissection highlights finer details of spatial patterning

Vesalius has the power to detect substructures from identified territories. For instance, in depth analysis of the isolated CA field (territory 18 in Fig. 2a) recovers all three CA field cell types namely CA1 pyramidal cells, CA2 pyramidal cells and CA3 pyramidal cells (Fig. 3a). UMAP projections of CA field clusters confirm that CA2 pyramidal cells have distinct transcriptional profiles (Fig. 2b). Vesalius identified *Pcp4*, *Rgs14* and *Necab2* as marker genes for CA2, which is consistent with recent proteomics study for CA2 against CA1[21]. The *in situ* hybridization (ISH) images against these genes taken from the Allen Brain Atlas[18] validated our prediction (Fig. 3c, Fig. S1). Further investigation identified that *Pcp4* showed a strong expression in the thalamus and a comparatively weaker expression in the CA2 layer (Fig. 3d). Other approaches using the transcriptome of all barcodes - even if they consider spatial neighborhood - may find it difficult to separate CA2 pyramidal cells (Fig. 2d). Territory isolation ensures that weak expression patterns are not lost in favor of overall stronger patterns by performing normalization, scaling and clustering on an isolated section of the ST assay.

Vesalius uncovers anatomical sub-compartments in mouse brain

Applying Vesalius, we found that the third ventricle and medial habenula territory revealed subtle and spatially driven expression patterns (Fig. 3e-f). The medial habenula is

characterized by two distinct clusters (Fig. 3e) and both clusters are spatially distinct (Fig. 3f, Medial Habenula and Medial Habenula – Low). These results suggest the medial habenula is compartmentalized with each region exhibiting a distinct expression profile. Additionally, there is a clear distinction between the upper third ventricle and the lower third ventricle (Fig. 3f).

To further investigate territory compartmentalization, we identified genes differentially expressed between compartments. In the medial habenula, we found 47 differentially expressed genes ($p<0.05$) including *Gabbr2* and *Calb2*. In comparison to *Nwd2*, a medial habenula marker, *Gabbr2* is localized in the upper compartment of the medial habenula while *Calb2* is localized in the lower compartment (Fig. 3g). This observation is consistent with a recent scRNAseq study followed by multiplexed FISH[22]**.** The third ventricle exhibited 109 differentially expressed genes between compartments. For example, *Nnat* is more strongly expressed in the lower third ventricle (Fig. 3f) and its expression pattern coincides with the ependymal cell layer that lines the ventricular system (Fig. 3h – left). Interestingly, we found a separate cluster of ependymal cells that distinguishes itself both in the UMAP projections (Fig. 3e) and spatial distribution (Fig. 3f). By contrast, the expression of *Enpp2* was located in the upper third ventricle (Fig. 3f) and is absent from the ependymal cell layer (Fig. 3h – right).

Spatially driven expression patterns in the embryonic eye

The Slide-seq V2 for embryonic eye (E12.5) (see Methods) also displayed subtle spatially driven patterns (Fig. 3i – right). Clusters associated with Anterior Lens Epithelial Cells show two distinct spatial patterns organized in a concentric fashion. UMAP projections suggest that there is a transcriptional shift occurring between clusters (Fig. 3i – left). This shift could describe transcriptional diversity within a cell type or even a previously uncharacterized cell type. We observed similar effect with Lens Vesicle cells where both UMAP clusters map to distinct regions in the embryonic eye.

We compared gene expression between Anterior Lens Epithelial cell layers and found 35 differentially expressed genes. For example, *Cryba4* was highly expressed in the inner layer while *Ccnd2* was expressed in the outer layer (Fig. 3j – top row). Similarly, differential gene expression analysis between Lens Vesicle cells returns 17 differentially expressed genes including *Pmel* and *Aldh1a1* (Fig. 3j).

Territory based differential gene expression reveals territory markers

Vesalius provides a convenient way to analyze differential gene expression between anatomical territories. As such, Vesalius may assist in the discovery of territory specific marker genes. We isolated the Dentate Gyrus – Granule Cell Layer (DG-GCL) and found 488 differentially expressed genes between the DG-GCL and the remaining hippocampus. Out of these 488 genes, we recovered many DG-GCL marker genes such as *C1ql2, Prox1, Lrrtm4*, and *Stxbp6*. Expression patterns and ISH images of *C1ql2* and *Stxbp6* support that Vesalius can easily recover anatomical territory markers (Fig. 4a).

Vesalius identified genes specific to the inner as well as outer layer of tissue structures

The detection of tissue territory enables investigation of genes expressed at the border or center of the tissue structure. For instance, barcode clustering of the DG-GCL revealed that this territory also included a thin layer of barcodes belonging to the Dentate Gyrus – Sub-granular zone (DG-SGZ) as seen in Fig. 4b. By comparing gene expression between both barcode clusters, we were able to extract tissue border specific expression patterns. For example, we found two genes *Cst3* and *Apoe* expressed at the border between the DG-GCL and the DG- SGZ. Allen Brain Atlas ISH images corroborate these results by displaying a higher expression of both genes at the border of the DG-GCL and the DG- SGZ (Fig. 4c). While increased expression of *Cst3* and *Apoe* could results from a high cellular density at the border, our results demonstrate how Vesalius can recover subtle spatially driven expression or cellular patterns and tissue border expression.

Vesalius highlights intra-tissue specific gene expression patterns

As anatomical territories are represented as images, we can apply morphological operators (see Methods) to isolated territories. Image representations of territories also lend themselves to territory manipulation such as territory layering (see Methods). We isolated, dilated and performed layer analysis for the Corpus callosum. Remarkably, we found high expression of *Stmn4* and *Kif5a* in the inner most layers (Fig. 4d). Mean expression of both genes gradually increased towards the core of the corpus callosum. The ISH images against Stmn4 taken from the Allen Brain Atlas showed a stripe of *Stmn4* expression (Fig. 4d). *Kif5a*

also shows a stripe and strong expression towards the center of the murine brain. These examples show that Vesalius can detect tissue morphology associated gene expression patterns.

Discussion

ST is a technology that can study transcriptomics of cells in association with tissue structure. ST has shown that the transcriptome of a cell is influenced by its spatial context and cellular micro-environment[23–25]. However, due to the cellular heterogeneity of tissues and the nature of sequencing-based ST, algorithmic development is essential to identify tissue structure from ST data.

Vesalius is an effective tool to perform *in silico* anatomization and molecular characterization without using supporting image data. Without using additional images, Vesalius successfully identified known tissue structures. Vesalius has the capacity to identify tissue structure marked by subtle expression changes. For instance, Vesalius identified CA2 pyramidal cell layers in mouse hippocampus (Fig. 3a-d), which can be lost without its use (Fig. 2d). Furthermore, Vesalius is sensitive enough to identify two medial habenula compartments (Fig. 3e-f) distinguished by 47 marker genes including *Gabbr2* and *Calb2*. It is of note that previous approaches required microdissection or systemic multiplexed FISH to identify the spatial distribution of these genes[21,22]. Vesalius enabled identification of these compartments from ST data without sophisticated experimental setup**.**

It is intriguing that Vesalius also identified genes expressed differentially in accordance with their location within the Corpus callosum. For instance, *Stmn4* and *Kif5a* showed expression levels gradually increasing towards the inner layer of Corpus callosum (Fig 4d). Vesalius also identified genes expressed in the outer layer of DG such as *Cst3* and *Apoe* (Fig 4c). Identification of layer specific genes would not be possible without using Vesalius.

In addition, Vesalius identified distinct spatial gene expression of *Cryba4*, *Ccnd2*, *Pmel* and *Aldh1a1* in the developing eyes (Fig. 3j). It is not clear if genes show spatial distribution depending on the location in the developing eyes, or if they may describe diverse cell sub-types. At least Vesalius is sensitive enough to capture this subtle spatial gene expression pattern.

The success of Vesalius stems from active incorporation of image processing techniques to the study of ST data. For this, Vesalius converted the transcriptomic landscape into 3 color channels. Instead of using unique colors for each identified cell types (which has been performed classically), Vesalius performed dimension reduction using PCA and assigned the PCs to the RGB color space. We show that image processing applied directly to gene expression data provides a convenient way to dissect anatomical structures from spatial transcriptomic data (Fig. 2). It is also of note that the dissection and isolation of anatomical territories prior to cell clustering provides further sensitivity. For instance, we found that the dissection of the CA field recovers of CA2 pyramidal layer (Fig. 3a-b), a layer that can be missed using conventional methods (Fig. 2d). Additionally, we show how the medial habenula and the third ventricle are both spatially compartmentalized despite being enriched with canonical markers associated with their respective tissue type (Fig. 3e-f). Furthermore, we illustrate how territory isolation can serve to recover tissue markers as described in the dentate gyrus (Fig. 4a). The image representation of territories enables straightforward ways to manipulate territories using morphological operators and territory layering. This approach can illustrate spatial variable gene expression within tissue sections (Fig. 4d).

Identification of tissue anatomy by Vesalius is unique compared with other algorithms developed for ST data. Seurat performs cell clustering and maps cell identities to their respective locations[9]. BayesSpace[10] extracts spatial clusters by incorporating local spatial neighbourhood into their model. In these cases, clusters represent spots with similar cell types and only recover tissue structures composed of homogenous cell populations (Fig. 2d). These spatial clusters can be distributed across entire tissue sections and do not necessarily represent a distinct anatomical structure. In this context, isolation of spatial domain and anatomical structure becomes impossible. While these approaches have been applied to 10X Genomics Visium data with low spatial resolution, they cannot be used for tissue architecture detection for higher resolution ST data. By contrast, Vesalius describes transcriptomic profiles in RGB color space and applies image processing techniques. The color associated to a territory represents the cellular composition (or transcriptional composition) of a tissue. Therefore, each territory may contain heterogenous cell populations and becomes available for in depth analysis of spatial patterning.

The tissue architecture that Vesalius identified well matched the previous annotation (Fig 2a-b), while the results from BayesSpace did not show clear tissue territories (Fig 2d). As it is not easy to quantitatively assess the performance using Slide-seq V2 datasets, we

used 12 Visium datasets with known annotation. Even with these lower resolution datasets, Vesalius showed a competitive performance compared with BayesSpace and better performance than Giotto. It is also noteworthy that the computing time for Vesalius is 30 to 50 times faster than other ST clustering algorithms. The running time is important as the spatial resolution increases: using Slide-seq V2, Vesalius completed in a fraction of the time taken by BayesSpace (~1 hour vs ~20 hours). Furthermore, Vesalius run time can be reduced by decreasing image resolution and utilizing Vesalius's parallel processing capabilities. For Slide-Seq V2 dataset, we could not even run Giotto successfully due to excessive running time and memory faults.

Extracting spatially resolved gene expression patterns has already been tackled by many and have shown how gene expression is often spatially compartmentalized[26–29]. However, these approaches cannot explicitly identify anatomical structures especially when a tissue is very heterogeneous. Comparing territories between each other or simply layering territories not only recovers spatially resolved gene expression patterns but also provides insight into the effect of cellular micro-environment. Indeed, while some genes might be spatially resolved, the transcriptome of the cells in which they are contained might vary significantly with respect to their localization and micro-environment. For instance, the expression of *Cst3* is wide spread but the dissection and manipulation of the dentate gyrus enable us to uncover its peculiar expression pattern at the border of the Dentate gyrus granule cell layer and the dentate gyrus sub-granular zone (Figure 4b-c).

The performance of image processing techniques is dependent on the quality of the images. Vesalius is best suited for high resolution spatial transcriptomic data[6–8]. Nevertheless, lower resolution methods such as 10X Genomics Visium[30] can also be analyzed by Vesalius. In fact, Vesalius is still able to outperform other tools and quickly provide anatomical territories (Fig 2e-f). However, the benefits of tissue dissection are diminished when sharp territory borders cannot be obtained. To compensate for lower resolution or lower throughput, Vesalius can utilize 3D UMAP projections instead of PCA loading values. UMAP projection further compress the data and may lose subtle expression patterns only visible in specific PC slices.

Image representation of territories provides a way of investigating spatial transcriptomic by putting the spatial component at the forefront of the analysis. Vesalius demonstrates that image processing applied to spatial gene expression is a promising avenue for the exploration of spatial transcriptomic data. Anatomical territories can be automatically

isolated and analyzed in great depth. Spatial territories provide insight into tissue markers as well as tissue border specific gene expression. Vesalius is a tool to perform high-resolution anatomization without sophisticated experimental setup.

References

1. Burgess, D. J. Spatial transcriptomics coming of age. *Nature Reviews Genetics* vol. 20 317 (2019).

2. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods* vol. 11 360–361 (2014).

3. Eng, C. H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).

4. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).* **348**, aaa6090 (2015).

5. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* vol. 353 78–82 (2016).

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

7. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science (80-.).* **363**, 1463–1467 (2019).

8. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).

9. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

10. Zhao, E. *et al.* Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* 1–10 (2021) doi:10.1038/s41587-021-00935-2.

11. Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 1–31 (2021).

12. Bergenstråhle, J., Larsson, L. & Lundeberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **21**, 1–7

(2020).

13. Giovanni Palla *et al.* Squidpy: a scalable framework for spatial single cell analysis. *bioRxiv* 2021.02.19.431994 (2021) doi:10.1101/2021.02.19.431994.

14. Peng, T., Chen, G. M. & Tan, K. GLUER: integrative analysis of single-cell omics and imaging data by deep neural network. *bioRxiv* 2021.01.25.427845 (2021) doi:10.1101/2021.01.25.427845.

15. Kurc, T. *et al.* Segmentation and Classification in Digital Pathology for Glioma Research: Challenges and Deep Learning Approaches. *Front. Neurosci.* **14**, (2020).

16. Gurcan, M. N. *et al.* Histopathological Image Analysis: A Review. *IEEE Rev. Biomed. Eng.* **2**, 147–171 (2009).

17. Vu, Q. D. *et al.* Methods for segmentation and classification of digital microscopy tissue images. *Front. Bioeng. Biotechnol.* **7**, 53 (2019).

18. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).

19. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* 1–10 (2021) doi:10.1038/s41587-021-00830-w.

20. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).

21. KJ, G. *et al.* Specific Proteomes of Hippocampal Regions CA2 and CA1 Reveal Proteins Linked to the Unique Physiology of Area CA2. *J. Proteome Res.* **18**, 2571–2584 (2019).

22. Hashikawa, Y. *et al.* Transcriptional and Spatial Resolution of Cell Types in the Mammalian Habenula. *Neuron* **106**, 743-758.e5 (2020).

23. Maniatis, S. *et al.* Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science (80-.).* **364**, 89–93 (2019).

24. Miller, B. F., Bambah-Mukku, D., Dulac, C., Zhuang, X. & Fan, J. Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomics data with nonuniform cellular densities. *Genome Res.* gr.271288.120 (2021) doi:10.1101/gr.271288.120.

25. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human

dorsolateral prefrontal cortex. *Nat. Neurosci.* 1–12 (2021) doi:10.1038/s41593-020-00787-0.

26. He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834 (2020).

27. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).

28. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: Identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).

29. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G. C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat. Biotechnol.* **36**, 1183–1190 (2018).

30. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* vol. 353 78–82 (2016).

31. Shekhar, S. & Xiong, H. Voronoi Tesselation. in *Encyclopedia of GIS* 1241–1241 (Springer US, 2008). doi:10.1007/978-0-387-35973-1_1464.

32. Barthelmé, S. & Tschumperlé, D. imager: an R package for image processing based on CImg. *J. Open Source Softw.* **4**, 1012 (2019).

33. Kanopoulos, N., Vasanthavada, N. & Baker, R. L. Design of an Image Edge Detection Filter Using the Sobel Operator. *IEEE J. Solid-State Circuits* **23**, 358–367 (1988).

34. Edgar, R. *et al.* LifeMap Discovery™: The Embryonic Development, Stem Cells, and Regenerative Medicine Research Portal. *PLoS One* **8**, 66629 (2013).

35. Franzén, O., Gan, L. M. & Björkegren, J. L. M. PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, (2019).

Acknowledgments

We would like to thank Konstantin Khodosevich for critical reading of the manuscript.

Funding

The Novo Nordisk Foundation Center for Stem Cell Biology is supported by a Novo Nordisk Foundation grant (NNF17CC0027852). This work is also supported by Lundbeck Foundation (R324-2019-1649, R313–2019–421) to KJW.

Author contributions

Conceptualization: PCNM, KJW

Methodology: PCNM, BWH, KJW

Investigation: PCNM, CL

Visualization: PCNM

Funding acquisition: KJW

Project administration: KJW

Supervision: KJW

Writing – original draft: PCNM, KJW

Writing – review & editing: PCNM, CL, BWH, KJW

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

Slide-seq V2 data sets were downloaded from the Single Cell Portal : https://singlecell.broadinstitute.org/single_cell/study/SCP815/highly-sensitive-spatial-transcriptomics-at-near-cellular-resolution-with-slide-seqv2#study-summary

Visium 10X DLPFC data was taken from Globus as specified by[20]

All ISH images used for validation were taken from the Allen Brain Atlas:

https://mouse.brain-map.org/

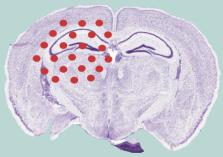
The Vesalius package and the code used for this analysis are available at: https://github.com/patrickCNMartin/Vesalius.

The Modified version of BayesSpace (see methods) to accommodate Slide-seqV2 data is available at:

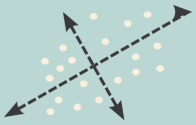
https://github.com/patrickCNMartin/BayesSpace

Marker gene tables are available in the supplementary data table: markerGenes.csv

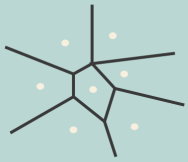
a. Embedding the transcriptome into RGB images



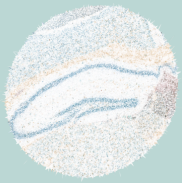
Spatial Transcriptomics



Dimensionality Reduction

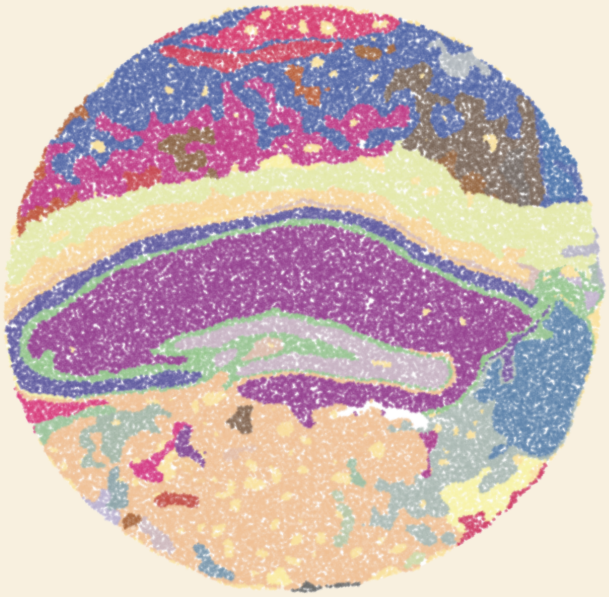


Coordinate Tiling



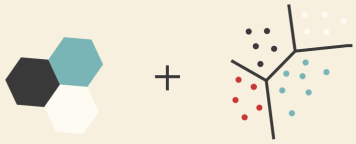
Embed Latent Space into RGB image

b. Image Analysis and Territory Isolation



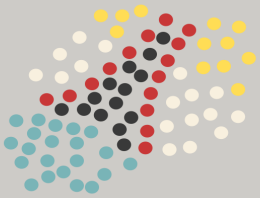
Histogram Equalization

Regularization & Smoothing



Segmentation & Territory Isolation

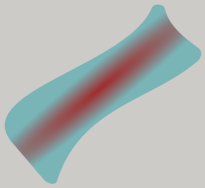
c. Territory based Spatial Transcriptomics Analysis



Territory Clustering

Territory Comparison

Tissue Border Expression



Morphology Driven Expression

Fig. 1. Describing tissue anatomy with Vesalius. (**a**) Vesalius embeds ST data into RGB colored images. This is achieved by preprocessing ST data with the intent of reducing the dimensionality of the data (via PCA or UMAP). The coordinates of each barcode are converted into rasterized tiles via the use of Voronoi tessellation. Finally, latent space values are embedded into the RGB color space and assigned to their respective tile. (**b**) Vesalius applies image analysis techniques to RGB images describing the transcriptional landscape of a tissue. Anatomical structures are isolated into separate territories. (**c**) Vesalius enables a territory-based ST framework including spatial territory clustering, territory comparison, tissue border expression, and morphology driven expression.

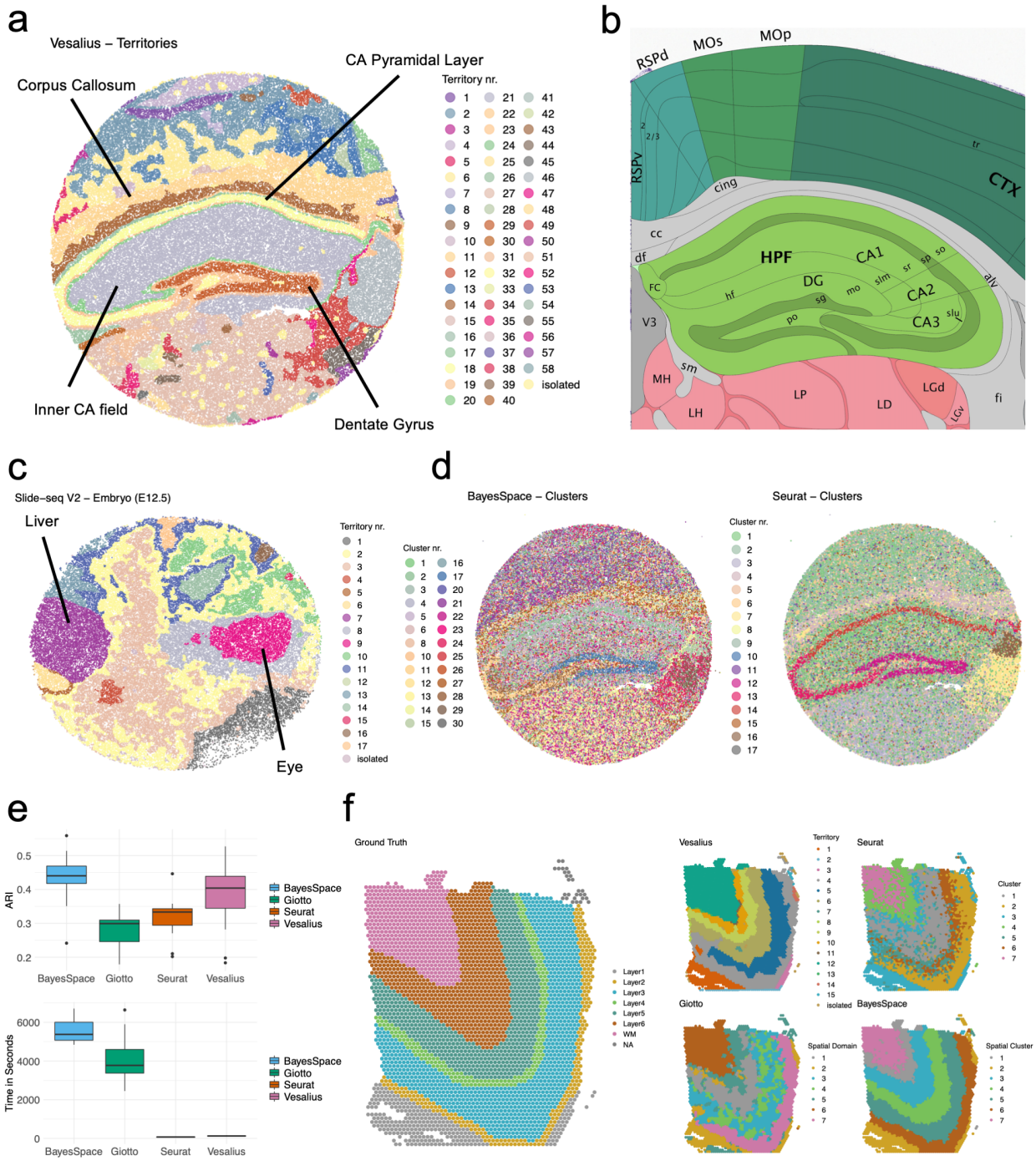


Fig. 2. Territory Isolation using Vesalius. (**a-b**) Vesalius accurately identifies anatomical structure in the mouse Hippocampus and surrounding brain. For example, Vesalius recovers the corpus callosum, the dentate gyrus, CA pyramidal layer and the inner CA field. Reference image taken from the Allen Brain Atlas [18] (**c**) Applied to Slide-seq V2 mouse embryo data, Vesalius recovers known structures such as the developing eye and the liver. (**d**) By contrast, only territories with homogenous cell populations are clearly distinguishable with BayesSpace or Seurat. (**e**) Vesalius performs well on the Visium 10X DLPFC data set compared to other

tools. While the ARI score for Vesalius is slightly lower than the best performing tool (BayesSpace), Vesalius compensates by being much faster. (**f**) Used as ground truth, DLPFC sample 151673 shows 7 clear layers. Layers as predicted by various spatial tool recover these layers with more or less accuracy. Vesalius outperforms both Seurat and Giotto.

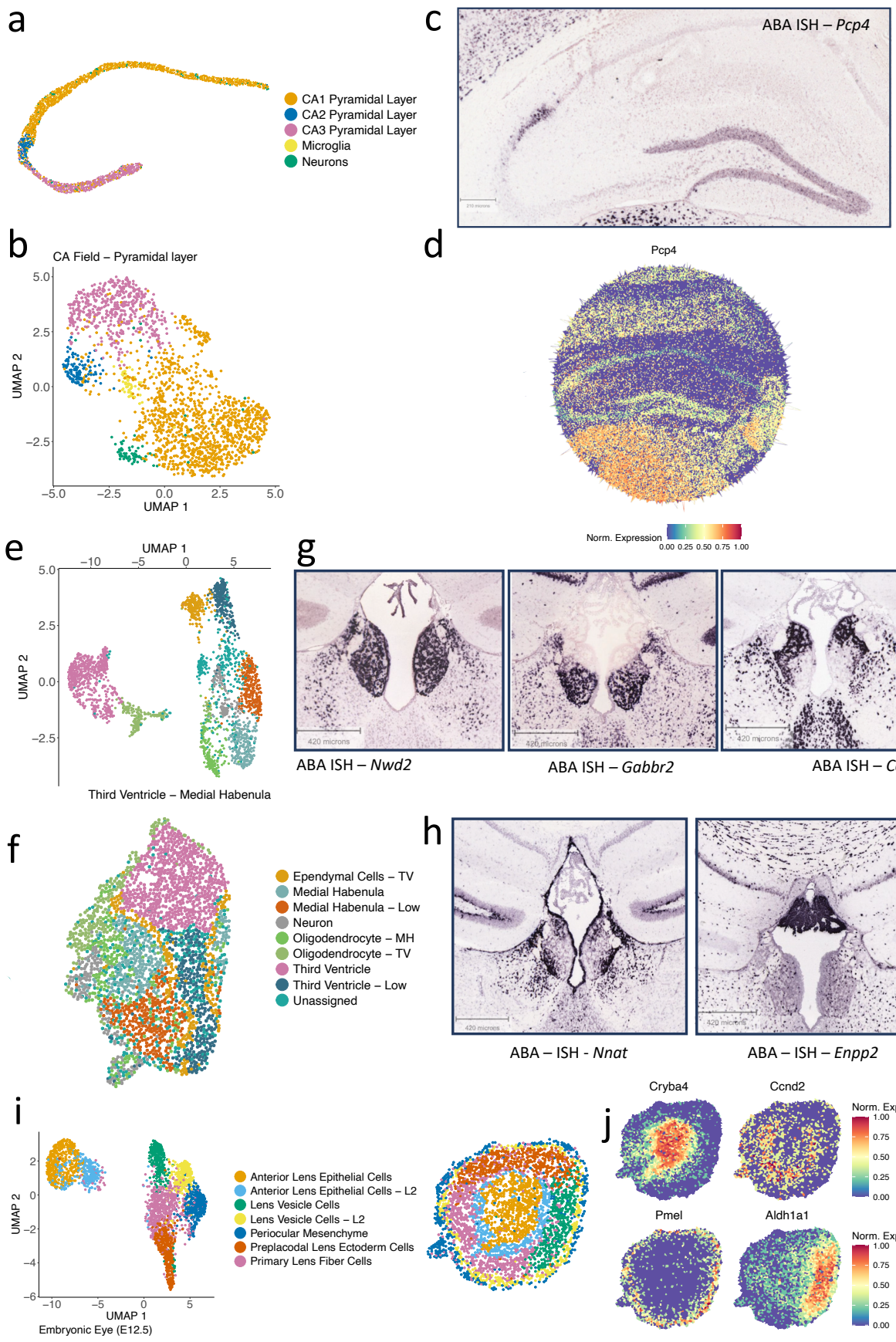


Fig. 3. In depth analysis of isolated territories. Clustering analysis can be carried out on isolated territories. (**a**) mapping of barcode clusters in the isolated CA field. Vesalius recovers all 3 CA pyramidal layers. (**b**) shows that all layers are indeed characterized by separated clusters in UMAP projections. (**c**) CA2 pyramidal layer was enriched with, among others, *Pcp4* a canonical CA2 layer marker. The ISH image taken from the Allen Brain Atlas corroborates the positioning of the CA2 layer within the isolated CA field. (**d**) *Pcp4* expression within the CA2 layer is lost in favor of stronger expression in the thalamus. (**e**) UMAP projections of the medial habenula and the third ventricle show that both tissues are characterized by distinct clusters. (**f**) mapping of barcode clusters in the isolated medial habenula and third ventricle show that the distinct clusters in (**e**) are also spatially distinct. (**g**) ISH image describing a medial habenula marker gene (*Nwd2*), a medial habenula lower compartment marker (*Gabbr2*), and a medial habenula upper compartment marker (*Calb2*). (**h**) ISH images of lower third ventricle marker (*Nnat*) and upper third ventricle marker (*Enpp2*). (**i**) UMAP projections and barcode cluster mapping of the embryonic eye (E12.5). (**j**) Differential gene expression analysis between Anterior Lens Epithelial Cell layers revealed that the expression of *Cryba4* was restricted to the inner layer while *Cnnd2* was expressed in the outer layer. Similarly, *Pmel* was expressed in the outer Lens Vesicle cell layer and *Aldh1a1* was expressed in the inner layer.

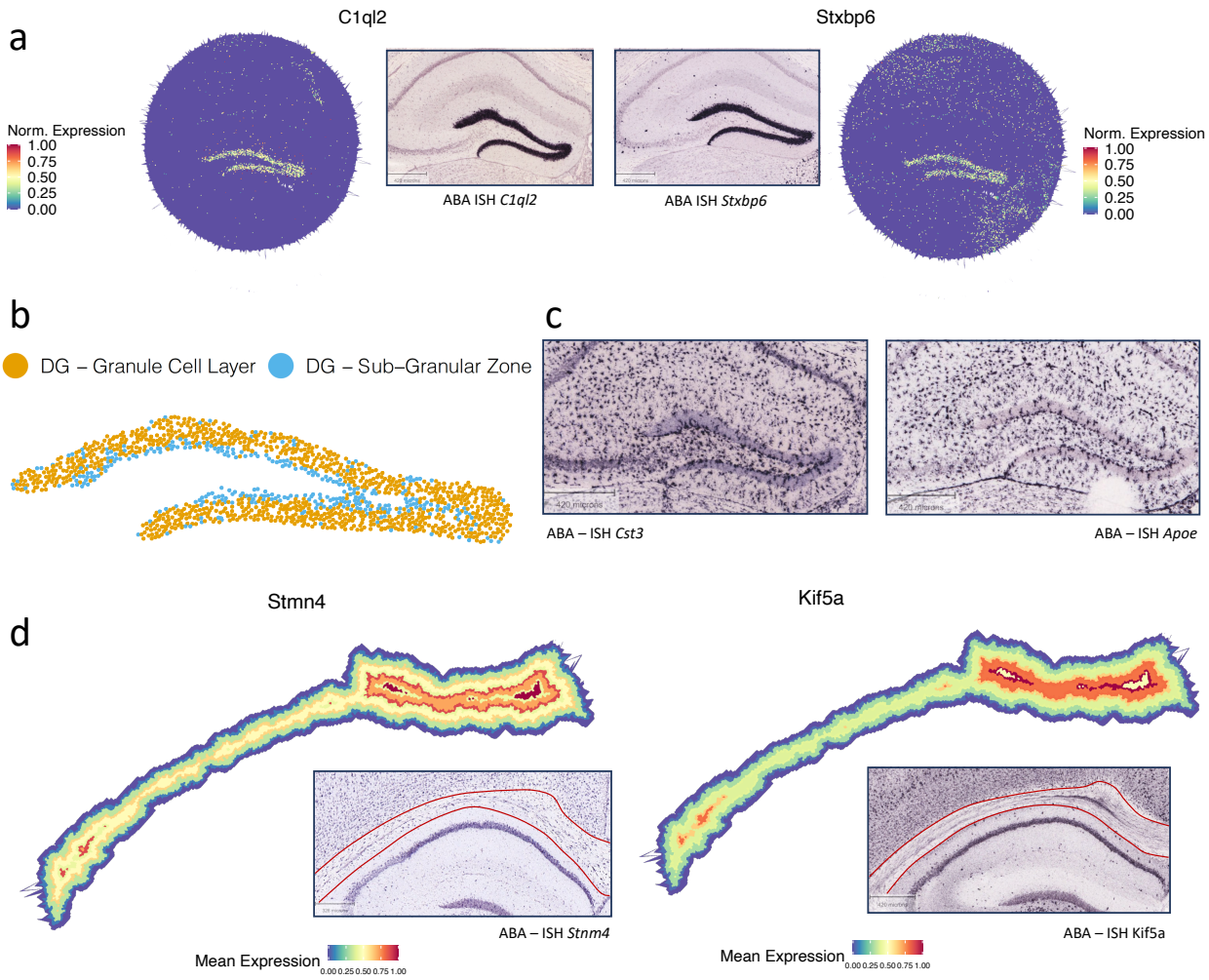


Fig. 4. Vesalius recovers territory marker genes, tissue border specific gene expression, and intra-tissue specific gene expression. (a) After extracting territory specific gene enrichment, we recovered many known Dentate Gyrus marker genes such as *C1ql2* and *Stxbp6*. ISH images taken from the Allen Brain Atlas corroborate these results. (b) Barcode clustering of the isolated dentate gyrus reveals transcriptional dissimilarity between Granule Cell layer and sub-granular zone. (c) Differential gene expression analysis between Dentate Gyrus layers displayed a high expression of *Cst3* and *Apoe* at the border between layers. *Cst3* and *Apoe* border expression is corroborated by Allen brain Atlas ISH images. (d) Layered expression pattern of *Stmn4* and *Kif5a* within the isolated Corpus Callosum show a higher expression at the center of the Corpus callosum. ISH images corroborate the spatial expression pattern of both genes. Corpus callosum contained within red lines.

Materials and Methods

Pre-processing

Pre-processing of data prior to Vesalius image building was handled by the Seurat package[9]. Slide-seq V2 data sets were loaded, log normalized and scaled using the default Seurat settings. Variable features (n=2000) were extracted prior to PCA. We used log normalization as the goal was to acquire general tissue territories for further analysis.

Vesalius – Color embeddings

The Vesalius algorithm embeds PCA loading values into the RGB color space. PCA is applied to the entire data set and slices are extracted from the PCA loading matrix. When referring to a "slice", we refer to 3 principle components in sequence starting from the first 3 PCs. Each slice always contains three PCs as one PC is required for each color channel (RGB). The first slice contains PC1 to PC3 and PC1 will be imbedded into the Red channel, PC2 into the Green channel and PC3 into the Blue channel. Every subsequent slice will follow the same principle.

For every barcode, we assign a combination of three numeric values one for each color channel.

$$B_{(i,c)} = \sum_{i=1} |L_{(i,c)}|$$

with $B_{(i,c)}$, barcode i in color channel c and L, the loading values associated to barcode i in color channel c. Notice that we take the absolute value of PCA loadings. In this case, we are only interested in variance and not *direction* of variance.

Barcode values are then ranged between 0 and 1 by min/max normalization. Barcode values then become:

$$B_{(i,c)} = \frac{B_{(i,c)} - \min(B_c)}{\max(B_c) - \min(B_c)}$$

Each color channel is normalized independently.

Vesalius also provides 3D UMAP embeddings. The principle remains similar as we extract and min/max normalize 3D UMAP projections for each dimension. Each dimension is assigned to a color channel. In the context of UMAP projections, there are no slices as all data is compressed into 3 channels.

Vesalius – Image processing and segmentation.

Once each barcode is assigned an RGB color code, Vesalius converts punctual coordinates into an image array. To overcome the sparsity of barcodes in 2D space, Vesalius computes a Voronoi diagram (also known as Dirichlet tessellation) for all barcodes and then rasterizes each tile[31]. Each barcode (and the associated RGB color code) is now described by all pixel contained in its tile. The overall resolution of the image array can be decreased by pooling barcodes that fall into the same pixel together.

Image arrays are handled by the *imager*[32] and *imagerExtra* R packages which contains a set of image processing method such as blurs, segmentation and image manipulation. Vesalius provides an iterative segmentation approach to reduce the color space of an image and extract territories. First, the image array is smoothed using either Gaussian blur, median blur, box blur or a combination of the aforementioned methods. Multiple smoothing rounds may be applied. Second, color values are clustered in k clusters using K-means clustering. The process of smoothing followed by clustering may be repeated over multiple values of k. In order to decrease computational time, the clustering is only applied to the center pixel value. The center pixel is defined by the pixel corresponding to the original barcode location before tessellation and rasterization. Vesalius also provides the option to smooth and segment images using all pixels instead. Using all pixels for segmentation produces sharper segments between homogeneous territories. However, this sharpness in homogenous territories may come at the cost of increased noise in heterogeneous territories.

Vesalius – Isolating Territories.

Color clusters and their associated barcodes are then separated or pooled into territories. Separation of territories occurs when color cluster territories are separated by a large enough (defined by the capture radius) distance in 2D space. Pooling of territories occurs if similar color cluster territories are not (necessarily) contiguous but in close proximity in 2D space. Barcodes in a common color cluster are pooled by capturing all barcodes that are within a capture radius of a seed barcode. The capture radius is defined as a proportion of the maximum possible distance between barcodes on the ST assay. This process is then applied to all captured barcodes until no more barcodes can been pooled into that local territory. If there are still

barcodes remaining in the color cluster, a new seed barcode is selected and the process is repeated until all barcodes have been separated/pooled into a territory.

It should be noted that this approach differs from the *k nearest neighbors'* algorithm as in this instance the capture radius is fixed and does not concern itself with number of neighbors.

Vesalius – Differential Gene expression and territory Markers

Marker genes and differentially expressed genes can be extracted from each territory. This process can be carried out in a threefold manner:

- Territory VS all other territories combined
- Territory VS all other territories individually
- Territory(ies) VS territory(ies)

In order to be considered for differential expression analysis, genes must pass a set of criteria. First, genes must be present in a certain percentage of barcodes in at least one territory (>10% of beads as default). Second, the log fold change must be above a certain threshold (logFC >= 0.25). It should be noted that this threshold is applied in the case of up-regulation as well as down-regulation. Remaining genes are tested for significant differential gene expression by using a Wilcoxon rank-sum test (Bonferroni corrected p-value < 0.05). Gene expression patterns can be visualized by using the *viewGeneExpression* function provide in the Vesalius package. Gene expression can be visualized over the entire slide or in an isolated territory. For visualization, gene expression is min/max normalized.

Vesalius – Territory dilation, erosion, filling and cleaning.

By using image representation of territories, Vesalius provides a convenient way to manipulate territories using image morphology. Vesalius encompasses dilation, erosion, filling, and cleaning into one function. We summarize image morphologies using a "morphology factor" described by a vector of integers. Positive integers increase territory size while Negative integers decreased territory size. Numerical vectors of positive and negative integers provide filling and cleaning morphologies.

Vesalius – Territory Layering and layered gene expression

Isolated territory layering is achieved by capturing territory edges and removing barcodes belonging to the edge and repeating the process until no more barcodes remain. First, the isolated territory is converted to a black and white image and X-Y sobel edge detection is applied[33]. All barcodes that share a pixel with the detected edge are pooled into a layer and removed from the territory. The edge detection process and pooling are repeated until all barcodes have been pooled into a layer. Differential gene expression between layers is carried out using a Wilcoxon rank-sum test (Bonferroni corrected p-value < 0.05 & logFC >= 0.2). Visualization of gene expression between layers is provided by the *viewLayeredExpression* in the Vesalius package. Layers are described by their normalized and averaged expression value.

Cell Clustering and Annotation

Clustering analysis of isolated territories was carried out using the Seurat package using default parameters. Clustering resolution was selected on the basis of UMAP granularity to ensure precise clustering while avoiding over clustering. Clusters and territories were manually annotated using their respective genetic markers. Markers were extracted from each cluster using the *FindAllMarkers* function provided by Seurat and the *extractClusterMarkers* function provided by Vesalius. *FindAllMarkers* compares clusters between each other while *extractClusterMarkers* compares clusters to all other barcodes. This distinction ensures that we recover subtle differences between cell types as well as territory specific gene expression.

We used the default Wilcoxon Rank-sum test for marker extraction. Identified markers were compared to Allen Brain Atlas[18] (https://mouse.brain-map.org/), the lifeMaps/geneCard database[34] (https://discovery.lifemapsc.com/in-vivo-development) or panglaodb database[35] (https://panglaodb.se/) to assign cell type to clusters and territories. Manual annotation of cell cluster was preferred over automated methods to ensure correct tissue annotations, rare cell type annotation and finally to maintain subtle spatially driven patterning.

Method comparison

To contrast territory isolation and cell clustering, we compared Vesalius to other spatial methods such as BayesSpace[10], Seurat[9] and Giotto[11]. BayesSpace required a minor modification to accommodate Slide-seq V2. We adapted the *.find_neighbors* function to select neighbors based on Euclidean distance in 2D space and select the 6 closest neighbors. We

added a new platform named "SS" for Slide-Seq. The forked and modified version of BayesSpace is available at: https://github.com/patrickCNMartin/BayesSpace.

Spatial methods were compared using the Visium 10X DLPFC data set taken from[20]. BayesSpace, Giotto, Seurat and Vesalius were run on all 12 samples and compute time was extracted for each sample. Tool performance was assessed using an Adjusted Rand Index (ARI). All code used for method comparison is available at: https://github.com/patrickCNMartin/Vesalius.

Supplementary Figures

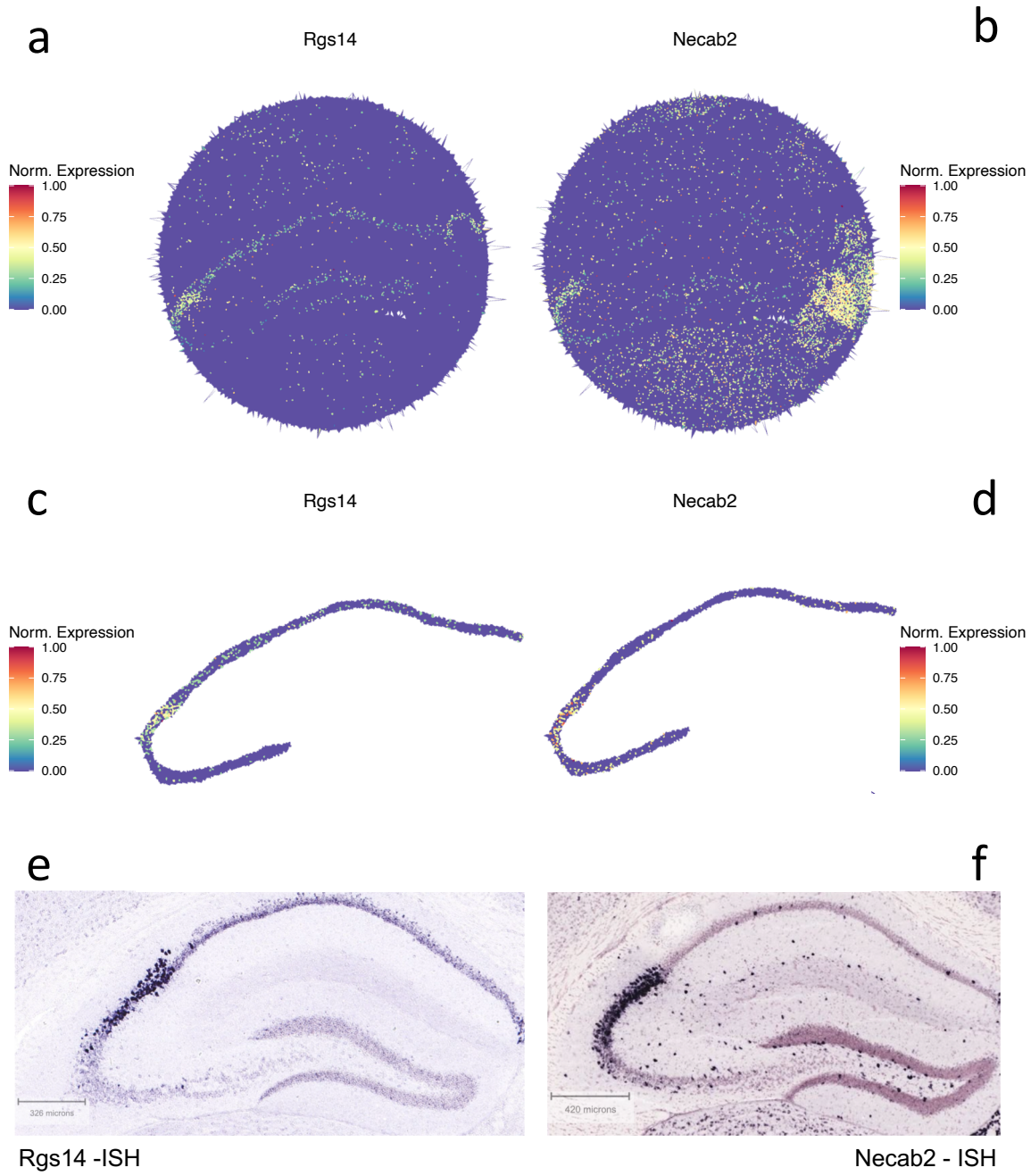


Fig. S1. Validation of CA2 field Marker genes. Vesalius recovered the CA2 layer and canonical CA2 layer marker genes such as *Rgs14* and *Necab2*. (**a**) shows that the overall expression of *Rgs14* is fairly weak throughout the entire but increase in the CA2 layer. (**b**) *Necab2* is only weakly expressed in the CA2 field. (**c**) and (**d**) show how the isolated CA field emphasizes the expression of both marker genes. (**e**) and (**f**) confirm that both genes are indeed expressed in CA2 layer as described by the ISH images taken from the Allen Brain Atlas.