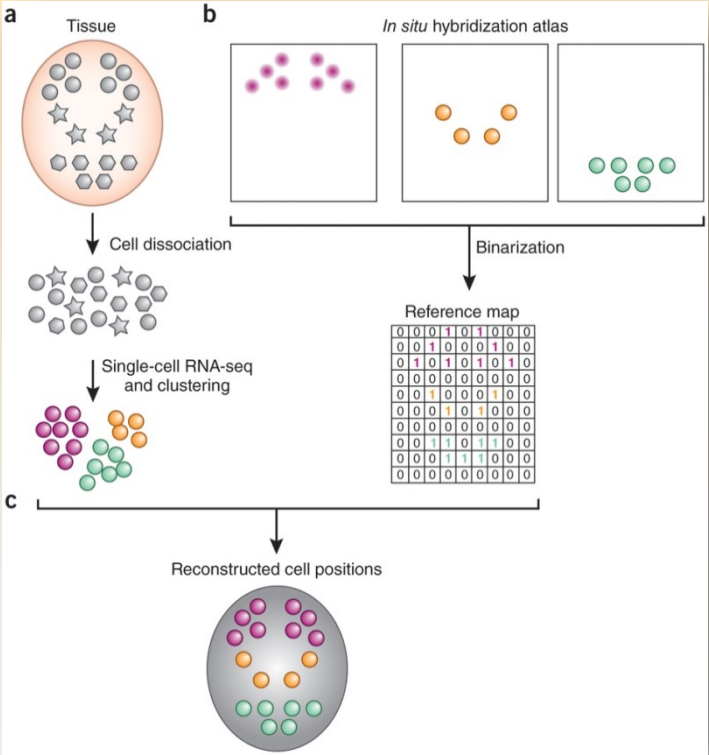
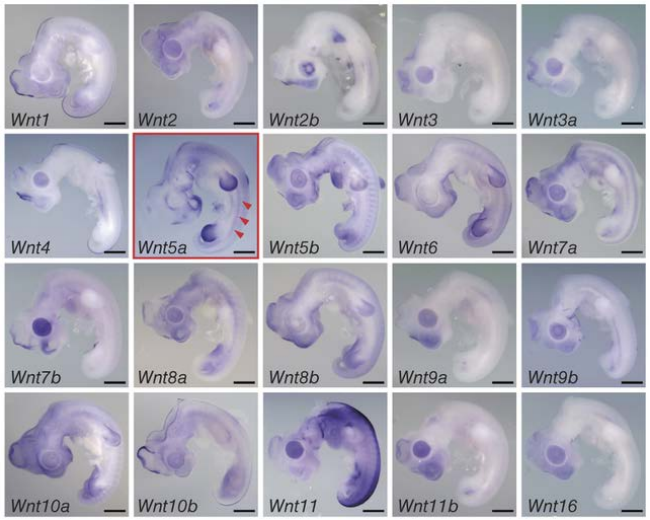
Novel computational methods to spatially map single-cell RNA-seq data to complex tissues

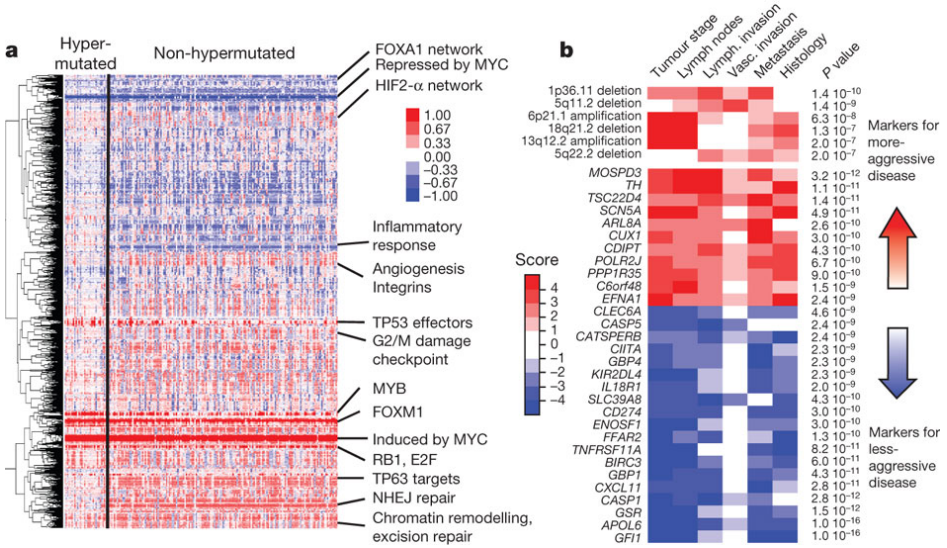


Audrey Fahrny
Technical Journal Club
26.06.2015

Faridani & Sandberg, Nat. Biotechnol. (2015)

Why study spatial heterogeneity in organisms





- Understanding the **biological significance** of complex cellular and tissue heterogeneity.
- Spatial context of gene expression information for cells is important for knowledge of cellular fates and function in **health and disease**.
- Understanding signaling networks
- Developmental biology

Wang et al. Nat. Genet. (2013); The Cancer Genome Atlas Network, Nature (2012)

Existing approaches to study cellular heterogeneity in organisms

Staining methods: **In situ hybridization** (ISH):

+ Allows gene expression to be assayed in many cells
- Limited to small number of transcripts/genes
- Marker analysis can localize only a handful of genes simultaneously within tissue section

Genomic profiling: **RNA sequencing** (RNA-seq):

+ Full transcriptome profiling
+ Single-cell resolution (scRNA-seq)
+ Global insight into cellular function, state and heterogeneity
- No spatial resolution→ lack info about cells' environment and localization

Additional general drawbacks:

- Selection bias: Rely on small set of predefined markers & cell purification
- Tissue processing → loss of signal

Existing experimental approaches for spatially resolved RNA-seq

Nat Methods. 2014 Feb;11(2):190-6. doi: 10.1038/nmeth.2804. Epub 2014 Jan 12.

Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue.

Lovatt D[1], Ruble BK[2], Lee J[3], Dueck H[4], Kim TK[3], Fisher S[4], Francis C[4], Spaethling JM[3], Wolf JA[5], Grady MS[5], Ulyanova AV[5], Yeldell SB[6], Griepenburg JC[6], Buckley PT[3], Kim J[7], Sul JY[3], Dmochowski IJ[2], Eberwine J[8].

Science. 2014 Mar 21;343(6177):1360-3. doi: 10.1126/science.1250212. Epub 2014 Feb 27.

Highly multiplexed subcellular RNA sequencing in situ.

Lee JH[1], Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM.

Existing experimental approaches for spatially resolved RNA-seq

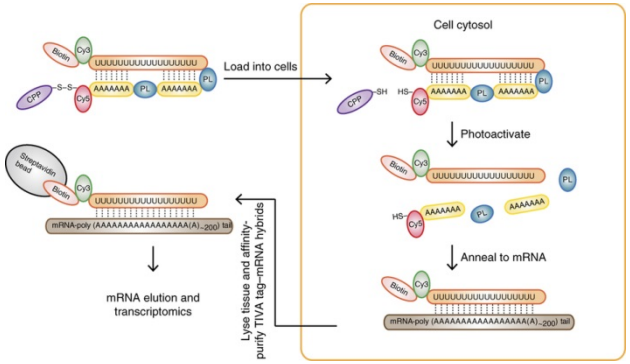
Nat Methods. 2014 Feb;11(2):190-6. doi: 10.1038/nmeth.2804. Epub 2014 Jan 12.

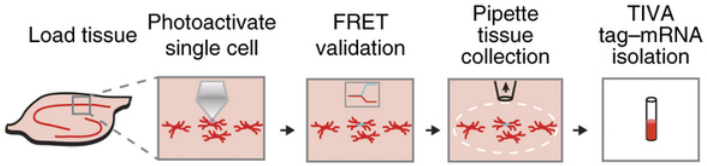
Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue.

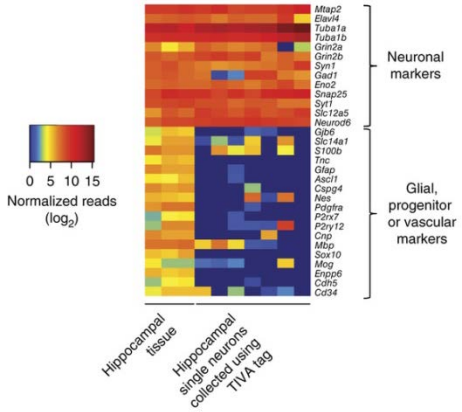
Lovatt D[1], Ruble BK[2], Lee J[3], Dueck H[4], Kim TK[3], Fisher S[4], Francis C[4], Spaethling JM[3], Wolf JA[5], Grady MS[5], Ulyanova AV[5], Yeldell SB[6], Griepenburg JC[6], Buckley PT[3], Kim J[7], Sul JY[3], Dmochowski IJ[2], Eberwine J[8].

Transcriptome in vivo analysis (TIVA): Photoactivatable TIVA tag enabling mRNA capture from single cells in live tissue, followed by RNA-seq.

+ Noninvasive approach for capturing mRNA from live single cells in their natural microenvironment.
+ Unambiguous determination of cells' spatial origin.
- Limited throughput: manual photoactivation and cell picking.
- Tag may exhibit selectivity to certain cell types.







Existing experimental approaches for spatially resolved RNA-seq

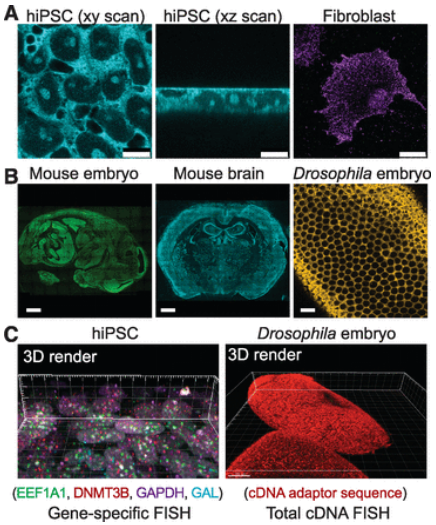
Science. 2014 Mar 21;343(6177):1360-3. doi: 10.1126/science.1250212. Epub 2014 Feb 27.

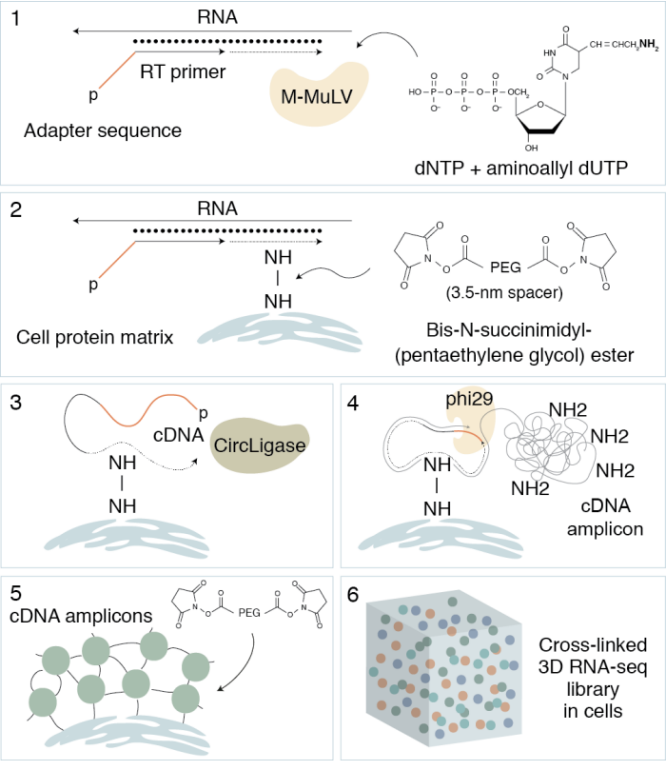
Highly multiplexed subcellular RNA sequencing in situ.

Lee JH[1], Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM.

Fluorescent in situ RNA-seq (FISSEQ): tagging RNA with random hexamers and carrying out RNA-seq in fixed cells.

+ Applicable to large variety of systems.
+ Reliable for analysis of small samples.
- Suboptimal for sampling larger tissues.
- Only cells present in same plane & close proximity can be assayed simultaneously.





Existing approaches for spatially resolved RNA-seq

TIVA & FISSEQ:

+ Unambiguous determination of cells' spatial origin.
- Applicable to variety of systems.
- Limited throughput/suboptimal for large tissues

Additional general drawbacks:

- Require highly specialized experimental tools.
- Do not yet offer widespread applicability of established scRNA-seq protocols.
- Currently of lower molecular sensitivity than scRNA-seq.

Computational approaches such as **Principal Component Analysis** (PCA): used to emphasize variation and bring out strong patterns in a dataset and partially recover spatial structure of tissues from single cell databases

+ Valuable for identification & characterization of cell types in mixed population.
- Give only very broad overview of spatial organization of assayed cells.
- Not well suited for spatially resolving novel cell types.

Spatial reconstruction of single-cell gene expression data

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier & Aviv Regev

Affiliations | **Contributions** | **Corresponding authors**

Nature Biotechnology **33**, 495–502 (2015) | doi:10.1038/nbt.3192

→ Seurat: computational model that infers cellular localization by integrating scRNA-seq data with *in situ* RNA patterns

Seurat

Seurat maps cells to their location by comparing the expression level of genes measured by scRNA-seq to their expression level in a tissue measured by ISH.

Model inputs:

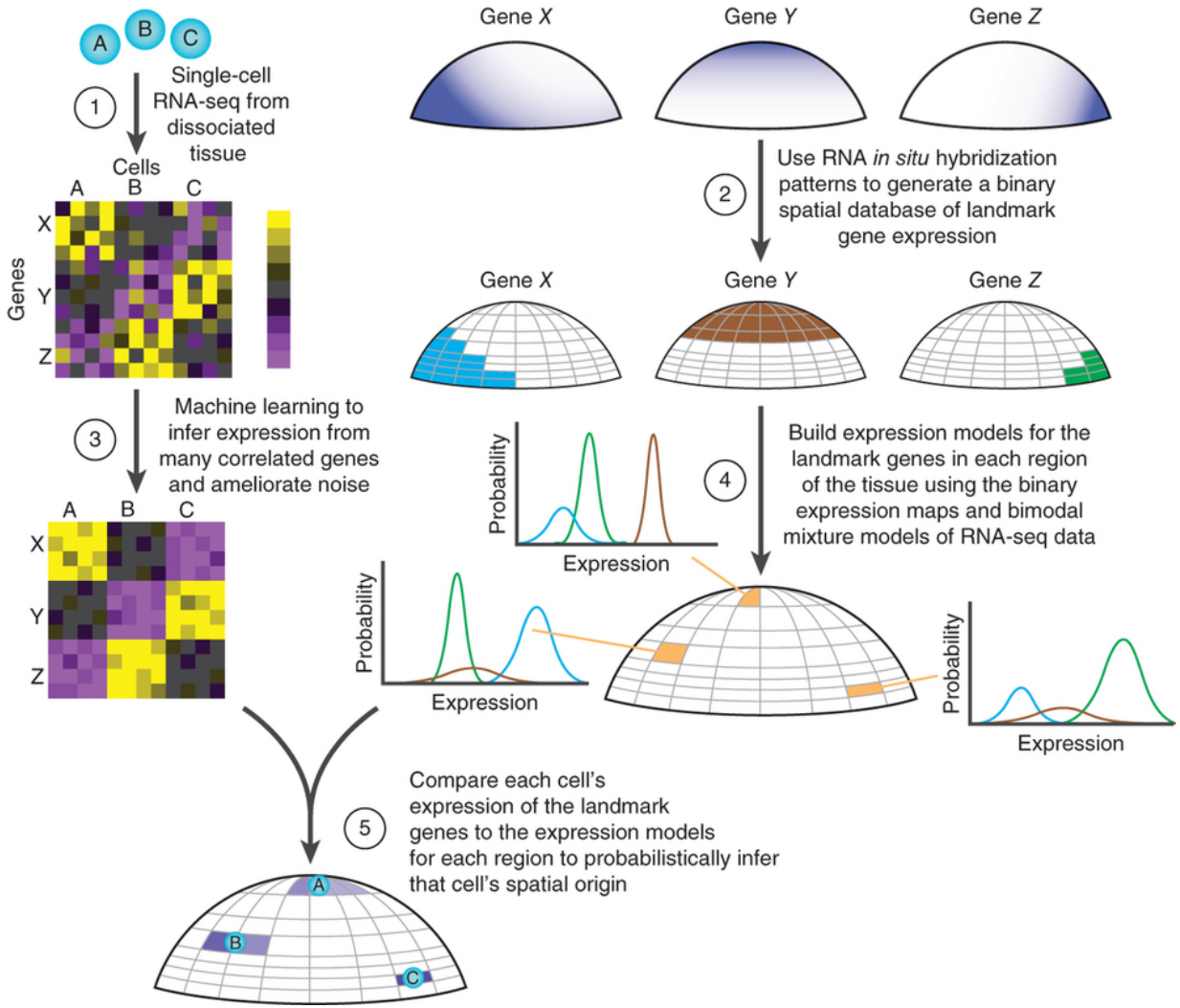
i. scRNA-seq data from dissociated cells

ii. ISH patterns for a small number of landmark genes

→ Subdivision of the tissue of interest into discrete spatial domains ('bins')

→ Landmark genes defined as 'on' or 'off' in each bin, as determined from published *in situ* stainings.

→ **Seurat then uses the single-cell expression levels of the landmark genes to determine in which bins a cell likely originated.**

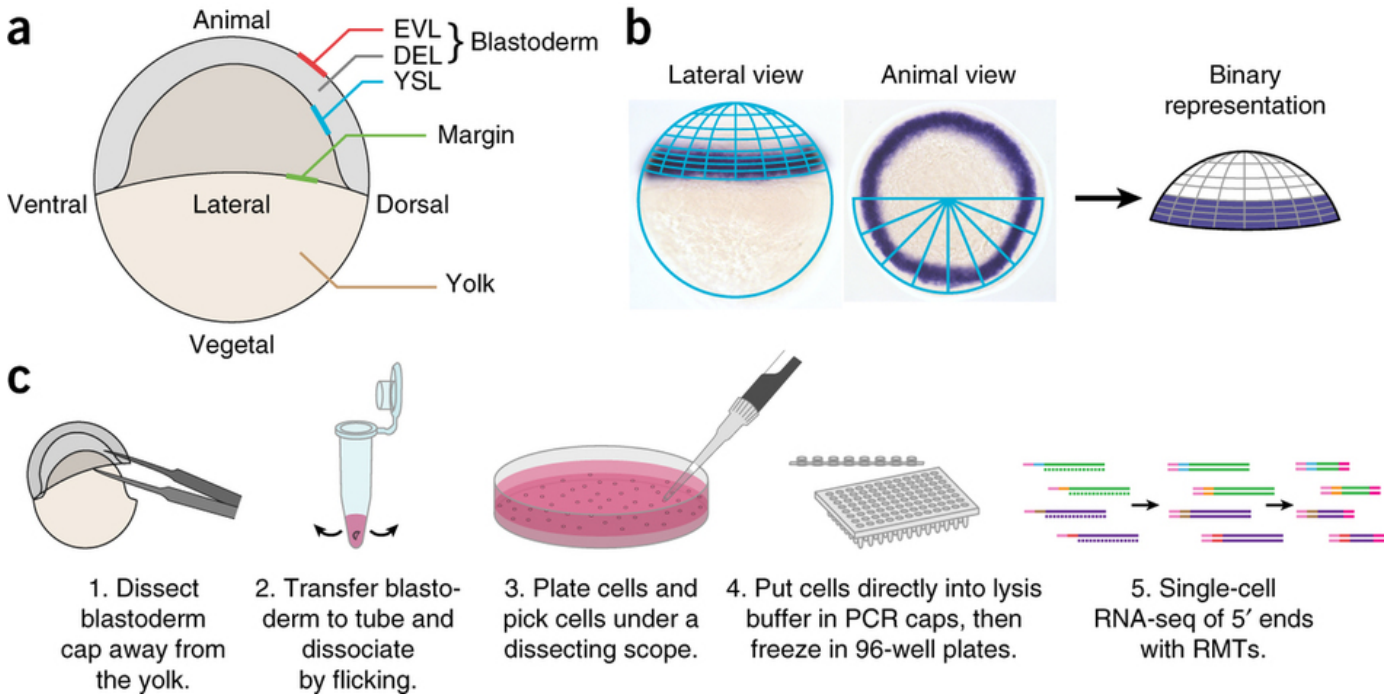


Application

- Applied to widely studied zebrafish embryo at the late blastula stage
- Extensive in situ patterns studied
- Applying Seurat to a data set of 851 dissociated single cells from zebrafish embryos → confirmed the method's accuracy and used it to predict and validate patterns where *in situ* data were not available and correctly localized rare cell population

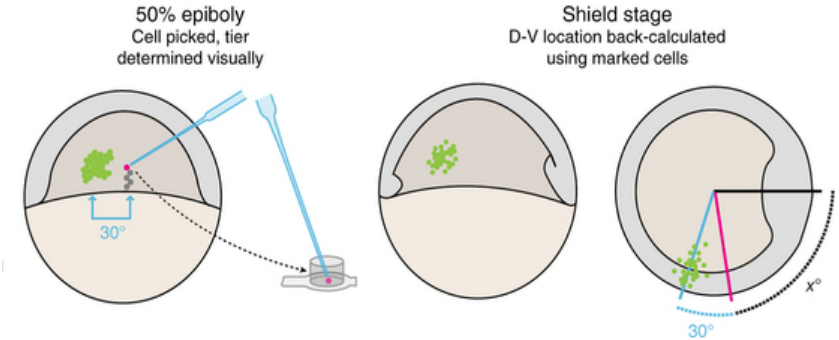
Workflow

1. scRNA-seq of 851 cells in developing zebrafish embryo.
2. Reference map constructed from colorigenic in situ data for 47 genes.
3. Run Seurat to determine cells' most probable localization in tissue of origin.



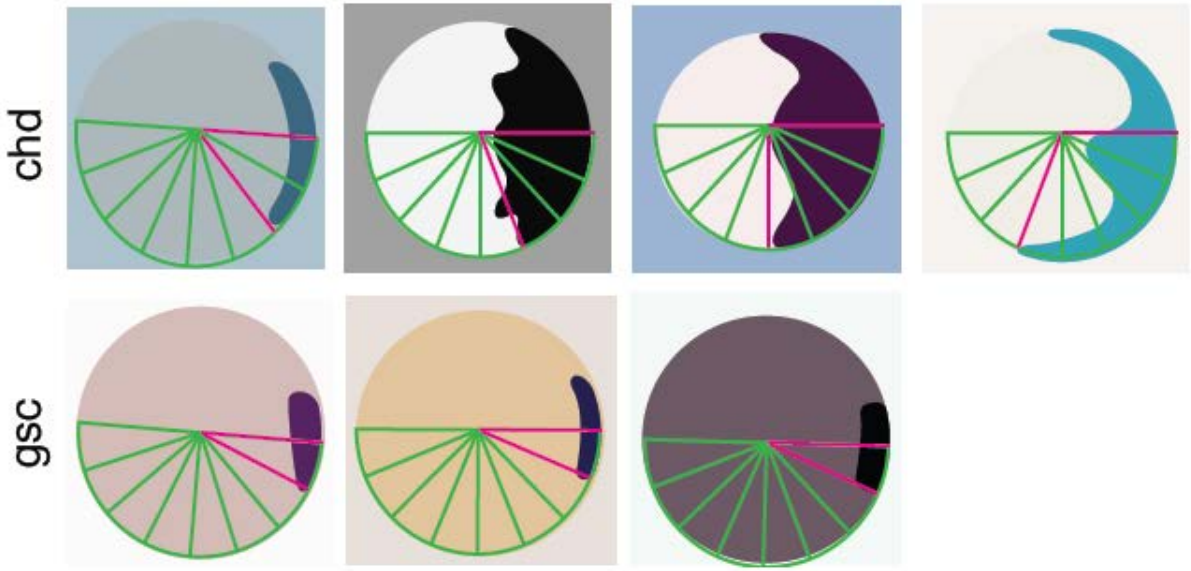
Controls:

- Cell seq. from biased dissociation protocol (enriches for embryonic margin cells).
- Cell seq. from manual isolation.



Building spatial reference map

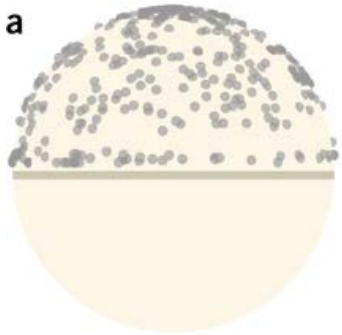
- Binary in situ land mark expression reference map from published data sets
- Variability in published data sets

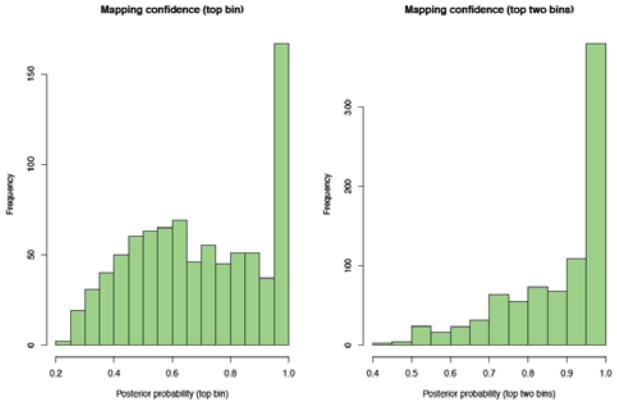


Validation: Spatial reconstruction of single-cell expression data

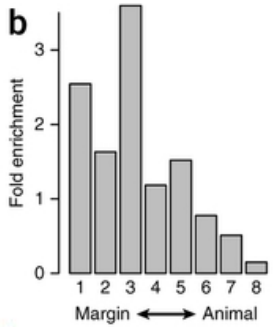
- Seurat maps cells throughout the embryo, consistent with the random distribution of the tissue
- Seurat **mapped majority of cells to 1-2 bins with high confidence** (p>0.9), (24% for a single bin, 59% for two bins, which are typically adjacent).
- **Control 1** (cells experimentally enriched for embryonic margin): Seurat's inferred locations overlapped considerably with the experimentally enriched area
- **Control 2** (manually isolated cells from intact embryos):
  - Seurat's inferred location within one bin of the registered location
  - Median distance error is 2 bins

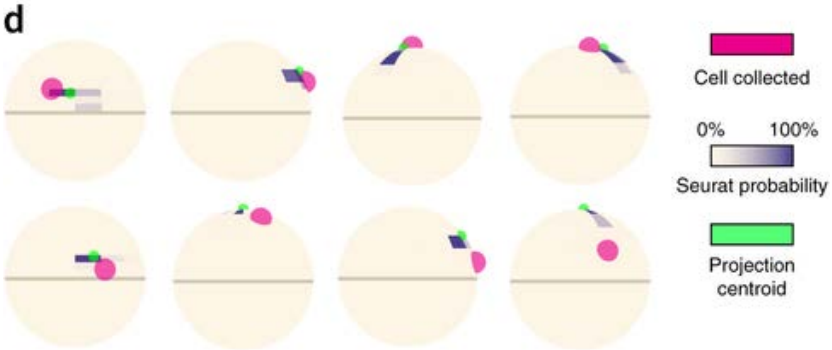
Cells from entire tissue

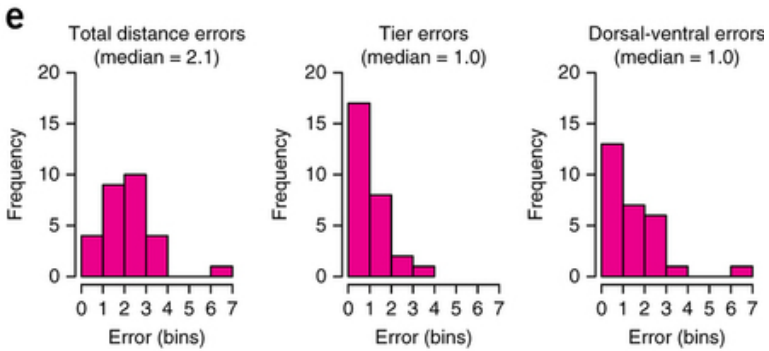




Cells from biased sampling (depleted for animal cap)



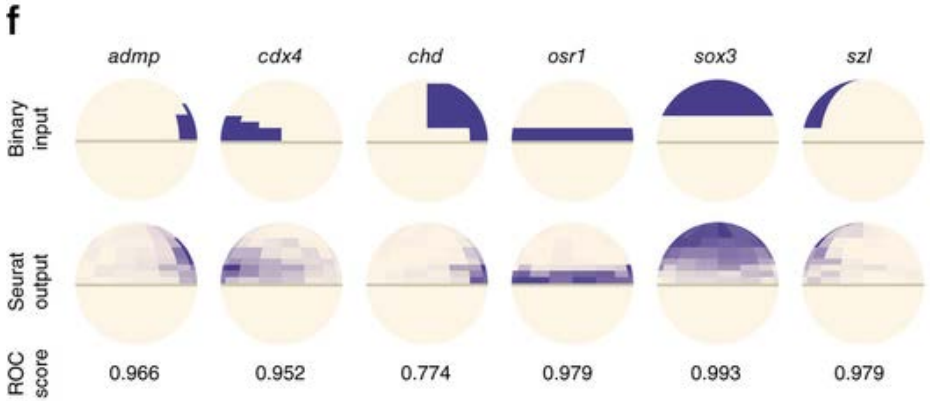


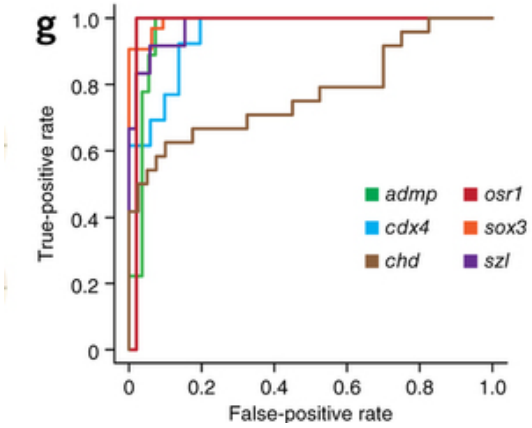


Validation: Spatial reconstruction of single-cell expression data

Re-inferred in situ pattern of landmark genes:

- Inferred patterns demonstrated remarkably **high overlap with experimental data** (median ROC = 0.96)
- 12 / 47 genes exhibiting near-perfect classification (ROC > 0.98).
- A rare subset of genes apparently performed poorly (e.g., *chd*) → literature revealed these genes had highly variable published *in situ* patterns.



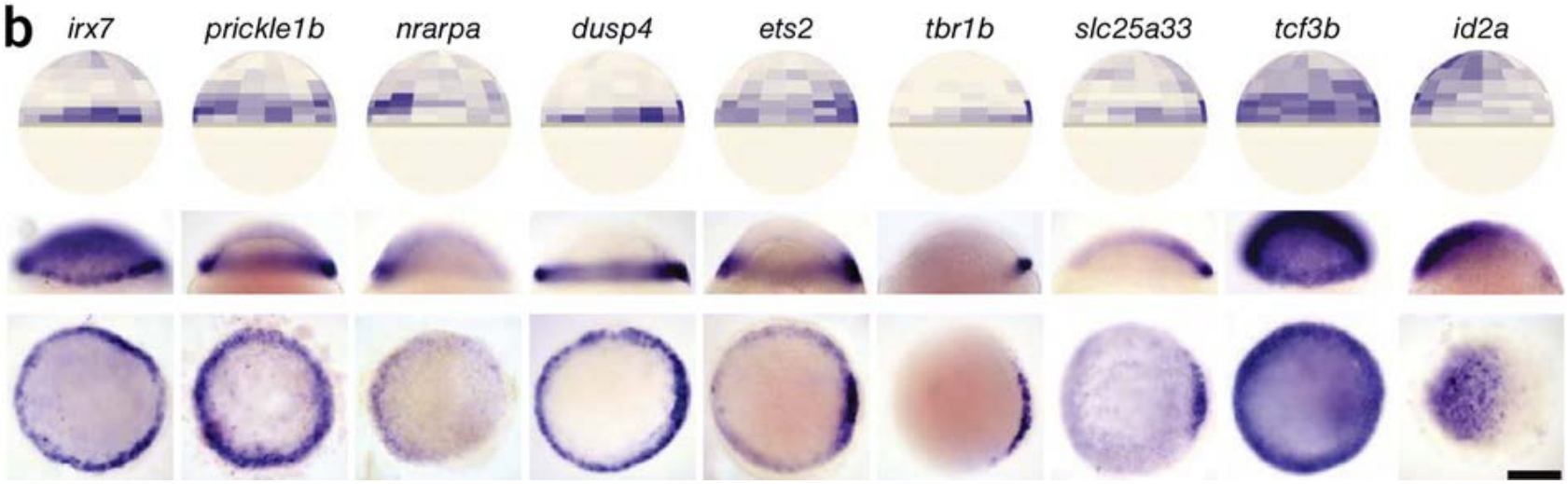


Seurat works even for genes with unknown expression patterns

Validation by RNA ISH of 14 genes without published expression patterns:

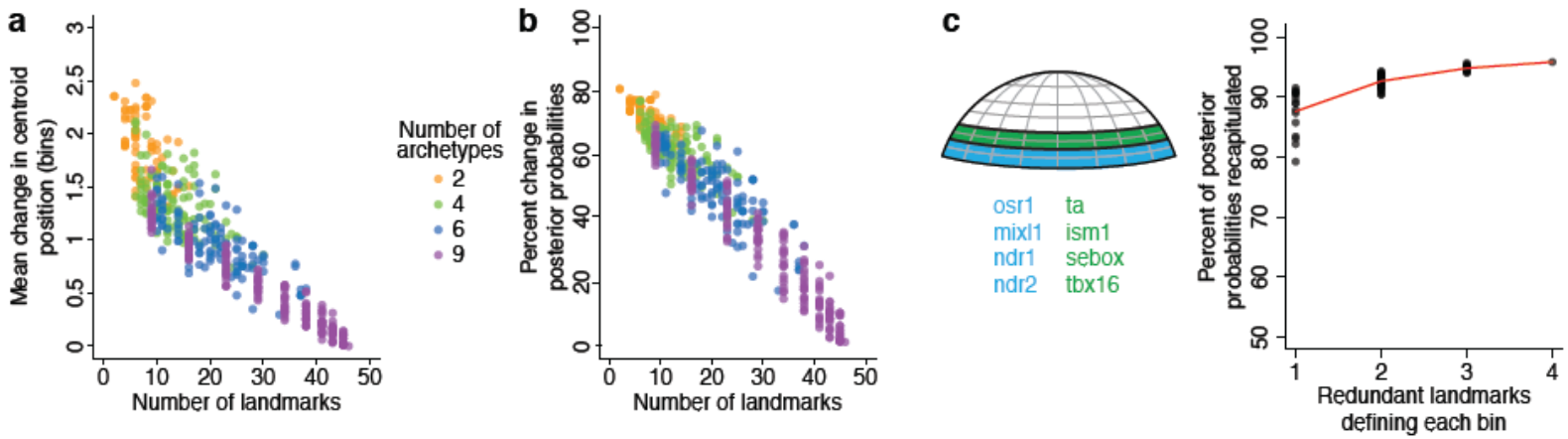
→ Experimentally determined *in situ* expression patterns exhibited overall high accordance with Seurat's predicted patterns

→ **Seurat can correctly transform scRNA-seq data into spatial predictions for genes whose expression patterns are not known**

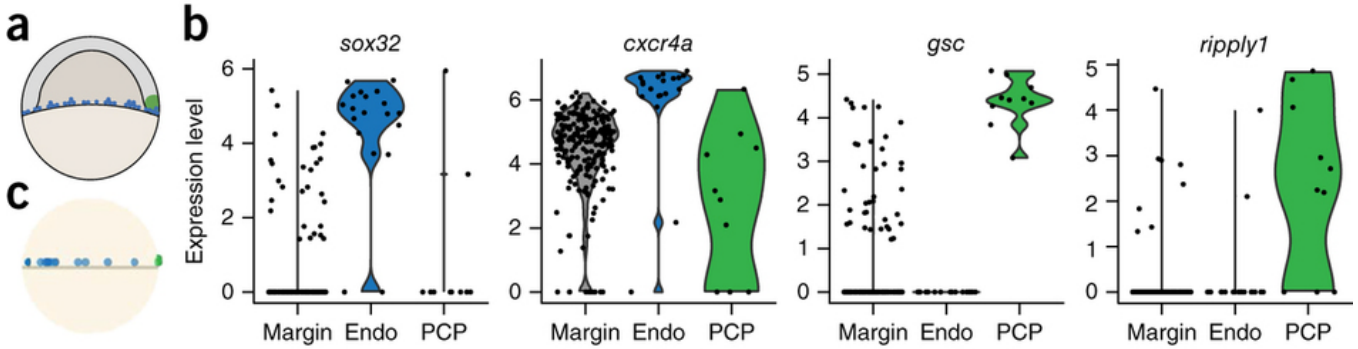


Spatially diverse landmark genes improve Seurat's mapping

- Stabilization of spatial mapping with inclusion of ≥30 landmark genes.
- Best when genes were sampled across all nine archetypes: spatially diverse landmark genes improve Seurat's mapping power.
- Having 2 genes with overlapping spatial expression patterns is valuable, additional redundancy has diminishing returns.



Seurat correctly localizes rare cell populations

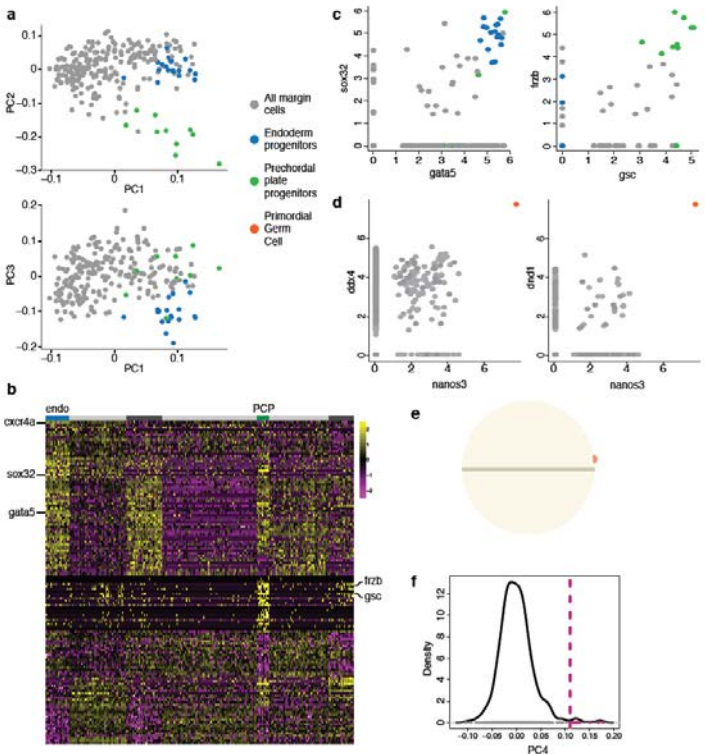


Green: Prechordal plate progenitors; ***Blue: Endodermal progenitors***; *Primordial germ cells (PGC)*

- 10 cells characterized by strong expression of the prechordal plate markers (*gsc* and *frzb)* → Seurat correctly mapped cells to dorsal-most embryonic margin.
- 19 putative endodermal progenitors defined by high expression levels of *sox32*, *cxcr4a* and *gata5* → Seurat scattered the endodermal progenitors across the lowest tier of the embryonic margin

- PGC cells (~1 per 500 cells in embryo) → Identified one cell that expressed extremely high levels of the canonical PGC markers *ddx4/vasa*, *nanos3*, and *dnd1* → Seurat mapped this cell to a mid-margin location, consistent with the distribution of these cells at this stage.

→ **Seurat successfully characterized the spatial distribution of known rare subpopulations with different characteristic localizations**.



Seurat discovers markers of rare subpopulations

- Used Seurat's spatial inferences in a **spatially aware marker selection strategy** (to avoid identifying boarder, nonspecific markers of the embryonic margin).

→ Successfully rediscovered multiple well- characterized prechordal plate progenitor markers and also found candidate markers that were not previously annotated in the prechordal plate, including *ripply1* and *ptf1a*.

ISH to validate new marker gene:

→ *In situ* hybridization for *ripply1* agreed with Seurat's prediction,

→ *ripply1*/*gsc* double *in situ* hybridization showed that *ripply1* is expressed only in a subset of *gsc*-expressing cells.

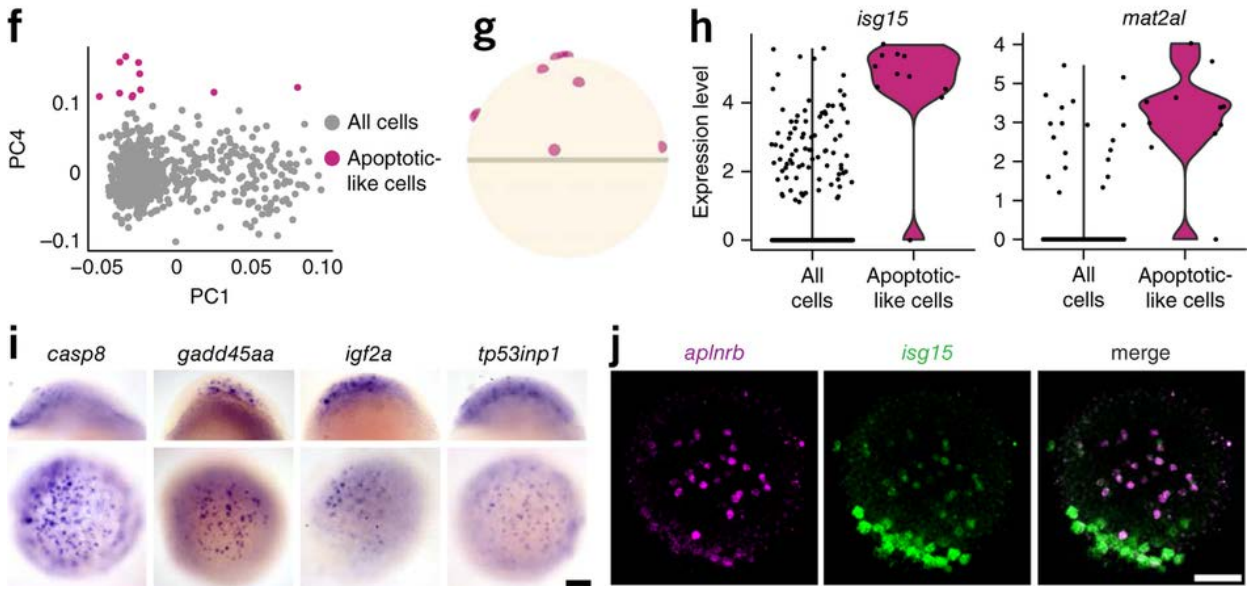
→ *ripply1* is a bona fide marker of the prechordal plate progenitors at 50% epiboly

→ **Spatially aware approach discovers markers of rare subpopulations**.

Seurat identifies dispersed, rare cell populations

- Searched for potentially novel subpopulations present in RNAseq data set → **12 cells** expressing genes hallmark of **apoptosis**, **cellular stress** and **cell signaling.**
- Seurat mapping: Apoptotic-like cells scattered throughout developing embryo, originating more frequently toward animal and ventral poles
- Not an artifact: cells identified in 10 separate embryos and in each experimental batch
- Number and specific locations different for each embryo, consistent with stochastic localization.
- *In situ* analysis for *foxo3b*, *aplnrb* and *isg15* interdependently confirmed their individual scattered expression.

→ **Identification of previously uncharacterized and stochastically localized population of "stressed cells"**.



Summary: Evaluation of Seurat's performance

- Seurat can transform scRNA-seq data into spatial predictions for both genes with known and unknown expression patterns
- Discovers markers for rare populations
- Identifies dispersed, rare cell populations

Limitations

- Seurat relies on the **spatial segregation of gene expression patterns** in a tissue in order to construct a reference map

→ may be challenging to apply it to tissues such as tumors where there is no guarantee of reproducible spatial patterning, or to tissues where cells with highly similar expression patterns are spatially scattered across a tissue (e.g., the adult retina).

High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin

Kaia Achim, Jean-Baptiste Pettit, Luis R Saraiva, Daria Gavriouchkina, Tomas Larsson, Detlev Arendt & John C Marioni

Affiliations | **Contributions** | **Corresponding authors**

Nature Biotechnology **33**, 503–509 (2015) | doi:10.1038/nbt.3209

Overview

- Approach: combines previously generated ISH-based gene expression atlases with unbiased single-cell transcriptomics
- Distinct cell types can be determined solely by expression of a few highly expressed transcription factors
- Applicable to any system with a reference gene expression database (RNA RISH data) of sufficiently high resolution

Spatial mapping approach

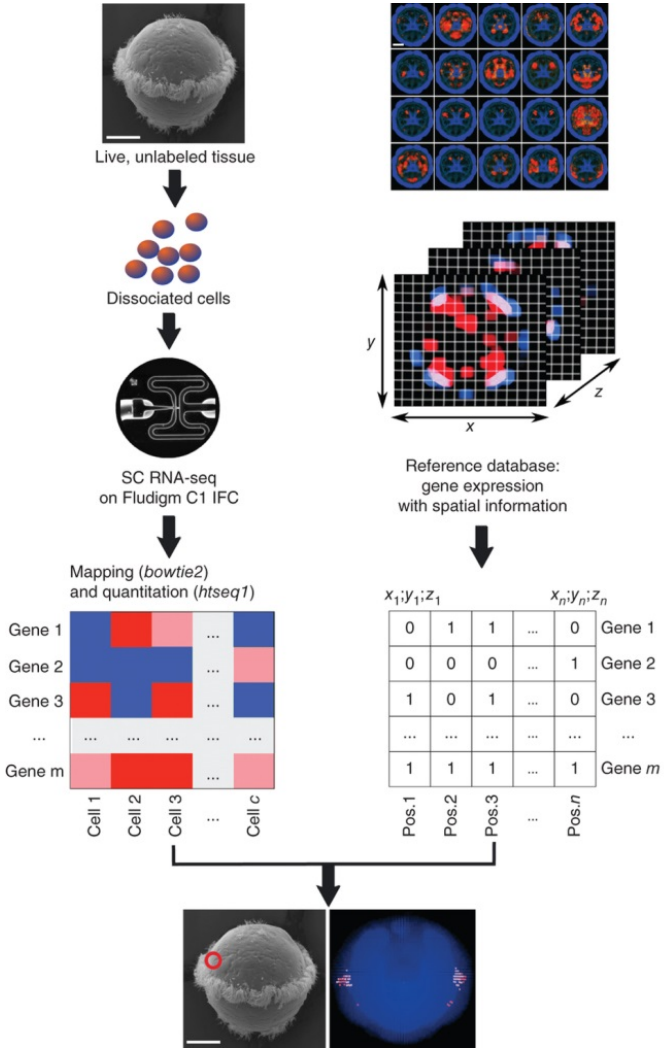


Table 1 List of existing ISH atlases

Species	Tissue	Database	Resolution (ISH)	Number of genes (ISH)
Mouse	Brain	http://mouse.brain-map.org/	Region (0.008 mm^3)	~20,000[a]
	Prenatal brain	http://developingmouse.brain-map.org/	Region (fine)	~2,000
	Developing embryo (E14.5)	http://www.genepaint.org/; http://www.eurexpress.org/	Region (fine)	16,193
Chicken	Developing embryo, various stages	http://geisha.arizona.edu/	Region	4,072
Xenopus laevis	Whole animal, various stages	http://www.xenbase.org/	Region (broad)	360[b]
Drosophila melanogaster	Whole animal, various stages	http://insitu.fruitfly.org/	Region (broad)	7,808
		http://bdtnp.lbl.gov/	Cell	95
Caenorhabditis elegans	Whole animal, various stages	http://www.wormbase.org/	Cell, cell group	3,363
Arabidopsis thaliana	Root	http://www.arexdb.org/	Cell	20,872[c]
Non-model species				
Human	Brain	http://human.brain-map.org/	Selected regions	~1,000
Zebra finch	Brain	http://www.zebrafinchatlas.org/	Region (fine)	187
C. intestinalis	Whole animal, various stages	http://www.aniseed.cnrs.fr/aniseed/	Region (fine, broad)	up to 2,600[d]
Marine invertebrates, 21 species	Whole animal, various stages	http://www.kahikai.org/index.php?content=genes	Region (broad)	306
P. dumerilii	Developing brain	Tomer *et al.*, 2010 (ref. 14); Pettit *et al.*, 2014 (ref. 17)	Subcellular[e]	168

- Gene expression atlas → binarized → matrix of ***n*** positions that each comprise presence and absence values (1 or 0, respectively) for ***m*** genes.
- For each sequenced cell ***c***, expression data for same set of ***m*** genes is compared to expression profiles at all ***n*** positions in the reference matrix and matched based on highest similarity.

- Table: ISH atlases exist for many species and developmental stages → broadly applicable.
- Can also use targeted ISH of marker gene screens as a mapping reference for RNA-seq data.

Application

- Gene expression patterns in developing brain of marine annelid, *P. dumerilii*.
- *P. dumerilii* is an important model system for studying bilaterian brain evolution.

- At 48h post-fertilization (hpf), the *P. dumerilii* larval brain is composed of a relatively low number of cells (~2,000)
  - Wide range of cell types (several types of differentiated neurons, sensory cells and proliferating progenitor cells).
  - Previously, whole-mount *in situ* hybridization (WMISH) was used to study the expression pattern of 169 differentially expressed candidate genes such as transcription factors, regulators of cell fate and body plan patterning, within the brain of *P. dumerilii*,

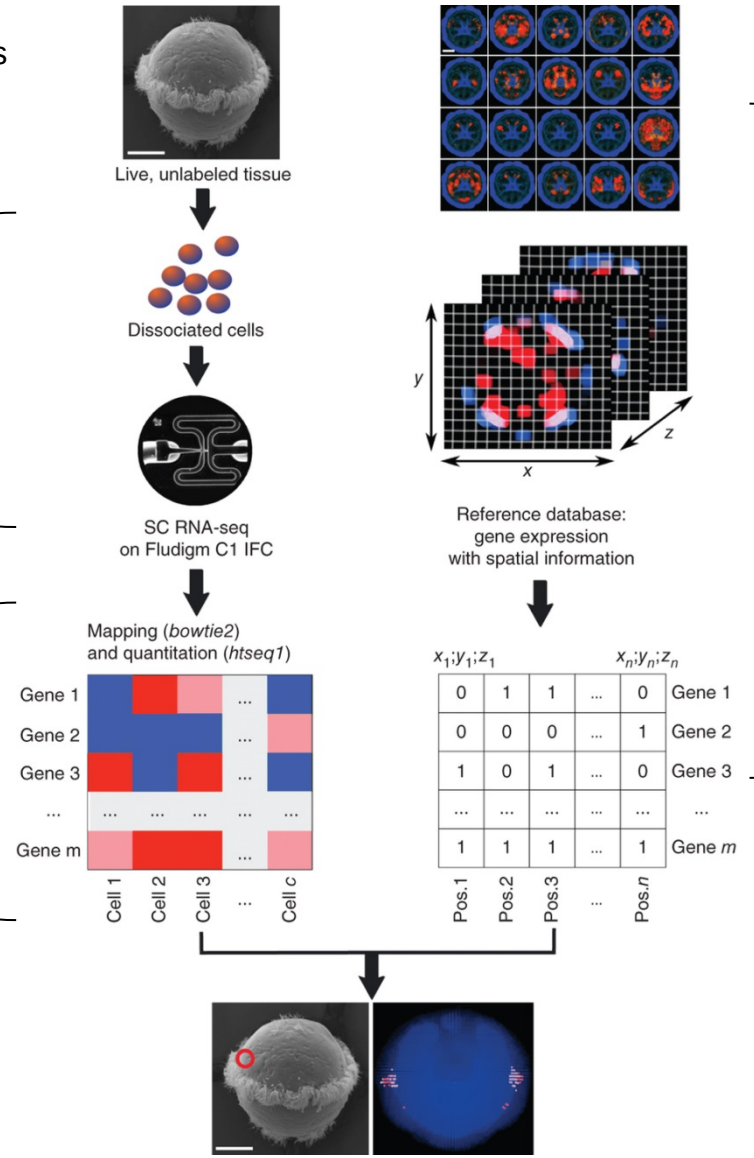
  → Facilitating the creation of a WMISH expression atlas

Work flow

P. dumerilii larval brains (at 48hpf)

Dissociation, cell capture, lysis, RT, cDNA amplification, next generation sequencing

Mapped 139 high-quality filtered cells' sequencing reads to *P. dumerilii* reference transcriptome (using *bowtie2*)



Spatial mapping:
- 169 genes included in reference ISH databases
- Removed genes with low-quality ISH signals
→ **reduced reference set of 72 genes**

ISH data set divided into 3um^3 voxels & binarized → matrix

Computational model:

1. Calculated a specificity score → convert score vector such that elements take values between 0 and 1→ **transformed specificity score**

2. Determined a correspondence score for each cell-voxel combinations

3. Determine significance of the cell-voxel correspondence scores using simulations → **determined empirical probability**

Mapping individual cells to precise, single location

Mapping results:

- Established likely location for 91% of cells in data set.
- Could map back majority (83%) of sequenced cells to a precise, single location
- Set of voxels to which each cell is mapped back are typically arranged in small, bilaterally symmetric and spatially coherent groups (Fig. a-b")

Broad mapping domains (ex. Fig. c):

- Indicative of relative molecular homogeneity of respective brain regions
- Augmenting reference atlas with genes that display variable patterns of expression should improve precision of mapping.

Effect of size of reference atlas on mapping:

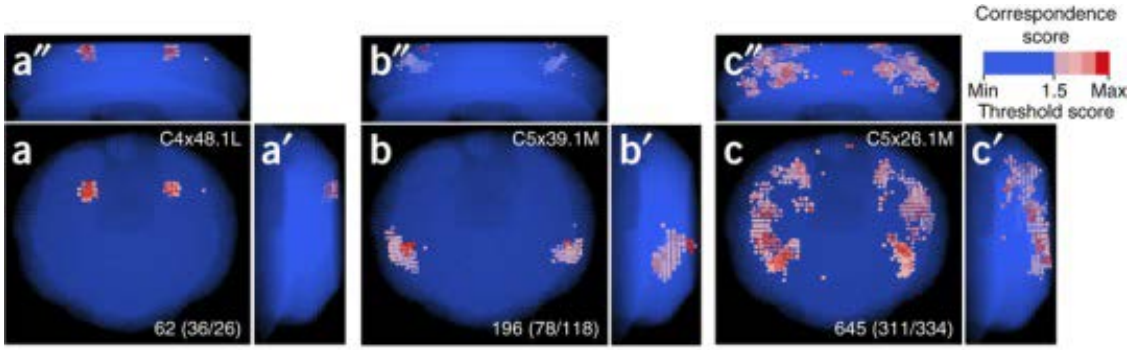
- Fraction of cells mapped back with medium & high confidence increased as a function of the number of reference genes

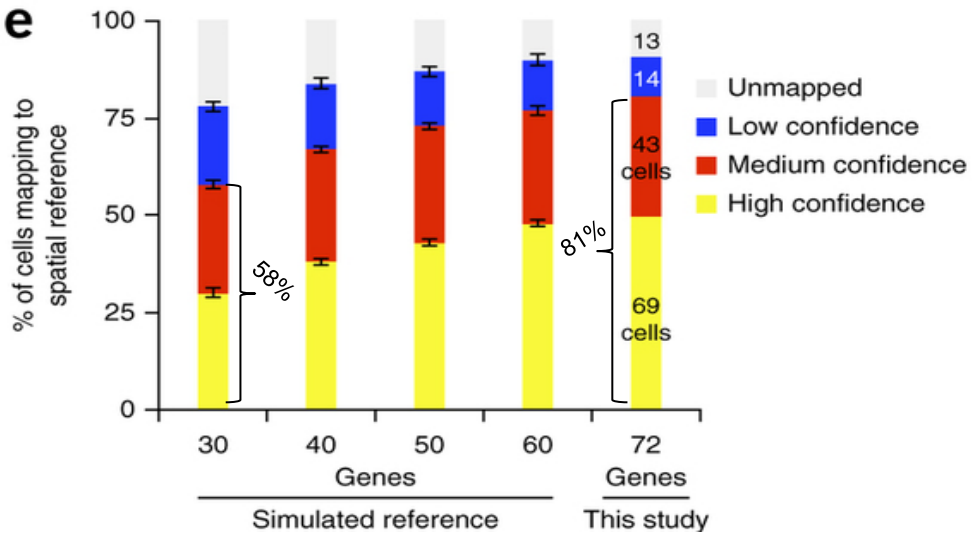
→ Only ~50-100 genes with spatially distinct patterns of expression needed to map cells to specific location with high degree of confidence

Single-cell precision mapping
(<150 voxels)
= 83% of seq. cells

Cell mapping to small brain region
(150-500 voxels)
= 13% of seq. cells

Broad mapping
(> 500 voxels)
= 4% of seq. cells

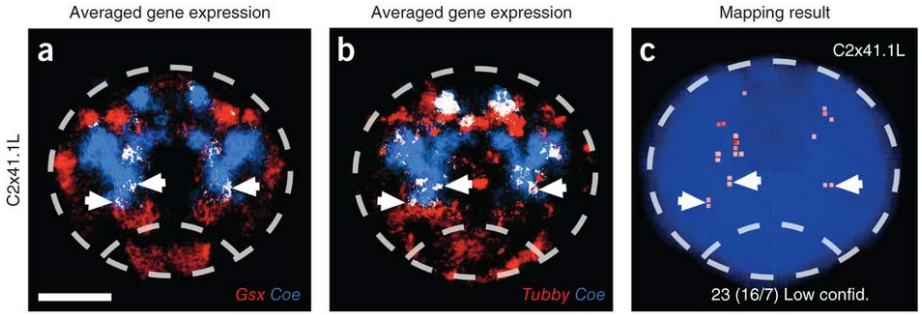




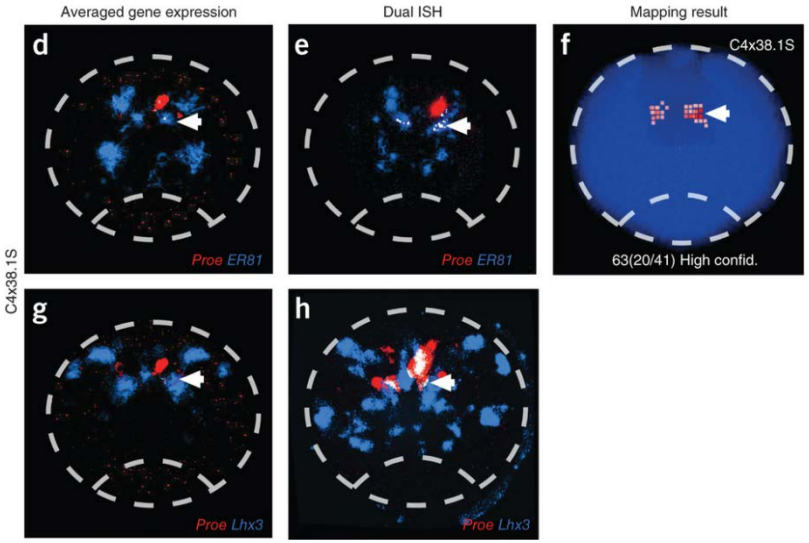
Mapping validation & associated challenges

Co-expression analysis for **genes that were co-expressed in scRNA-seq data but not represented in the binarized ISH dataset:**

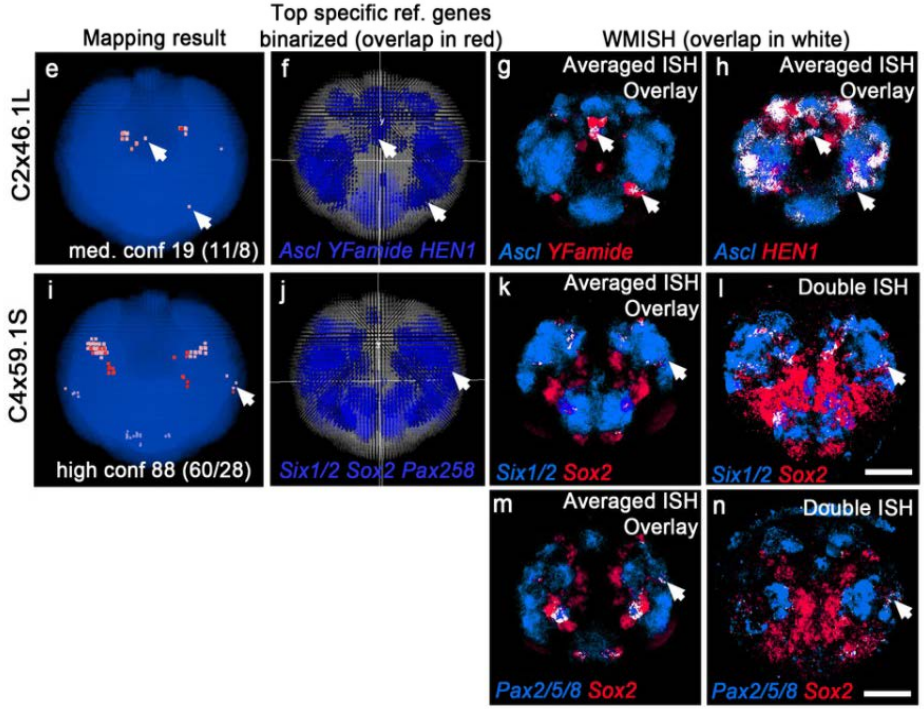
Ex.: Overlaying averaged, non-binarized ISH images revealed areas of co-expression:



Ex.: Mismatch not explained by overlaying → Dual ISH confirms colocalization:



*Ex.: Overlaying averaged, non-binarized ISH images revealed **larger overlap** of genes:*



1. Reference matrix used averaged expression patterns
   → **Averaging & binarization** of ISH images can lead to loss of information → **false 'presence'** & **'absence'** calls in binarized reference spatial matrix
   → Altering binarization threshold can overcome this problem and improve the reference.

2. Imperfections in ISH database
   → Misannotation of gene expression value

3. Technical noise in scRNA-seq → bias for particular genes (ex.: → erroneously high specificity score)

Conclusion: Approach robust to technical challenges

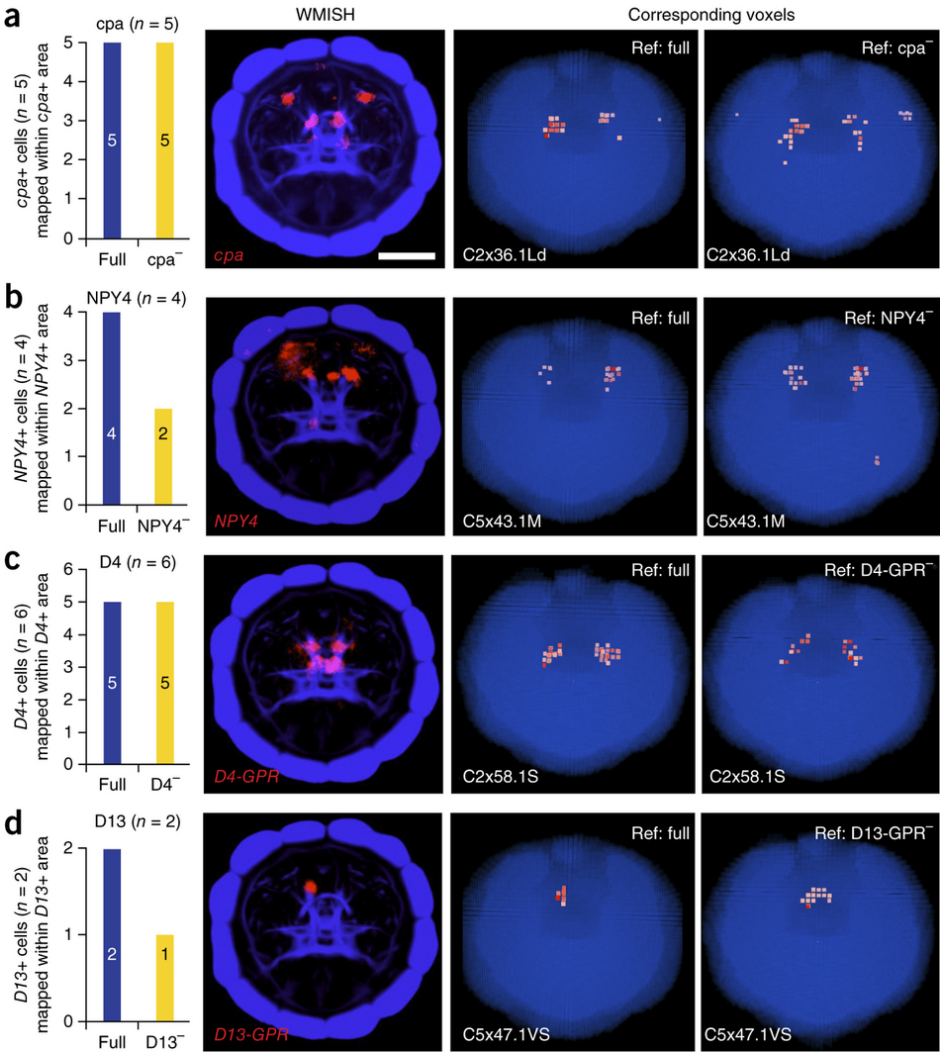
Validation using reference-independent marker genes

- Removed gene from reference matrix and compared the mapping results with those generated with the full reference.
- Mapping successful = statistically significant overlap between voxels to which it was mapped back and the expression domain of the selected marker gene.

Results:

- 14/17 cells displayed concordant results with both references
- 1/17: marginally sig. when full reference used, statistically insig. overlap when using reduced reference
- 2/17: no mapping back to loci when using reduced reference (retrospectively found respective gene weakly expressed)

→ Approach provides a tool for identifying genes co-expressed with known markers, thus revealing new biological insights.

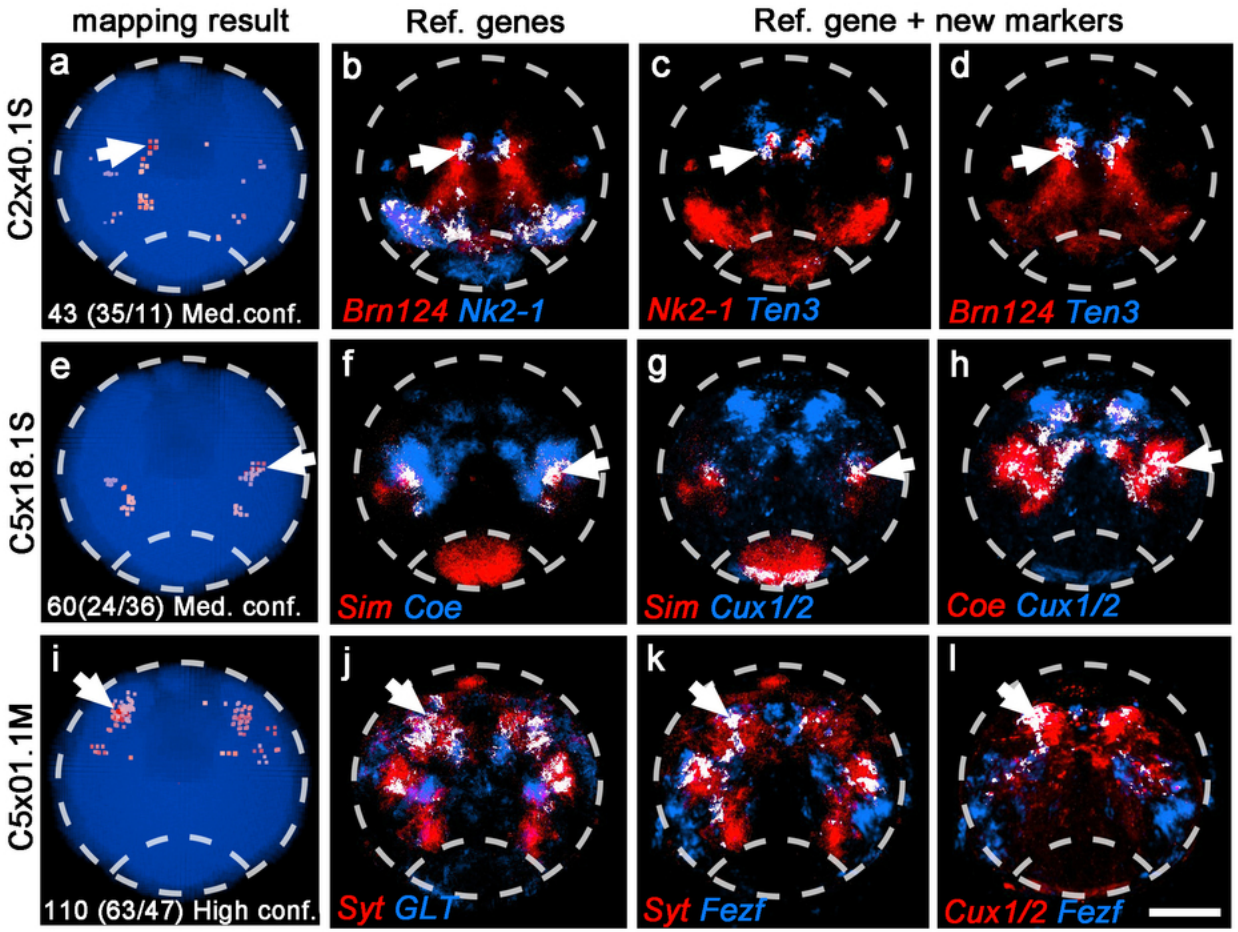


Identifying genes co-expressed with known markers

Registered new WMISH patterns for 3 genes (*Ten3, Cux1/2, Fezf*) expressed in subset of cells in scRNA-seq dataset, then assess spatial mapping

→ *Ten3, Cux1/2, Fezf* each co-expressed with known reference genes in the location indicated by spatial mapping

→ New marker genes identified from the scRNA-seq experiment **independently validated the spatial mapping** and could be used to **further refine the reference atlas**



Summary

- Developed a computational approach that combines a spatially referenced ISH atlas with single-cell transcriptome profiles generated using scRNA-seq to map each cell back to the tissue under study.
- Profiling ~7% cells in P. dumerilii brain (randomly distributed throughout tissue of interest), 81% of cells were mapped back to a relatively precise location.
- Validated results computationally & using ISH
- Does not require a priori cell labeling (unlike TIVA)

+ Can assay cells from across relatively large tissue simultaneously (unlike FISSEQ)
+ High throughput in contrast to TIVA & FISSEQ
+ Broadly applicable in contrast to FISSEQ
+ Can be used to identify new tissue-specific genes

- Spatial origin of cells assayed by TIVA & FISSEQ can be determined unambiguously
- Approach depends critically on the quality (resolution, accuracy & information content) of the reference atlas and scRNA-seq data

+ Even without a cellular resolution reference ISH atlas (majority of cases), cells can be mapped back to small and restricted spatial domains using this approach

Conclusion & Outlook

- Two computational models to accurately map location of individual cells in complex tissue
- Identify expression patterns
- Localize rare cell population and identify novel markers
- Extract valuable information from existing data sets & databases (scRNA-seq; ISH)

Outlook

- Produce a high-quality spatial map of a tissue:
  - Standard techniques could suggest the most relevant landmark genes to establish a preliminary input spatial map
  - Generate an unbiased spatial reference map with emerging techniques that perform low-input RNA-seq on cryosections (eg RNA tomography).

Thank you for your attention