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

Kaia Achim[1,2,5], Jean-Baptiste Pettit[1,4,5], Luis R Saraiva[1,3], Daria Gavriouchkina[2,4], Tomas Larsson[2], Detlev Arendt[2] & John C Marioni[1–3]

Understanding cell type identity in a multicellular organism requires the integration of gene expression profiles from individual cells with their spatial location in a particular tissue. Current technologies allow whole-transcriptome sequencing of spatially identified cells but lack the throughput needed to characterize complex tissues. Here we present a high-throughput method to identify the spatial origin of cells assayed by single-cell RNA-sequencing within a tissue of interest. Our approach is based on comparing complete, specificity-weighted mRNA profiles of a cell with positional gene expression profiles derived from a gene expression atlas. We show that this method allocates cells to precise locations in the brain of the marine annelid *Platynereis dumerilii* with a success rate of 81%. Our method is applicable to any system that has a reference gene expression database of sufficiently high resolution.

High-throughput single-cell RNA-sequencing by automated reaction processing in multiwell plates[1] or by microfluidics devices[2] is becoming an established experimental technique[3]. The sample collection step in these protocols results in the dissociation of tissue and therefore loss of spatial information. However, investigating the molecular composition of individual cells in the context of spatial location is important, especially when studying primary cells. This is particularly relevant when looking at complex tissues, such as the early embryo or the brain, where multiple heterogeneous cell types are located in close proximity.

Single-molecule fluorescence *in situ* hybridization (FISH)[4] has been widely used to quantitate transcript numbers at single-cell resolution within the context of a tissue of interest. This allows gene expression to be assayed in many cells but it can only be applied to a small number of genes. More recently, methods for high-throughput, spatially resolved single-cell RNA-seq have been developed using *in vivo* marking

and picking of selected cells from predefined spatial coordinates[4], or *in situ* amplification of cellular transcriptomes on tissue sections[5]. Although these approaches facilitate analysis of all expressed genes in each captured cell, they are limited in terms of the number of cells that can be labeled and then processed (**Supplementary Note 1**). An alternative approach is to carry out unbiased single-cell RNA-sequencing (where cells are dissociated without knowledge of their spatial location) followed by computational approaches, such as Principal Component Analysis, to partially recover the spatial structure of the tissue of interest[6]. However, such approaches do not facilitate a fine-grained reconstruction whereby the location of each cell can be precisely determined.

To overcome these limitations, we propose an integrated approach that combines previously generated *in situ* hybridization (ISH)-based gene expression atlases with unbiased single-cell transcriptomics (**Fig. 1**). Notably, ISH atlases exist for many species and developmental stages (**Table 1**) making our approach broadly applicable. Moreover, beyond the model systems for which comprehensive maps are typically available, targeted ISH of tens of marker genes is commonly used to study spatially restricted patterns of gene expression in systems that provide key insights into evolution and development[7–9]. Such screens could also be used as a mapping reference for RNA-seq data, and, for several 'nonmodel' species (e.g., zebra finch or the ascidian *Ciona intestinalis*), high-quality gene expression data are already being assembled into structured databases (**Table 1**).

RESULTS

Experimental design and quality control

We used the developing brain of a marine annelid, *P. dumerilii,* to demonstrate the utility of our approach. *P. dumerilii* is an important model system for studying bilaterian brain evolution[10,11] so obtaining spatially registered transcriptomic profiles of cells within its brain is valuable for understanding how cell types in the vertebrate brain evolved.

At 48 h post-fertilization (hpf), the *P. dumerilii* larval brain is composed of a relatively low number of cells (~2,000). However, this small number of cells harbors a wide range of cell types, including several types of differentiated neurons, sensory cells and proliferating progenitor cells[12–16]. 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*, thus facilitating the creation of a WMISH expression atlas[14,17]. Subsequently, this WMISH data set was divided into 3-μm[3]

Figure 1 Overview of the spatial mapping workflow using the *P. dumerilii* scRNA-seq data set. We used an existing gene expression atlas (right column) to link scRNA-seq data (left column) from cells extracted from the developing brain of *P. dumerilii* (top left) with spatial coordinates. The gene expression atlas was binarized, resulting in a matrix of *n* positions that each comprise presence and absence values for *m* genes. For each sequenced cell *c*, expression data for the same set of *m* genes was compared to expression profiles at all *n* positions in the reference matrix and matched based on highest similarity. An example of the likely position for one cell is indicated in the two images at the bottom by the red circle in the bottom left (ventral view of *P. dumerilii* larva) and red voxels in the apical view at bottom right. The upper right panel of this figure is modified from ref. 17. Scale bars, 50 µm.

voxels and binarized, resulting in a matrix (**Fig. 1**) where genes are arranged in the rows and spatial coordinates (voxels) in the columns[17]. Using a previously described approach[17], we removed genes with low-quality WMISH signals, resulting in a reduced reference set of 98 genes. Considering the 3-µm^3 voxel size and the observation that cell diameters range from 4 to 25 µm, most cells in the *P. dumerilii* brain are expected to correspond to 11–174 voxels. An entry in the matrix (one voxel) is set equal to 1 if the corresponding gene is expressed in the voxel of interest and set equal to 0 otherwise.

To generate the single-cell RNA-sequencing (scRNA-seq) data, we dissociated *P. dumerilii* larval brains, and followed that by cell capture, cDNA synthesis and amplification on the C1 Single-Cell Auto Prep IFC. In total, we sequenced 213 samples, of which 155 (73%) corresponded to single cells (as judged by visual inspection of the captured cells) with the remainder consisting of single dead cells, wells containing multiple cells and empty wells, which were sequenced as a negative control (**Supplementary Table 1**).

For each cell, the scRNA-seq data were mapped to the *P. dumerilii* reference transcriptome augmented with the ERCC spike-in molecule sequences[18]. The quality of the data was evaluated by calculating the percentage of all reads mapping to the reference transcriptome and to the ERCC spike-in sequences (**Supplementary Fig. 1**). Samples where >10% of reads mapped to the ERCC sequences and <10% of reads mapped to the reference transcriptome were excluded from further analyses. In addition, samples containing no cells or multiple cells were excluded, leaving 139 high-quality filtered cells (QF cells) that were used in downstream analyses (**Supplementary Table 1**). We note that 139 cells corresponds to ~7% of all cells in the larval brain of *P. dumerilii*.

To assess whether the 139 QF cells represented a random sample of cells from across the whole brain, we compared the proportion of sequenced cells in which each gene was expressed with the proportion of voxels showing expression of the same gene in the binarized reference atlas (**Supplementary Fig. 2**). This revealed that the ratios display good concordance (**Supplementary Fig. 2**; Spearman's Rho = 0.48, $P < 2 \times 10^{-5}$ Spearman's Rank Test), thus providing confidence that the sequenced cells represent a broad sample of cells from across the *P. dumerilii* brain.

Mapping of individual cells

To determine the spatial origin of each sequenced cell we applied a three-step approach. First, for each cell, we calculated a specificity score that indicates to what extent each gene is specifically expressed in that cell relative to all other cells (**Supplementary Note 2**). For each cell, its score vector was then transformed using a logistic function such that its elements took values between 0 and 1, where a value near 1 indicates that a gene is highly specific to the cell of interest.
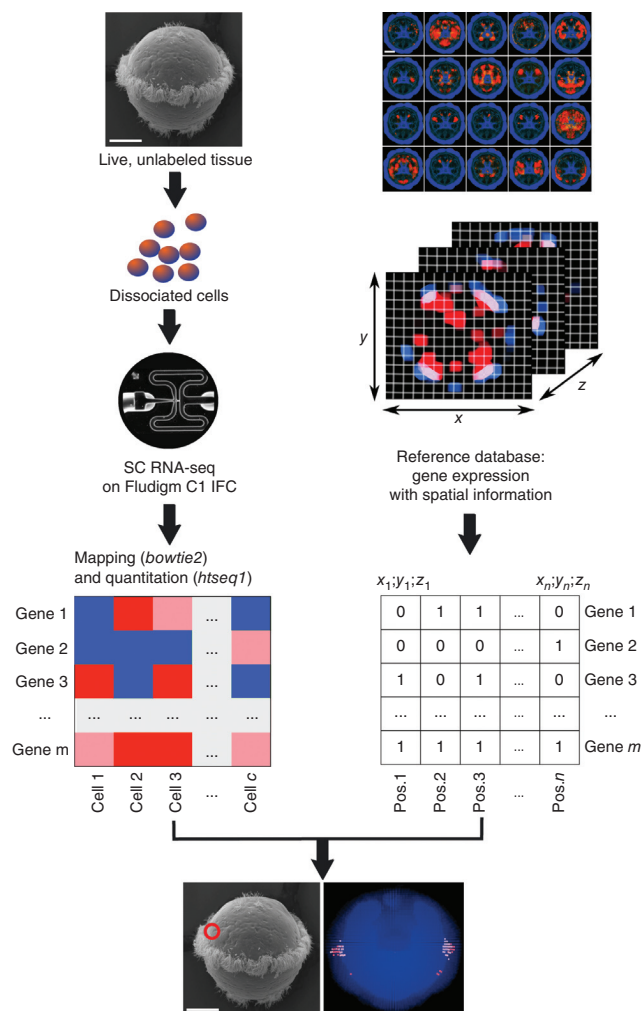
Subsequently, a 'correspondence score' was determined for each cell-voxel combinations. For each gene expressed in the scRNA-seq data and present in the WMISH atlas (72/98 genes; others correspond to genes expressed at low levels or expressed exclusively in cells that we did not capture), a match or mismatch between the scRNA-seq and the WMISH data resulted in the transformed specificity weight being added or subtracted from the cell-voxel score, respectively. This yielded a cell-voxel correspondence score for all combinations of cells and voxels. Third, we used simulations to determine, for each cell, the significance of the cell-voxel correspondence scores (**Supplementary Figs. 3** and **4**). More specifically, we generated a randomized data set by permuting the specificity scores 100 times for each cell and mapped these simulated cells back to the reference. Based upon this simulated data set, we determined the empirical probability that each true cell was mapped back to a particular number of voxels at a given correspondence score threshold.

We mapped 69 cells (50%) with high confidence (at least 21 voxels with a score >1.5 corresponding to <10% chance of occurrence in the simulated data), 43 cells (31%) with medium confidence (at least 16 voxels with a score > 0.5 corresponding to <30% chance of occurrence in the simulated data) and 14 cells (10%) with low confidence (>11 voxels with a score > 0 corresponding to <50% chance of being observed in the simulated data; **Supplementary Figs. 3** and **4**). We considered the voxels with the highest confidence for each sample as being the most likely loci from which the cell assayed by means

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

ISH atlases for an array of model and non-model species are listed, alongside information about the number of genes analyzed, the resolution of possible co-expression analyses in each system, tissues and developmental stages for which the atlas exists, and web links to the databases. In most systems, cellular resolution is impossible to achieve owing to the variability in the cell positions between individuals. In these cases, we labeled region-specific resolution as "fine" when expression analysis is performed on tissue sections and imaged at high resolution, and "broad" when only whole-mount images are available.

[a]Co-expression analyses at 0.008 mm^3 resolution are available for 4,104 genes with brain-wide expression[23]; [b]manually curated set of whole-mount *in situ* hybridization images are available for 360 genes in *Xenopus*; [c]including a microarray data set of different cell populations in *A. thaliana* root; [d]number of genes varies for different stages, usually ~200-2,000 genes are shown; [e]three-dimensional stacks of 169 ISH images for *P. dumerilii* larval brains allow co-expression analyses at resolution of 3 µm^3.

of scRNA-seq originated (**Supplementary Table 2**). Altogether, we established a likely location for 91% of cells in our data set (**Fig. 2d**), with the median number of voxels per mapped cell being 54 (interquartile range: 30 – 110).

Notably, the set of voxels to which each cell is mapped back to are typically arranged in small, bilaterally symmetric and spatially coherent groups (**Fig. 2a-a″, 2b-b″** and **Supplementary Table 2**). This is expected, because the voxel size in our model is much smaller than one cell, and the *P. dumerilii* brain shows strong bilateral symmetry at 48 hpf. Given that the captured cells are between 8 and 17 µm in diameter (**Supplementary Note 2**) and the voxels are all 3 µm^3, our results suggest that we can map back the majority (83%; number

of corresponding voxels <150, **Fig. 2d**) of sequenced cells to a precise, single location (considering bilateral symmetry; **Fig. 2a-a″**). The remaining cells map back to either a small number of voxels (13%; number of corresponding voxels between 150 and 500; **Fig. 2b-b″**) or, in three cases, to a broader domain (4%; number of corresponding voxels >500; **Fig. 2c-c″**). Broad mapping domains are indicative of relative molecular homogeneity of the respective brain regions, at least when considering the genes included in the reference. In this case, augmenting the reference atlas with genes that display variable patterns of expression in the cells mapped back to such regions should improve the precision of the mapping. A summary of the locations across the brain to which all cells are mapped back can be found in

Figure 2 Examples of the spatial mapping of single cells. (**a**) An example of a cell mapped at single-cell precision (<150 voxels). 83% of cells in our data set fit this mapping profile. (**b**) An example of a cell mapping to a small brain region. 13% of cells in our data set fit this mapping profile. (**c**) An example of a broad mapping where many voxels score above the threshold. 4% of cells in our data set fit this mapping profile. Note that for mapping profiles shown in **b** and **c**, smaller high-score voxel groups (dark red) can be identified within the large number of above-threshold voxels. Such high-score domains likely correspond to the precise positions of the cell. The samples shown in **a**–**c** are mapped with high confidence. (**a**–**c**) apical view of the brain; (**a′**–**c′**) lateral view; (**a″**–**c″**) dorsal view. Voxels with a correspondence score < 1.5 are colored blue. Upper right corner: cell ID; lower right: number of voxels above threshold (in parentheses: voxels localized in the left/right-side of the brain, respectively).
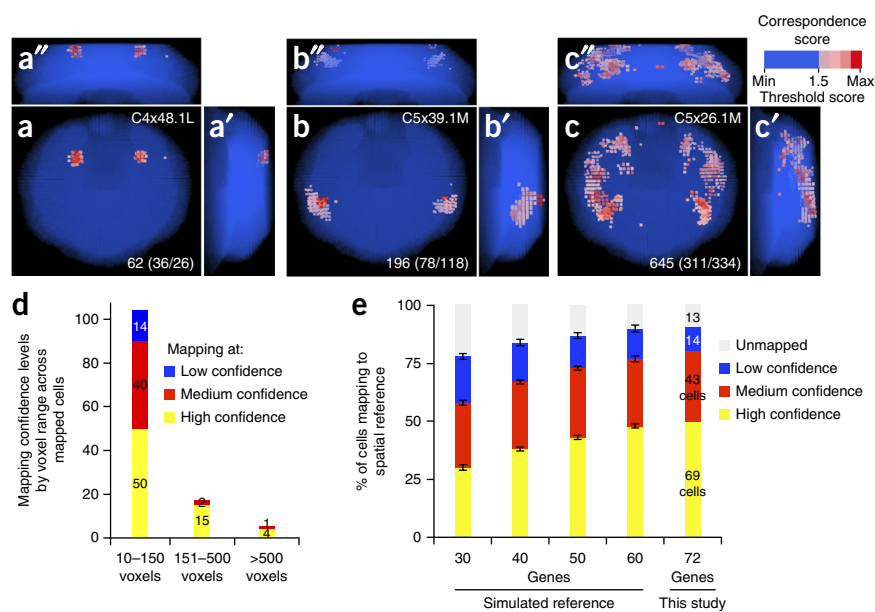
(**d**) Distribution of mapped cells across the defined confidence levels stratified by the number of voxels to which each cell is mapped back. (**e**) Percentage of cells mapping back to each category (high/medium/low/unmapped) for the reference used in this study (72 genes out of the 98 genes in the full *P. dumerilii* spatial reference atlas that were expressed in at least one of the sequenced cells; right-hand side) and the mean (± s.e.m.) number of cells assigned to each category when simulated references containing different numbers of genes (30, 40, 50 and 60) are used.

Figure 3 Validation of co-expressed genes in scRNA-seq data. Co-expression analysis for combinations of cell-specific genes that were not represented in the binarized ISH data set. (**a**,**b**) Digital overlay of averaged ISH images of 48 hpf *P. dumerilii* larvae (image registration and average pattern calculations were performed on 3–6 animals per gene[14]) showing the co-expression of *Gsx* and *Tubby* with *Coe* (**a**,**b**). (**c**) Visualization of the mapping result for the respective cell (C2x41.1L). Labeling scheme as in **Figure 2**. In addition, the mapping confidence level is indicated for each sample. (**d**–**h**) Comparison of averaged ISH overlay images and dual ISH on individual *P. dumerilii* larvae. Dual ISH on 48 hpf *P. dumerilii* larvae showing cells co-expressing *Proenkephalin* and *ER81* (**e**) and *Proenkephalin* and *Lhx3* (**h**), undetected in both the averaged ISH images (**d**,**g**, respectively) and the binarized reference matrix. Colocalization of genes is shown in white. (**f**) The position of the cell (C4x38.1S) indicated by spatial mapping corresponds to the position of the cells co-expressing the genes analyzed by ISH. The outline of *P. dumerilii* brain is drawn with dashed line. White arrows point to the location of co-expressing cells. Scale bar, 50 μm.
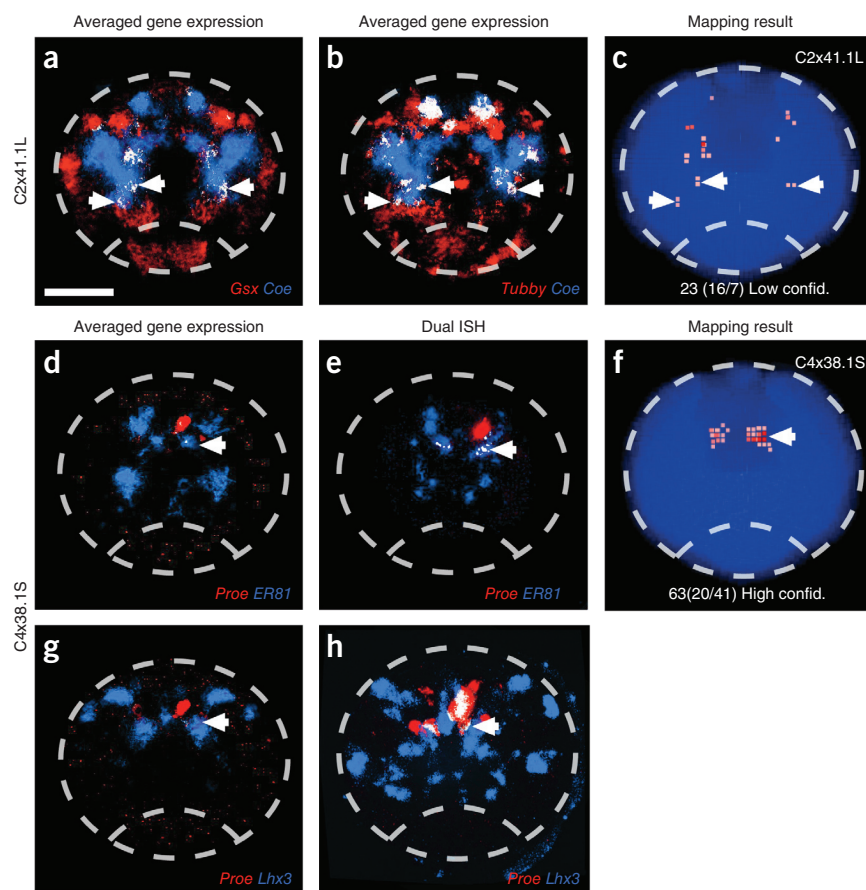
Supplementary Figure 5. The individual cell mappings can be viewed at the following link: http://www.ebi.ac.uk/~jbpettit/map_viewer/?dataset=examples/coord_full.csv&cluster0=examples/resultsBio.csv.

The size of the reference atlas

To investigate how the number of reference genes affected the mapping, we generated simulated spatial reference atlases, each containing a subset of the genes used in the initial atlas (30, 40, 50 and 60 gene subsets were used), and mapped each cell back to these simulated references. We observed that the fraction of cells mapped back with medium or high confidence increased as a function of the number of reference genes. On average, 58% of cells were mapped back with medium or high confidence when only 30 reference genes were used, in contrast to the 81% of cells that were mapped back when the full reference set of 72 genes (corresponding to genes included in the WMISH atlas and expressed in the scRNA-seq data) was deployed (**Fig. 2e**). This suggests that a relatively small number of genes (between 50 and 100) with spatially distinct patterns of expression are needed to map cells to a specific location with a high degree of confidence (Discussion). Critically, pre-existing ISH resources often contain more genes (**Table 1**) than the *P. dumerilii* reference used in this study.

Mapping validation and the associated challenges

Overall, our approach enables the majority of cells to be mapped back with medium to high confidence. However, even for these cases, there exist discrepancies where, for example, a highly specific gene is not expressed in the region to which a cell is mapped (**Fig. 3**). This can occur as our approach uses a relatively large reference set and is therefore robust to such mismatches assuming that other, highly specific, genes exist. However, understanding the factors that lead to such discrepancies between the ISH and scRNA-seq data sets provides valuable insights into the quality of both data sets.

First, our WMISH database used averaged expression patterns, which can lead to both false 'presence' and 'absence' calls in the reference matrix, especially for genes with variable expression across individuals. Indeed, in 13 cells where one of the four most specific genes was never co-expressed in the same domain as the remaining three genes, we observed that overlaying the averaged, nonbinarized ISH images revealed areas of co-expression (**Fig. 3a–c** and **Supplementary Fig. 6**). Altering the binarization threshold can overcome this problem and improve the reference.

Second, imperfections in our WMISH database, such as missing expression domains due to probe design, low gene expression levels in particular cells or insufficient signal development, can also lead to misannotation of a gene expression value. Illustrating this, we chose four additional cells where one of the top four most-specific genes was not co-expressed with any of the other three most-specific genes and where the mismatch was not explained by overlaying the nonbinarized WMISH patterns. In all four cases we performed new dual ISHs between the mismatch gene and at least one of the remaining highly specific genes. The domains of dual ISH co-labeling overlapped the locations indicated by the spatial mapping in all cases (**Fig. 3** and data not shown), affirming the initial mapping.

Finally, scRNA sequencing is susceptible to substantially more technical noise than bulk sequencing approaches[19]. It is possible that such biases might lead to a disproportionately large number of reads being associated with a particular gene in a given cell and, consequently, a high specificity score. As the quality of scRNA-seq data improves, this will become less of a problem.

Figure 4 Comparison of mapping success using highly specific marker genes. (**a–d**) Comparison of the mapping success before and after a highly specific marker gene is removed from the reference. For each gene, we considered cells showing specific expression of that gene in the scRNA-seq data (specificity score > 4). From these cells, we assessed the number of successfully mapped cells against the full reference atlas (blue bar) and after removing the respective gene from the reference (yellow bar). The expression pattern of the gene is shown on the right (WMISH). Scale bar, 50 μm. One example of the mapping result (corresponding voxels in red) against the full reference (ref: full) and the reference lacking the respective gene (ref: removed, gene indicated in the top right corner) is shown for each case.
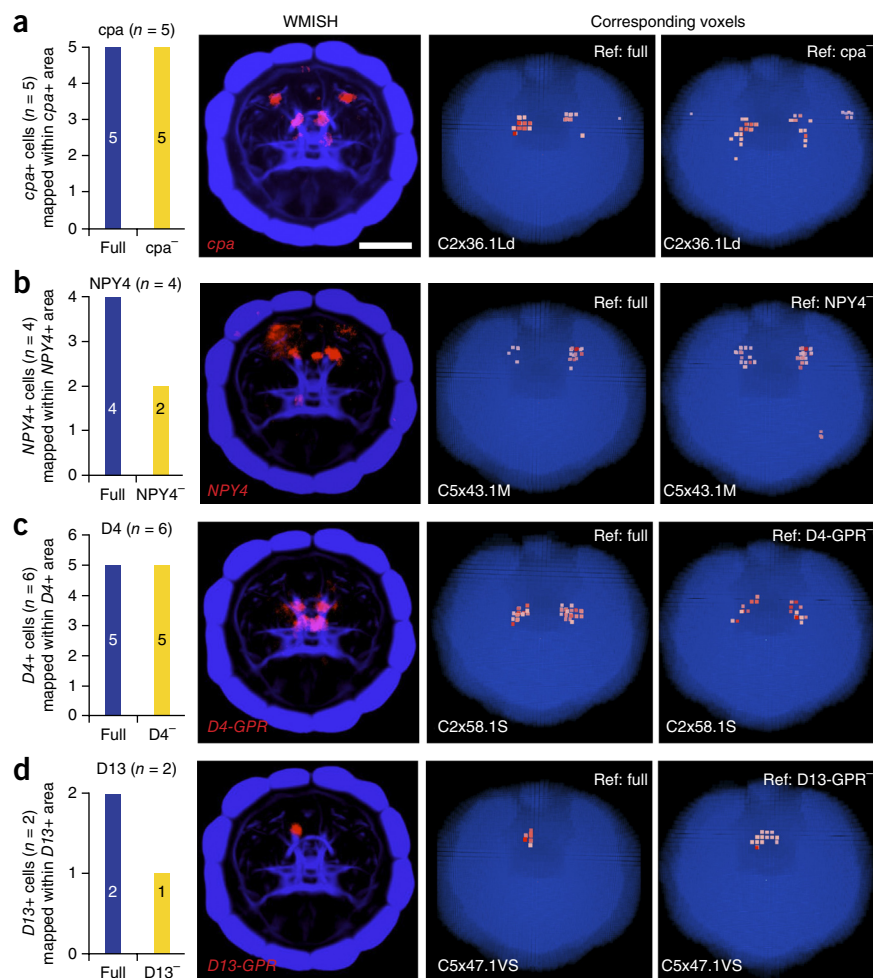
Nevertheless, we identified the probable loci of origin for 81% of cells with high or medium confidence, suggesting that our approach is robust to technical challenges.

Validation using reference-independent marker genes

To validate the mapping, we selected four genes expressed in a small number of voxels (each gene expressed in <2% of voxels) that were expressed specifically in few cells (specificity score > 4 in each cell). We removed each gene, in turn, from the reference and compared the mapping results obtained using the reduced reference with those generated with the full atlas (**Fig. 4**).

For each cell in which the gene of interest was specifically expressed, we considered mapping successful when there was statistically significant overlap (Fisher Exact Test; P-values were corrected for multiple testing using the Benjamini-Hochberg approach; significance threshold = 0.05) between the voxels to which it was mapped back and the expression domain of the selected marker gene (**Supplementary Table 3**). Of the 17 cells tested, 14 displayed concordant results with both references (13/14 of these cells were mapped back to domains that showed statistically significant overlap with the relevant marker gene; for 1 of the 14 cells, no significant overlap was observed when either reference was used). For the remaining three cells, one was marginally significant when the full reference was used but showed a statistically insignificant overlap when the reduced reference was employed. The final two cells were not mapped back to voxels expressing the gene of interest only when the reduced reference was used. However, in both cases, closer inspection of the reference atlas revealed weak expression of the respective gene in the areas matching the predicted position of these unsuccessfully mapped cells (data not shown).

Our approach provides a tool for identifying genes co-expressed with known markers, thus revealing new biological insights. To demonstrate this, we registered new WMISH patterns for three genes that were expressed in a subset of cells in the scRNA-seq data set, which suggests they each have a spatially restricted expression pattern. We then assessed whether the expression domain of each marker conforms to the spatial mapping. We found that *Ten3* (**Supplementary Fig. 7c,d**), *Cux1/2* (**Supplementary Fig. 7g,h,l**) and *Fezf* (**Supplementary Fig. 7k,l**) were each co-expressed with known reference genes in the locations indicated by spatial mapping (**Supplementary Fig. 7a,e,i**, white arrows). In all cases, combination with the new marker defined the location even more precisely than the known genes. Thus, new marker genes identified from the scRNA-seq experiment independently validated the spatial mapping and could be used to further refine the reference atlas.

These validations strongly support the quality of our mapping approach and demonstrate how it can be used to exploit a relatively small, spatially referenced *in situ* atlas alongside scRNA-seq to identify new marker genes that can form the basis of downstream studies.

DISCUSSION

In this manuscript, we 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. We demonstrated the utility of our approach using cells taken from the brain of the marine annelid, *P. dumerilii*, at 48 hpf. Profiling over 7% of cells in its brain, we observed that 81% of cells were mapped back to a relatively precise location. We validated our results both computationally and using ISH for genes that displayed variability in expression in the scRNA-seq data.

Recently, alternative approaches for profiling the transcriptomes of spatially referenced cells have been proposed[5,20]. Transcriptome *in vivo* analysis (TIVA)[20] allows individual cells to be fluorescently labeled within a tissue, visualized and then sequenced after capture. In practice, this is extremely useful when a particular cell, or small

number of cells, is of interest. However, the throughput of this approach is potentially limited by the manual photoactivation and picking steps and also the number of tags that can be employed simultaneously.

By contrast, fluorescent *in situ* RNA-seq (FISSEQ)[5] sequences individual cells *in situ* (i.e., directly on cells or tissue mounted on a coverslip). Such an approach can, in theory, be broadly applied, although the practical challenges are substantial. Moreover, only cells that are relatively close to one another can be assayed in the same experiment.

Our approach is complementary to these experimental strategies. Fundamentally, we do not aim to maintain the tissue structure through the experimental process, but aim to recover this structure *post hoc* by combining the scRNA-seq data with a pre-existing spatially referenced ISH atlas. Unlike TIVA, we do not need to label cells a priori and, unlike FISSEQ, we can assay cells from across a relatively large tissue simultaneously. However, the spatial origin of the cells assayed by TIVA and FISSEQ can be determined unambiguously, which may have advantages in certain circumstances.

From a computational perspective, our method could be extended in a number of ways. First, instead of using a binarized threshold for determining whether a gene is expressed from the ISH data, we could treat the reflection data for each gene as a continuous variable. Second, the *in situ* data used in our study have a very high resolution, with 11–174 voxels making up a typical cell. As discussed below this is somewhat atypical, with most resources having substantially lower resolution. Nevertheless, when such high-resolution information exists, one option is to first build a cellular model and to map each cell back to this new atlas. Finally, our method penalizes mismatches when a gene is expressed in the scRNA-seq data and not in the reference atlas. This is justified by noting that current scRNA-seq protocols display a substantial degree of technical noise, meaning that calling a gene as unexpressed in a given cell is challenging[19]. As scRNA-seq protocols improve, a two-way penalization criterion may become more appropriate (**Supplementary Note 3**).

Our approach depends critically upon the quality of the reference atlas. Consequently, it is important to filter out genes with low-quality ISH patterns, because these can create co-expression domains that in reality do not exist, but can lead to false-positive mappings. In practice, this can be done using automated approaches that process and filter out low-quality images[21]. Additionally, the number of genes in the reference database is of high importance. In this paper we used simulations to demonstrate that, as expected, increasing the number of genes in the reference set leads to a monotonic increase in the number of cells mapped back with high confidence. However, somewhat surprisingly, we observed that the increase in the number of confidently mapped cells increased only slightly when the reference set was increased from 60 to 72 expressed genes. One important consideration when determining whether an ISH database provides a suitable reference is the information content of the genes contained therein, with genes expressed in a spatially restricted and nonoverlapping manner providing the most utility.

The precision of mapping is also affected by the resolution of the ISH atlas. Indeed, the majority of species listed in **Table 1** lack a cellular resolution reference atlas. However, even without such an atlas, cells can be mapped back to small and restricted spatial domains using our method, thus facilitating important biological insights including the identification of new tissue-specific genes. Illustrating this, Satija *et al.*[22] have developed an analogous computational approach, which they apply to assign individual cells to locations within the gastrulating zebrafish embryo using a low-resolution ISH atlas.

As with the ISH atlas, potential problems can also arise during the scRNA-seq data generation. At present, scRNA-seq is a relatively noisy technology, with only ~20–40% of molecules being captured per cell in a typical experiment. Consequently, for genes expressed at low to moderate levels, quantification of expression is challenging. Moreover, some technologies require that cells fall within a specific size window before processing and sequencing. This might result in under-representation of particular cell types, thus seriously affecting the utility of approaches such as that described herein. The use of microwell plates and droplet technology, as well as protocol improvements, promise to overcome these limitations.

METHODS
Methods and any associated references are available in the online version of the paper.

Accession codes. Raw sequencing data are available from ArrayExpress, accession number E-MTAB-2865/.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS
K.A. developed most experimental protocols, performed experiments, helped analyze and interpret the data and wrote the manuscript. J.-B.P. developed the computational approach, analyzed the data and helped write the manuscript. L.R.S. performed single-cell RNA-sequencing experiments. D.G. initiated the single-cell work in *Platynereis*, developed the protocol for dissociating cells from the *P. dumerilii* brain and tested single-cell RNA sequencing protocols. T.L. generated the reference transcriptome and analyzed gene expression. D.A. conceived the project, supervised the study and helped write the manuscript. J.C.M. conceived the project, supervised the study and wrote the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Jaitin, D.A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
2. Tan, S.J. *et al.* A microfluidic device for preparing next generation DNA sequencing libraries and for automating other laboratory protocols that require one or more column chromatography steps. *PLoS ONE* **8**, e64084 (2013).
3. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
4. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. Methods* **10**, 1127–1133 (2013).
5. Lee, J.H. *et al.* Highly multiplexed subcellular RNA sequencing *in situ*. *Science* **343**, 1360–1363 (2014).
6. Durruthy-Durruthy, R. *et al.* Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution. *Cell* **157**, 964–978 (2014).
7. Garcia-Solache, M., Jaeger, J. & Akam, M. A systematic analysis of the gap gene system in the moth midge Clogmia albipunctata. *Dev. Biol.* **344**, 306–318 (2010).

8. Biffar, L. & Stollewerk, A. Conservation and evolutionary modifications of neuroblast expression patterns in insects. *Dev. Biol.* **388**, 103–116 (2014).
9. Pani, A.M. *et al.* Ancient deuterostome origins of vertebrate brain signalling centres. *Nature* **483**, 289–294 (2012).
10. Ferrier, D.E. Evolutionary crossroads in developmental biology: annelids. *Development* **139**, 2643–2653 (2012).
11. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* **9**, 868–882 (2008).
12. Arendt, D., Tessmar-Raible, K., Snyman, H., Dorresteijn, A.W. & Wittbrodt, J. Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain. *Science* **306**, 869–871 (2004).
13. Tessmar-Raible, K. *et al.* Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. *Cell* **129**, 1389–1400 (2007).
14. Tomer, R., Denes, A.S., Tessmar-Raible, K. & Arendt, D. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell* **142**, 800–809 (2010).
15. Randel, N., Bezares-Calderon, L.A., Guhmann, M., Shahidi, R. & Jekely, G. Expression dynamics and protein localization of rhabdomeric opsins in Platynereis larvae. *Integr. Comp. Biol.* **53**, 7–16 (2013).
16. Marlow, H. *et al.* Larval body patterning and apical organs are conserved in animal evolution. *BMC Biol.* **12**, 7 (2014).
17. Pettit, J.B. *et al.* Identifying cell types from spatially referenced single-cell expression datasets. *PLOS Comput. Biol.* **10**, e1003824 (2014).
18. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
19. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
20. Lovatt, D. *et al.* Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* **11**, 190–196 (2014).
21. Puniyani, K., Faloutsos, C. & Xing, E.P. SPEX2: automated concise extraction of spatial gene expression patterns from Fly embryo ISH images. *Bioinformatics* **26**, i47–i56 (2010).
22. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, doi:10.1038/nbt.3192 (14 April 2015).
23. Grange, P. *et al.* Cell-type-based model explaining coexpression patterns of genes in the brain. *Proc. Natl. Acad. Sci. USA* **111**, 5397–5402 (2014).

ONLINE METHODS

Single cell dissociation and capture protocol. *P. dumerilii* larvae were incubated in filtered natural seawater (FNSW), at 18 °C for 47 h to reach the desired stage. For single-cell dissociation, 80–100 47 hpf larvae were picked, washed in Ca- and Mg-free artificial sea water (CMF-SW) and incubated for 5 min in 0.5% Pronase (Roche cat # 10165921001), 1% sodium thioglycolate (Sigma T0632) in CMF-SW. After this treatment, the trunks and epispheres separated. 50–80 epispheres were randomly picked, rinsed in 1:1 FNSW:1× PBS, followed by 1× PBS, and incubated for 1 min in 150 µg/ml Liberase (Roche, cat # 05401119001) in 1× PBS. Epispheres were rinsed in 1× PBS twice, transferred to a nonstick Eppendorf tube and triturated by pipetting. Dissociated cells were washed twice in 200 µl of 1× PBS and concentrated by centrifuging (1 min, 200*g*). Cells were resuspended in 20 µl of 1× PBS, of which 5–15 µl was loaded on the capture chip (Fluidigm cat # 100-5760).

Cell capture, lysis, reverse transcription and cDNA amplification were performed on the C1 IFC for mRNA-seq (10–17 µm 'chip', Fluidigm cat # 100-5760) on Fluidigm C1 Single-Cell Auto Prep System.

Following the company's recommendation, we used SMARTer Ultra Low Input RNA Kit for on-chip reverse transcription and the ADVANTAGE-2 PCR kit (Clontech cat # 634832) for on-chip PCR. We added 1 µl of 1:4,000 dilution of ERCC spike-in RNA (Ambion, cat # 4456740) to the C1 lysis mix; this concentration equals 1 µl of 1:800,000 dilution per cell. Sequencing libraries were prepared using the Nextera XT DNA Sample Preparation kit (Illumina cat # FC-131-1096) and 100 bp paired-end sequences were generated using the Illumina HiSeq2000 platform.

Mapping and quantitation of next-generation sequencing data. We used *bowtie2* (ref. 24) to map the raw sequencing reads to the *P. dumerilii* reference transcriptome (http://4dx.embl.de/arendt/publicdata/Publicdata.html). Expression counts for each gene were obtained using HTSeq1 (refs. 25,26), and normalized by the total number of reads mapped to the transcriptome in order to account for sequencing depth (tpm). For the spatial mapping, we limited our gene expression data set to a curated set of 169 genes that had spatial expression information in the reference ISH database. Of these, we further excluded 71 expression patterns that had previously been shown to be of moderate to poor quality[17], leaving a final set of 98 reference genes in our ISH set. The raw count data for this gene set in all cells are provided in **Supplementary Table 4**.

Sequences of the reference genes in *P. dumerilii* WMISH atlas. Nucleotide sequences for the genes included in the *P. dumerilii* WMISH reference atlas used in this study are provided as **Supplementary Sequences**. Gene sequences were obtained from the following sources: previous publications: WntA (ref. 27), rOpsin, Six1/2, Pax6 (ref. 28); Rx[12]; DHE3.R30 (*P. dumerilii* EST IB0AAA32CF06EM1), Dll, GLT, NK21, Tubby (*P. dumerilii* EST IB0AAA40DH05EM1)[29]; Hb80 (ref. 30); ChAT, Chx10, Dbx, Gsx, Islet, Lhx2, Sim, VAChT[31]; Otp, Phc2, Syt, Vax[13]; Ngn[32]; FVRI, FVRIamide[33], Prox1, Sox2 (ref. 34); miR.277, miR.7 (ref. 35); Gli[36]; Ascl, BF1 (Foxg1), Brn124, COE, Dach, Emx, ER81.H85.H86, Ets3.H71, Svp, Tll (Tlx), TrpHyd, Wnt5, Wnt8 (ref. 14); Trp2 (Steinmetz *et al.*[37]); DLamide; FLamide, FVMamide, FVamide, HIGA, LYamide, NPY (NPY4), RYamide, SPY, WLD, YFamide[38]; RGWamide[39]; VWamide (MIP)[40], Pax258 (ref. 41); rOpsin3 (ref. 15); Proenkephalin[42]; FGFR, GBRL2.Y85, KLF.H38, Tektin.2.R68 (ref. 16); Bsx, Not[43]; NCBI nucleotide database: Dpn.H35 (HES3, GenBank KC999041), HIGA (GenBank KF515947), Otx (GenBank AJ278856), MLDneuropeptide (GenBank KF515945), NGEWneuropeptide (GenBank KF515948); Arendt lab *P. dumerilii* EST sequencing library: CALM.R29 (EST IB0AAA31DB02EM1), cpa (EST IB0AAA35BD12FM1), Dek.H2 (EST IB0AAA15CC12EM1), HEN1.Y61 (EST IB0AAA56YF10EM1), LDB3.R10 (EST IB0AAA19CC08EM1), MyoD.H29 (EST IB0AAA28CH12EM1), p53.63.73.like.H45 (EST IB0AAA34AG01EM1), PRVA-Y49 (Parvalbumin-alpha, EST IB0AAA53YF24FM1), Tekitin.3.R7 (EST IB0AAA18BD02EM1), Tolloid.Y68 (EST IB0AAA57YH01EM1), ZFAT.like. H97 (EST IB0AAA37DA10EM1); Lhx3 (provided by Kristin Tessmar-Raible), Pnr (provided by Maria Antonietta Tosches), Eya and Sepiapterin reductase (provided by Keren Guy).

Gene specificity. Given the set of 98 reference genes and the set of cells that were sequenced, we define an $C \times M$ read count matrix, D, where $D_{c,m}$ describes

the normalized number of reads mapped to cell *c* for gene *m*. Subsequently, for each cell-gene combination, we define its specificity ratio, $r_{c,m}$ as:

$$r_{c,m} = \frac{D_{c,m}}{\frac{1}{C}\sum_{a=1}^{C} D_{a,m}}$$

The specificity scores for each gene in the quality-filtered cells are provided in **Supplementary Table 5**.

Computing the correspondence score. To map back each sequenced single cell to its localization in the reference ISH data set, we developed a scoring system where we calculate the correspondence between every cell-voxel combination.

This scoring system compares the binary vector of expression provided by the WMISH data with a binarized version of the expression pattern for each cell sequenced. To binarize the expression vectors, we used a threshold of ten reads above which a gene was considered expressed.

The score $S_{e_C, e_{ref}}$ between the binary expression vector e_c from single cell *c* and e_{ref} from voxel *ref* in the ISH data set is defined as:

$$S_{c,ref} = \sum_{m=1}^{M} f_{r_{c,m}}(e_{c,m}, e_{ref,m})$$

with

$$f_{r_{c,m}}(e_{c,m}, e_{ref,m}) = \begin{cases} t(r_{c,m}), & e_{c,m} = e_{ref,m} = 1 \\ -t(r_{c,m}), & e_{c,m} = 1, e_{ref,m} = 0 \\ 0, & Otherwise \end{cases}$$

and

$$t(r_{c,m}) = \frac{r_{c,m}}{1 + r_{c,m}}$$

This scoring scheme is designed to assess the correspondence between a single cell and each reference voxel with regard to the specificity ratio of each gene for the considered single cell. The specificity scores are transformed to fall in the interval [0,1] following an algebraic function, *t*, which avoids giving too much weight to exceptionally specific genes and quickly reduces the weight of nonspecific genes that may hinder the precision of the mapping.

The correspondence scores for each cell-voxel combination in the quality-filtered cells are provided in **Supplementary Table 6**.

The method presented above penalizes only mismatches when the gene is expressed in the RNA-seq data and not in the reference atlas (one-way penalization). It is possible to penalize similarly mismatches that arise when a gene is not found to be expressed by the RNA-seq but expressed in the reference atlas (two-way penalization) as detailed in **Supplementary Note 3** and **Supplementary Table 7**. The one-way penalization is justified by noting that current scRNA-seq protocols display a substantial degree of technical noise, meaning that accurately calling a gene as unexpressed in a given cell is challenging.

Selecting the score threshold based on simulated data. For a single cell *c*, once the scores against every voxel in the reference data set are computed and sorted, we need to define a score threshold above which we consider the voxels as the potential area where the single cell came from.

To find this threshold, we conducted a simulation study by generating 13,900 (100 per sequenced cell) random "simulated single cells." Each simulated single cell is created by randomly shuffling the specificity scores for all genes in each sequenced cell.

We then apply the mapping method to this set of randomly generated samples and summarize the results for different thresholds. **Supplementary Figure 3** shows the proportion of simulated cells with at least *n* voxels (*x* axis) above a certain threshold (different lines on the plot). From this null distribution, we chose three different thresholds corresponding to different levels of confidence in the mapping. (i) High confidence mapping for cells that have at least 21 voxels scoring higher than 1.5. The probability of a simulated cell mapping back with the same criteria is less than 10%. (ii) Medium confidence

mapping for cells with at least 16 voxels scoring higher than 0.5. The probability of a simulated cell mapping back with these criteria is less than 30%. (iii) Low confidence mapping for cells that have at least 11 voxels scoring higher than 0. For these cells, the mapping can be considered as providing a strong hypothesis about the location of these cells, although the confidence in the mapping is relatively low (<50% of obtaining such a result by chance in the null set). For comparison, **Supplementary Figure 4** shows the same proportion plot for different thresholds but for the sequenced cells.

R scripts for the analysis. Curated R scripts and example analysis workflows for the computation of gene specificity ratios, correspondence scores and the thresholds for confidence values are provided as **Supplementary Scripts**. Alternatively, the scripts can be found at GitHub (https://github.com/jbogp/nbt_spatial_backmapping).

***In situ* hybridization, imaging and image registration.** *In situ* hybridization (ISH) and dual ISH on 48 hpf *P. dumerilii* larvae was performed as described previously[44]. All the colocalization patterns shown were replicated in at least four animals.

The *P. dumerilii* genes analyzed by ISH were: *ER81* (ref. 14), *Fezf*[16], *Lhx3* (produced by K. Tessmar-Raible), *Proe*[42], *Cux1/2 and Ten3*. For the synthesis of mRNA probes against *Proe*, *Cux1/2* and *Ten3*, wild-type *P. dumerilii* RNA was reverse transcribed using SuperScriptIII reverse transcriptase (Life Technologies, cat. # 18080044), and amplified by PCR using TaKaRa ExTaq DNA polymerase (Clontech, cat. # RR001A) and the following gene specific primers:

Proe (5′-CATTTGCAAGTTCCGAGGTT-3′ and 5′-GCTTGTCACTGG TTGGTTCC-3′),

Cux1/2 (5′-CTGCCCTTGAAGAGGAGTTG-3′and 5′-GACTCCAACGG TTCGATGAT-3′),

Ten3 (5′-ATCTGTAAGCCAGGCTGGAA-3′ and 5′-GGTCGCAAGT GACCGTTTAT-3′).

The resulting PCR fragments were cloned into pCRII-TOPO vector (Life Technologies, cat # K4610-20). For the synthesis of ISH probes, cDNA plasmids were linearized and antisense RNA probes were transcribed using SP6 or T7 RNA polymerase (Roche, cat. #11487671001 and 10881775001, respectively) and DIG RNA-labeling mix (Roche, cat. #11277073910)Top of FormBottom of Form or Fluorescein RNA labeling mix (Roche, cat. # 11685619910).

For imaging of ISH samples, samples were mounted in 97% 2,2′-thiodiethanol (Sigma, cat. # 166782) and imaged on Leica TCS SP8 confocal microscope, using a combination of fluorescence and reflection[45] microscopy. The colocalization analyses and image post-processing was performed using Fiji[46] software. The figure panels were compiled using Adobe Illustrator and Adobe Photoshop software. Image registration and gene expression averaging was performed as previously described[47].

24. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
25. Anders, S., Pyl, P.T. & Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
26. Fonseca, N.A., Petryszak, R., Marioni, J. & Brazma, A. iRAP - an integrated RNA-seq analysis pipeline. *bioRxiv* http://dx.doi.org/10.1101/005991 2014.
27. Prud'homme, B., Lartillot, N., Balavoine, G., Adoutte, A. & Vervoort, M. Phylogenetic analysis of the Wnt gene family. Insights from lophotrochozoan members. *Curr. Biol.* **12**, 1395 (2002).
28. Arendt, D., Tessmar, K., de Campos-Baptista, M.I., Dorresteijn, A. & Wittbrodt, J. Development of pigment-cup eyes in the polychaete Platynereis dumerilii and evolutionary conservation of larval eyes in Bilateria. *Development* **129**, 1143–1154 (2002).
29. Raible, F. *et al.* Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii. *Science* **310**, 1325–1326 (2005).
30. Kerner, P. *et al.* The expression of a hunchback ortholog in the polychaete annelid Platynereis dumerilii suggests an ancestral role in mesoderm development and neurogenesis. *Dev. Genes Evol.* **216**, 821–828 (2006).
31. Denes, A.S. *et al.* Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell* **129**, 277–288 (2007).
32. Simionato, E. *et al.* atonal- and achaete-scute-related genes in the annelid Platynereis dumerilii: insights into the evolution of neural basic-Helix-Loop-Helix genes. *BMC Evol. Biol.* **8**, 170 (2008).
33. Jekely, G. *et al.* Mechanism of phototaxis in marine zooplankton. *Nature* **456**, 395–399 (2008).
34. Kerner, P., Simionato, E., Le Gouar, M. & Vervoort, M. Orthologs of key vertebrate neural genes are expressed during neurogenesis in the annelid Platynereis dumerilii. *Evol. Dev.* **11**, 513–524 (2009).
35. Christodoulou, F. *et al.* Ancient animal microRNAs and the evolution of tissue identity. *Nature* **463**, 1084–1088 (2010).
36. Dray, N. *et al.* Hedgehog signaling regulates segment formation in the annelid Platynereis. *Science* **329**, 339–342 (2010).
37. Steinmetz, P.R., Kostyuchenko, R.P., Fischer, A. & Arendt, D. The segmental pattern of otx, gbx, and Hox genes in the annelid *Platynereis dumerilii. Evol. Dev.* **13**, 72–79 (2011).
38. Conzelmann, M. *et al.* Neuropeptides regulate swimming depth of Platynereis larvae. *Proc. Natl. Acad. Sci. USA* **108**, E1174–E1183 (2011).
39. Conzelmann, M. & Jekely, G. Antibodies against conserved amidated neuropeptide epitopes enrich the comparative neurobiology toolbox. *Evodevo* **3**, 23 (2012).
40. Conzelmann, M. *et al.* Conserved MIP receptor-ligand pair regulates Platynereis larval settlement. *Proc. Natl. Acad. Sci. USA* **110**, 8224–8229 (2013).
41. Backfisch, B. *et al.* Stable transgenesis in the marine annelid Platynereis dumerilii sheds new light on photoreceptor evolution. *Proc. Natl. Acad. Sci. USA* **110**, 193–198 (2013).
42. Jekely, G. Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proc. Natl. Acad. Sci. USA* **110**, 8702–8707 (2013).
43. Tosches, M.A., Bucher, D., Vopalensky, P. & Arendt, D. Melatonin signaling controls circadian swimming behavior in marine zooplankton. *Cell* **159**, 46–57 (2014).
44. Tessmar-Raible, K., Steinmetz, P.R., Snyman, H., Hassel, M. & Arendt, D. Fluorescent two-color whole mount *in situ* hybridization in Platynereis dumerilii (Polychaeta, Annelida), an emerging marine molecular model for evolution and development. *Biotechniques* **39**, 460, 462, 464 (2005).
45. Jekely, G. & Arendt, D. Cellular resolution expression profiling using confocal detection of NBT/BCIP precipitate by reflection microscopy. *Biotechniques* **42**, 751–755 (2007).
46. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
47. Asadulina, A., Panzera, A., Veraszto, C., Liebig, C. & Jekely, G. Whole-body gene expression pattern registration in Platynereis larvae. *Evodevo* **3**, 27 (2012).