Choice of alignment pipeline strongly influences clustering quality of scRNA-seq datasets

Inbal Shainer[1] and Manuel Stemmer[1,2]

[1] Max Planck Institute of Neurobiology, Am Klopferspitz 18, 82152 Martinsried

[2] corresponding author

Abstract

Background

Single-cell RNA sequencing (scRNA-seq) has quickly become one of the most dominant techniques in modern transcriptome assessment. In particular, 10X Genomics' Chromium system, with its high throughput approach, turn key and thorough user guide made this cutting-edge technique accessible to many laboratories using diverse animal models. However, standard downstream processing, including the alignment and cell filtering pipelines might not be ideal for every organism or tissue. Here we applied an alternative strategy, based on the pseudoaligner kallisto, on twenty-two publicly available single cell sequencing datasets from a wide range of tissues of eight organisms and compared the results with the standard 10X Genomics' Cell Ranger pipeline.

Results

In most of the tested samples, kallisto outperformed Cell Ranger in sequencing read alignment rates and total gene detection rates. Although datasets processed with Cell Ranger had higher cell counts, outside of human and mouse datasets, these additional cells were routinely of low quality, containing low gene detection rates. Thorough downstream analysis of one kallisto processed dataset, obtained from the zebrafish pineal gland, revealed

clearer clustering, allowing the identification of an additional photoreceptor cell type that previously went undetected. The finding of the new cluster suggests that the photoreceptive pineal gland is essentially a bi-chromatic tissue containing both green and red cone-like photoreceptors and implies that the alignment and processing pipeline can affect the discovery of biologically-relevant cell types.

Conclusion

While Cell Ranger favors higher cell numbers, using kallisto results in datasets with higher median gene detection per cell. We could demonstrate that cell type identification was not hampered by the lower cell count, but in fact improved as a result of the high gene detection rate and the more stringent filtering. It is thus beneficial to favor high quality cells and accept a lower cell count, leading to an improved classification of cell types.

Keywords

Single-cell RNA-Sequencing, Cell Ranger, kallisto, zebrafish, pineal gland, alignment, opsin, 10X Genomics

Background

Single-cell transcriptome sequencing (scRNA-seq) has rapidly become one of the most popular tools for dissecting the transcriptomic states of individual cells in a tissue of interest. It can be applied to virtually any biological sample as long as a reference genome is available. Among the available scRNA-seq techniques, the Chromium (10X Genomics) platform is probably the most widely used at this point. Thanks to its user-friendly design and very well-

documented workflow, it has quickly emerged as the top choice for many researchers and clinicians (Cao et al., 2019; Davie et al., 2018; Kölsch et al., 2020; Packer et al., 2019; Pandey et al., 2018; Peuß et al., 2020; Shainer et al., 2019; Wang et al., 2020). Its droplet-based design and simple workflow make it the ideal technique for surveying hundreds to thousands of cells in a single experiment, without needing prior knowledge of the system (Svensson et al., 2018; Zheng et al., 2017). In addition to its hardware, 10X Genomics also provides a whole suite of tools called Cell Ranger for processing sequencing reads (demultiplexing, alignment, filtering, dimensionality reduction and visualization of clusters) (10xgenomics.com). Within this package, STAR (Dobin et al., 2013) is applied to align the millions of short reads, typically produced by Illumina sequencers. This aligner has been shown to work accurately and reliably and is widely used.

Besides STAR and other classic aligners like GSNAP or TopHat2 (Dobin et al., 2013; Kim et al., 2013; Wu and Nacu, 2010), there have been recent advances in the development of so called pseudoaligners. Instead of trying to exactly evaluate the alignment of each base in a given read, pseudoaligners only focus on the potential identity of the target transcript (Bray et al., 2016). Among the few existing tools like Salmon, Sailfish or kallisto (Bray et al., 2016; Patro et al., 2014; Patro et al., 2017), kallisto comes with a suite of complementary tools for processing and filtering scRNA-seq reads, making it as simple to use as Cell Ranger. Importantly, the computational resources required for kallisto are marginal in comparison to Cell Ranger/STAR, with an entire scRNA-seq run, can be aligned on a standard laptop within tens of minutes (Bray et al., 2016) instead of hours for Cell Ranger/STAR. The alignment process is sped up by a magnitude by splitting up the reads into *k*-mers and matching them using hash tables, while the accuracy is maintained by constructing a transcriptome de

Bruijn graph (Bray et al., 2016; Melsted et al., 2019). Kallisto can also align reads coming from bulk RNA-seq, and the introduction of the Barcode-UMI-Set (BUS) format allows processing and comparison of scRNA-seq data originating from various sources (Melsted et al., 2019). In addition to the alignment of the sequencing reads by either STAR or kallisto, the transcripts have to be associated with their respective cell barcodes and the unique molecular identifiers (UMIs) have to be counted (Klein et al., 2015). As a last step, empty droplets are removed from the dataset before one can proceed to downstream processing (Lun et al., 2019; Macosko et al., 2015). All these functions are conveniently included in the Cell Ranger pipeline, but are usually not obvious to the user.

While Cell Ranger is an all-in-one package, the kallisto pipeline consists of three parts. Kallisto (Bray et al., 2016) aligns the sequencing reads, while bustools (Melsted et al., 2019) associates the reads to the respective cell barcodes, collapses the UMIs, counts the identified transcripts and creates the cell to gene matrix. In the third part we use DropletUtils (Griffiths et al., 2018) to rank barcodes according to detected UMI levels and define the inflection point within the resulting knee-plot as a threshold between droplets that are thought to contain cells and empty droplets. The individual packages are all open-source and the process can be inspected in detail (Bray et al., 2016; Griffiths et al., 2018; Melsted et al., 2019). We refer to this entire pipeline as “kallisto”.

Although based on the same principles, the underlying approach of kallisto and Cell Ranger differ and so are potentially the resulting datasets. Here we show that alternative alignment strategies can indeed make a difference in the derived biological conclusions, especially when working with organisms other than human or mouse. Using kallisto we found a strong

increase in alignment rates (percent of reads aligned to reference transcriptome) and consequent gene detection rates within most of the tested Chromium v2/v3 datasets (zebrafish, killifish, cavefish, drosophila, c. elegans, ciona, mouse, human). Additionally, we show that the choice of reference genome and differences in thresholding empty droplet removal has strong effects on correct filtering and resulting cell counts. Finally we demonstrate, using a dataset obtained from the zebrafish pineal gland (Shainer et al., 2019), that alignment with kallisto and cell filtering with DropletUtils (Griffiths et al., 2018) drastically improve clustering accuracy and reveal new cell types.

Results

To uncover whether the choice of alignment tool affects the downstream results, we analyzed twenty-two published single-cell sequencing datasets from eight different organisms (Table 1). The first difference between the two approaches was the overall higher alignment rates of sequencing reads to the transcriptome for kallisto (on average 7.2% increase) with the only exception of drosophila dataset s1 (Fig. 1a). Furthermore, total gene detection rates were increased in the kallisto samples in comparison to Cell Ranger, with the exception of the *C. elegans* datasets. Most of the identified genes were shared between the two pipelines, but for teleost and mammal samples, kallisto detected considerably more genes than Cell Ranger (Fig. 1b and Additional file 1: Figure S1). When looking at the median gene count (MGC) and median UMI count (MUC) per cell, we found again increases in the majority of samples processed with kallisto, with one exception: mouse sample mm-neuron-2k-v2 (Fig. 1c, d). In contrast, cell counts were lower for all samples processed with kallisto, except for the human samples and mouse sample mm-neuron-1k-v3 (Fig. 1e).

Both classic aligners as well as pseudoaligners require an annotated reference transcriptome for alignment, which is typically generated from a gene transfer format (GTF) file and corresponding genomic resource. To assess the influence of transcriptome reference choice on alignment results, we ran our zebrafish samples also along a recently improved zebrafish transcriptome annotation (zta) v4.3.2 (Lawson et al., 2020) (Fig. 2). In all the zebrafish datasets, zta v4.3.2 improved the alignment and gene detection rates, while differences in cell counts were minor (Fig. 2). Similar to results seen for the alignment to the Ensembl reference (101), kallisto analyzed datasets had higher MGC and MUC when aligned to zta 4.3.2 in comparison to Cell Ranger. Unlike classic aligners, kallisto also allows alignment to a pure transcriptomic input. This is an advantage for species, where a GTF file and genomic resource are not available. So, in this case we also included a reference extraction from Ensembl's BioMart. Although based on the same Ensembl version, the reference constructed from BioMart contains more transcript annotations. However, it only marginally improved gene detection rates in comparison to the genome annotation (Fig. 2). This analysis shows that the choice of reference can impact total gene detection, MGC, MUC and cell count for both kallisto and Cell Ranger, but regardless the reference choice, Cell Ranger provided higher cell counts, while kallisto outperformed in total gene detection, MGC and MUC. Additionally, the flexibility of kallisto enables a wider choice of options of alignment references.

Next, we looked at the cellular distribution in relation to their detected genes. Interestingly, all samples processed with Cell Ranger included a population of cells at the lower end of MGCs, at around 300 to 500 genes per cell (Fig. 3 and Additional file 1: Figure S2 and S3).

The only exceptions here were the human samples (Fig. 3f and Additional file 1: Figure S2 and S3). This effect was also less pronounced in the mouse datasets (Additional file 1: Figure S2 and S3). The kallisto pipeline on the other hand excluded most to all of these low MGC cells (Fig. 3 and Additional file 1: Figure S3).

What is considered a cell in scRNA-seq data is based on finding an appropriate cutoff between what are thought to be empty droplets and droplets containing a cell. Since this is commonly measured only indirectly by looking at transcript levels beyond a certain threshold in the knee plots, the numbers of cells are heavily dependent on the applied filter (Lun et al., 2019; Macosko et al., 2015). Thus, in a next step we wanted to test, whether the datasets processed with the kallisto pipeline still contain cells with higher MGCs, when the thresholds are set to the same levels as Cell Ranger. Therefore, we did not identify the inflection point automatically using DropletUtils, but forced kallisto to reach the same cell counts as Cell Ranger ('kallisto forced'), which resulted in an overlap between the majority of cells (Additional file 1: Figure S4).

Kallisto forced still showed overall better MGCs and MUCs per sample, however less pronounced with one more exception (dm-brain-s1) (Additional file 1: Figure S5) and total gene detection slightly increased (Additional file 1: Figure S6). Interestingly, the population of cells containing low gene count, that was previously only detected within the Cell Ranger data, was now evident (Additional file 1: Figure S7). Hence, the differences appear to stem from the filtering process taking care of the empty droplet removal.

To test whether the noticeable differences in gene or cell number detection between the two pipelines alter the ability to characterize the cellular composition of the tissue in the

downstream analysis, we performed principal component analysis (PCA) and clustering of the zebrafish pineal sample number two (dr-pineal-s2), which was previously studied in detail by one of the authors (Shainer et al., 2019).

In the previous analysis of this dataset (Shainer et al., 2019), dr-pineal-s2 was aligned to the zebrafish GRCz10 genome assembly (Ensembl release 90) with Cell Ranger version 2.0.2. A total of 2,266 cells were detected (Shainer et al., 2019) and eight cell types, including the pineal gland rod- and cone-like photoreceptors (PhR), five other pineal cell types (retinal pigment epithelium-like, Müller-like glia, neurons, macrophages/microglia and blood cells) as well as habenula neurons (Shainer et al., 2019).

In the current analysis, the dr_pineal_s2 dataset was aligned with the new Cell Ranger version (v5) to the latest Ensembl release (101), which resulted in more than double of the total number of cells (6,769) (Fig. 1b and Additional file 1: Figure S2 and S8a).
The dataset was filtered using similar parameters as in the original publication (see Methods) and the observed cell population with low gene detection in the Cell Ranger object was still present (Additional file 1: S8). These cells clustered into 13 types (Fig. 4a, b and Additional file 1: Figure S9a, b and Additional file 2: Table S1), which included the previously identified clusters as well as two types of habenula cells, erythrocytes, fibroblasts, vascular endothelial cells, leukocytes and epithelial cells (Fig. 4a, b). The additional types of cells, although not identified in the previous study (Shainer et al., 2019), are possibly a result of the higher cell count and improved Cell Ranger version, and expected to reside in the pineal as well as in any other brain tissue.

Analysis of the same dataset using kallisto, resulted in a total number of 5,810 cells (Fig. 1b and Additional file 1: Figure S3 and S8b), and the population of cells with the low gene count was excluded. These cells clustered into 14 types (Fig. 4c, d and Additional file 1: Figure S9e, f and Additional file 2: Table S1), which included all the types identified by the Cell Ranger analysis, and an additional cone-like PhR cluster (Fig. 4c, d). The two cone-like PhR clusters differed in the opsins they expressed, with one expressing *opn1lw1* and *opn1lw2*, the red opsins (long wavelength sensitive), and the other expressing *parietopsin*, a green opsin. To further explore the discrepancy of the cluster identification between the pipelines, we checked for *parietopsin* expression in the Cell Ranger data (Fig. 4b). Although *parietopsin* was identified within the cone-like PhRs, no distinction between the red and the green cones is made even after more than doubling Seurat's FindClusters resolution parameter (Additional file 1: Figure S9c, d), which can increase the number of clusters, while preserving dimensional reduction. We refrained from changing any other standard parameter in order to maintain similarity of the pipelines. Analysis of the top differentially-expressed (DE) genes within the cone-like PhRs, revealed that for the kallisto dataset, the gene *col14a1b* (*collagen, type XIV, alpha 1b)* was the most DE gene in the red cone-like PhRs (Fig 4d and Additional file 2: Table S1). This gene was not detected at all by the Cell Ranger alignment, and therefor did not exist as a DE gene (Fig 4b and Additional file 2: Table S1).

To test whether the higher total gene detection rate in the kallisto dataset was the sole cause for identification of the new cluster, or whether the population of cells with the low gene count in the Cell Ranger data hampered its detection, we also performed downstream analysis of the kallisto forced dataset (Additional file 1: Figure S8g). This data combines the total gene detection rate of kallisto, as well as the population of cells with the low gene

count of Cell Ranger. We did not detect the new cluster of the green cone-like PhR in the kallisto forced at the beginning. However, increasing the clustering resolution from 0.9 to 1.2 revealed the new cluster (Additional file 1: Figure S9g, h). With this resolution, the gene *col14a1b* was the most DE gene in the red cone-like PhR for the kallisto forced as well (Additional file 1: Figure S9e, f, g, h). These findings suggest that not only the additional gene affected the identification of the new cluster, but that the population of cells with low gene count was detrimental to the cell type detection and required a higher resolution to reach the same results.

Discussion

We compared two scRNA-seq read alignment and filtering pipelines; Cell Ranger, the standard tool and part of the 10x Genomics package and the lightweight pseudoaligner, kallisto, on a diverse range of scRNA-seq datasets across the animal phylum. We specifically focused on the latest and most used Chromium chemistries (v2/v3) and compared over a range of different species. Furthermore, we aimed to compare directly with the standard parameters of the Cell Ranger pipeline, as this is the most common way to prepare Chromium scRNA-seq data.

Kallisto outperformed the standard Cell Ranger pipeline in overall alignment rates of the reads to the transcriptome, total gene detection rates, MGC & MUC, for most datasets. One of the reasons could be the ability of kallisto to discern reads mapping to multiple loci in the reference (Bray et al., 2016). This feature allows better distinction of paralogous genes (Deschamps-Francoeur et al., 2020), which, for example, are present at high degrees in all teleost species, due to their shared whole genome duplication event (Glasauer and

Neuhauss, 2014). Handling multi-mapped reads has been shown to increase alignment accuracy, resulting in better transcript detection and quantification (Deschamps-Francoeur et al., 2020). One of the most popular approaches to correctly count multi-mapped reads is expectation maximization, which has been included already in several tools like RSEM (Li and Dewey, 2011), Salmon (Patro et al., 2017) as well as kallisto (Bray et al., 2016; Melsted et al., 2019). Cell Ranger on the other hand discards any read that maps to more than one locus in the reference. This might result in the lower total gene detection of Cell Ranger in comparison to kallisto, and influences the DE genes defining the clusters, as in the case of *col14a1b,* which was detected by kallisto in the red cone-like PhRs.

Two recent publications comparing STAR and kallisto (Du et al., 2020; Vieth et al., 2019) showed that applying STAR (Dobin et al., 2013) results in better gene detection than kallisto, depending on the applied scRNA-seq technology and reference. This stands in contrast to our findings. However, the authors compare over different scRNA-seq technologies and only tested human or mouse datasets (Du et al., 2020; Vieth et al., 2019) without looking at other, non-mammalian samples, in which the differential filtering, observed in our study, was detected. We also focus on samples exclusively prepared with the latest 10X Genomics' v2/v3 chemistries, in which the improved capture rate leads to a bigger pool of available transcripts. Furthermore, we compare kallisto directly to Cell Ranger, the commonly used pipeline, which could very well lead to discrepancies in filtering and the outcome of downstream processing.

When looking at cell counts after thresholding out empty droplets, the numbers were higher for most of the samples when run with Cell Ranger. As mentioned before, the number of

cells is defined by a threshold drawn between potentially empty and cell containing droplets (Lun et al., 2019; Macosko et al., 2015). This process is quite different between Cell Ranger and kallisto. The cut-off in Cell Ranger was generally much higher, favoring higher cell counts, but with the potential integration of cells with low gene detection rates (Fig. 3), as seen consistently in all the non-mammalian datasets (Additional file 1: Figure S2). Due to the underlying technology and the lack of ground truth about the exact number of intact cells passing the droplet formation and library preparation, it is difficult to test, whether these are real cells or cells of lower quality (Ilicic et al., 2016). The number of these cells were reduced or not present in the mouse and human samples. One of the reasons could be that the standard parameters of Cell Ranger work better for mouse and human datasets or that these samples were of higher quality to begin with (Fig. 3).

To find out whether the cell population with low gene detection rates impacts the scientific conclusions of underlying biology, we performed clustering analysis of the zebrafish pineal dataset (Fig. 4). The pineal cellular composition in non-mammalian vertebrates shows similarities to the retina. It is a photoreceptive tissue that regulates circadian rhythms and seasonal changes and therefore contains rod-like and cone-like PhRs (Ekström and Meissl, 1997; Falcón et al., 2006). Using kallisto, we were able to further differentiate the cone-like PhRs into two clusters with one distinguished by red opsins (long wavelength sensitive) and the other by green opsins (Fig. 4). This finding suggests that the pineal gland could be a bi-chromatic photoreceptive tissue, containing cones sensitive to different wavelengths. Mutual exclusive expression of *opn1lw1* and *parietopsin* in the pineal cone-like PhRs was recently demonstrated (Cau et al., 2019), and thus, the additional cluster found with kallisto reflects a true biological finding, previously undetected when using the Cell Ranger platform

with the standard parameters. In contrast to the Cell Ranger data, the additional cluster was detected with kallisto forced but only when increasing the clustering resolution (Additional file 1: Figure S9). Merely due to the increased overall gene detection, it became possible to identify the additional cluster, while it was the altered filtering of the kallisto pipeline that reduced the noisiness in the data and improved cell type detection.

Although all the processing was performed using standard conditions in order to ensure the best comparison between datasets as possible, several steps in the pipelines can be tuned further, which can result in the detection of additional genes and clusters. Thus, we cannot rule out the detection of the green cone-like PhRs with Cell Ranger under non-standard analysis parameters. For instance, one could work with the unfiltered output of Cell Ranger and change the filtering threshold to the knee point instead of the inflection point or use tools like EmptyDrops (Lun et al., 2019) for this step. Additionally, other parameters in the downstream analysis can be adjusted such as the normalization method, number of variable genes, number of principal components or clustering algorithm. This however might not be applicable for the routine use of scRNA-seq, which is becoming increasingly popular, and requires an advanced knowledge.

The tuning of these parameters becomes even more challenging when working with tissues where the underlying biology is not well understood and the composition of cell types is unknown. It is thus beneficial to work with an optimized entry point of the analysis for both the unexperienced user as well as when working with previously unstudied tissue. Looking at the here tested diverse datasets, kallisto provides a better entry point to analyze scRNA-seq data, resulting in increased data quality and clustering accuracy. Besides the mentioned

multi-mapping ability, another big advantage is its flexibility concerning the reference input. It only needs a transcriptomic reference as a starting point, which for some organisms might be the only available option or the better annotated version. In addition, kallisto is extremely fast and light weight and can be executed on standard machines (Bray et al., 2016; Melsted et al., 2019). It is under constant maintenance and allows the simple integration of various scRNA-seq platforms, making it extremely versatile (Melsted et al., 2019). Furthermore, we find it as easy to use as Cell Ranger and can be easily integrated into a custom workflow (see Methods section).

Conclusions

Comparison of the alignment results between Cell Ranger, the standard 10X Genomics-based tool, and the alternative pipeline kallisto, across a wide spectrum of tissues and organisms, revealed improved read alignment and gene detection rates with kallisto across almost all samples. Cell Ranger on the other hand consistently favored higher cell numbers, which mostly included cells of lower quality. Thorough analysis of one of the datasets revealed that clustering analysis is more accurate and biologically meaningful, when kallisto is applied. In order to achieve accurate clustering, it is better to have high quality datasets with high gene detection rates, even if this results in fewer total cells. Depending on the origin of the dataset, we suggest to run alternative pipelines side by side and judge on an individual basis. High gene detection and stringent filtering positively impacts on cell type classification, which is the primary goal for most of the scRNA-seq experiments. Thus, choosing the best possible option is crucial.

Methods

Datasets

A list of all the used datasets and accession numbers can be found in Table 1. The fastq files were directly used for the alignments with no extra trimming of the reads. We stick to the most standard parameters and transcriptomic references (Ensembl genomes & annotations) for Cell Ranger and kallisto to simulate standard usage of these pipelines.

Alignment with Cell Ranger

All datasets were aligned on a cluster node with Cell Ranger version 5.0. The respective genome references and gene transfer format (GTF) files were obtained from Ensembl version 100/101 and prepared with Cell Ranger's mkref function. The alignment was run with standard parameters.

Alignment with kallisto

The index files for the kallisto workflow were generated using either "kb ref" (kb-python 0.24.4; (https://github.com/pachterlab/kb_python) to create the index from the same genome and gtf files (Ensembl version 100/101) as for the Cell Ranger references. For creating the index files of just transcriptomic input, "kallisto index" was used (https://pachterlab.github.io/kallisto/). The alignment was then performed with "kb count" of the kb-python package. This package conveniently wraps kallisto and bustools into one. For downstream analysis (e.g. Seurat) it is more convenient to work with gene names instead of gene IDs. In order to conveniently change this, we created a helper script to modify the final matrix accordingly. This python script also enables batch processing of several datasets from a sample sheet in an easy to handle command-line interface (CLI) (https://github.com/mstemmer/kb-helper).

Empty droplet filtering

While Cell Ranger takes care of the empty droplet filtering, the kallisto datasets were imported into R with busparse 1.3.0, their UMI counts were ranked using DropletUtils 1.9.0 and the empty droplets were removed by defining the inflection point on the resulting knee plot (lower cutoff = 500). The Cell Ranger matrices were loaded from their standard filtered output. We provide the corresponding R scripts with our kb helper tool (https://github.com/mstemmer/kb-helper).

Downstream processing with seurat

The matrices were converted into objects for seurat v4.0, applying the standard filtering options (min.cells = 3; min.features = 200). Total cell & gene detection, median gene counts & median UMI counts were then extracted from the seurat objects. For better visualization purposes, the cell distribution plots were roughly filtered (nFeature_RNA < 6000 & nCount_RNA < 25000). The dataset dr-pineal-s2 was chosen for downstream analysis of the data. Therefore, the dataset was further filtered (nFeature_RNA > 200 & nCount_RNA < 15000 & percent_mito < 30) (Additional file 1: Figure S8), similar to the original publication (Shainer et al., 2019), with the exception of percent_mito that was not used originally. The data was then log normalized and scaled (using Seurat's default parameters). Variable features were detected (Seurat's default parameters) and used for principal component analysis. PCs 1-20 were used as dimensions of reduction to compute the k.param nearest neighbors (Seurat's FindNeighbors function). Clustering analysis was performed with resolution of 0.9 or 1.2 (Seurat's FindClusters function). UMAP was used to visualize the datasets in 2D, with the same input PCs as the clustering analysis (Seurat RunUMAP

function). The top markers for each cluster were computed (Seurat's FindAllMarkers functions) and used for cluster identifications (Fig. 4, Additional file 1: Figure S9 & Additional file 2: Table S1). The markers were compared to the original publication (Shainer et al., 2019), while newly detected clusters were identified by comparing their marker genes to what is known in literature. The clustering analysis resulted in sub-clusters of the known pineal clusters (rod-like PhRs, RPE-like and Müller glia-like) as well as the habenula *kiss1* neurons and the leukocytes. The sub-clusters which had similar markers, varying only in expression levels, were merged to simplify the visualization. The sub-clusters before merging, and their respective markers, are illustrated in (Additional file 1: Figure S9).

Declarations

Ethics approval and consent to participate

Not applicable. No animals were used in this study.

Availability of data and materials

All data are freely accessible (Table 1). The script for running our kallisto pipeline can be found here: https://github.com/mstemmer/kb-helper

Competing interests

The authors declare no competing interests.

Funding

The Alexander von Humboldt foundation research fellowship (I.S).

Authors' contributions

I.S. wrote and structured the manuscript and analyzed zebrafish pineal dataset. M.S. designed study and analysis pipeline and wrote and structured the manuscript.

Acknowledgements

We thank Herwig Baier and the entire Baier lab for their support and constructive feedback. We especially thank Shachar Sherman, Gregory Marquart, Antonio Miguel Fernandes and Ashley Parker for their careful review of the manuscript.

Tables & Figures

Table 1

sample ID	organism	common name	tissue	10X version	FACS sorted	source	reference	ensembl reference version
dr-RGC-larva-s1	*D. rerio*	zebrafish	Retinal ganglion cells (RGCs)	2	yes	GSE152842	Koelsch et al., 2020	101
dr-RGC-adult-s17	*D. rerio*	zebrafish	RGCs	2	yes	GSE152842	Koelsch et al., 2020	101
dr-pineal-s1	*D. rerio*	zebrafish	pineal gland	2	no	SRR8315377/SRR8315378	Shainer et al., 2019	101
dr-pineal-s2	*D. rerio*	zebrafish	pineal gland	2	no	SRR8315379/SRR8315380	Shainer et al., 2019	101
nf-embryo-s1	*N. furzeri*	killifish	tail blastema	2	no	SRR11886701	Wang et al., 2020	101
nf-embryo-s2	*N. furzeri*	killifish	tail blastema	2	no	SRR11917464	Wang et al., 2020	101
am-hk-cave	*A. mexicanus*	mexican cavefish	head kidney	3	yes	GSE128306/L39134	Peuß et al., 2020	100
am-hk-surf	*A. mexicanus*	mexican cavefish	head kidney	3	yes	GSE128306/L39132	Peuß et al., 2020	100
dm-brain-s1	*D. melanogaster*	drosophila	brain	2	no	SRR6327103	Davie et al., 2018	100
dm-brain-s2	*D. melanogaster*	drosophila	brain	2	no	SRR6327104	Davie et al., 2018	100
ce-embryo-s1	*C. elegans*	C. elegans	embryo	2	no	SRR8611991	Packer et al., 2019	100
ce-embryo-s2	*C. elegans*	C. elegans	embryo	2	no	SRR8611992	Packer et al., 2019	100
ci-larva-s1	*C. intestinalis*	ciona	larva	2	no	SRR9051005	Cao et al., 2019	100
ci-larva-s2	*C. intestinalis*	ciona	larva	2	no	SRR9051006	Cao et al., 2019	100
hs-pbmc-4k-v2	*H. sapiens*	human	pbmc	2	no	*	10X Genomics	100
hs-pbmc-8k-v2	*H. sapiens*	human	pbmc	2	no	*	10X Genomics	100
mm-neuron-2k-v2	*M. musculus*	mouse	neurons	2	no	*	10X Genomics	100
mm-neuron-9k-v2	*M. musculus*	mouse	neurons	2	no	*	10X Genomics	100
hs-pbmc-10k-v3	*H. sapiens*	human	pbmc	3	no	*	10X Genomics	100
hs-pbmc-1k-v3	*H. sapiens*	human	pbmc	3	no	*	10X Genomics	100
mm-neuron-10k-v3	*M. musculus*	mouse	neurons	3	no	*	10X Genomics	100
mm-neuron-1k-v3	*M. musculus*	mouse	neurons	3	no	*	10X Genomics	100
						* https://support.10xgenomics.com/single-cell-gene-expression/datasets		

Figure 1

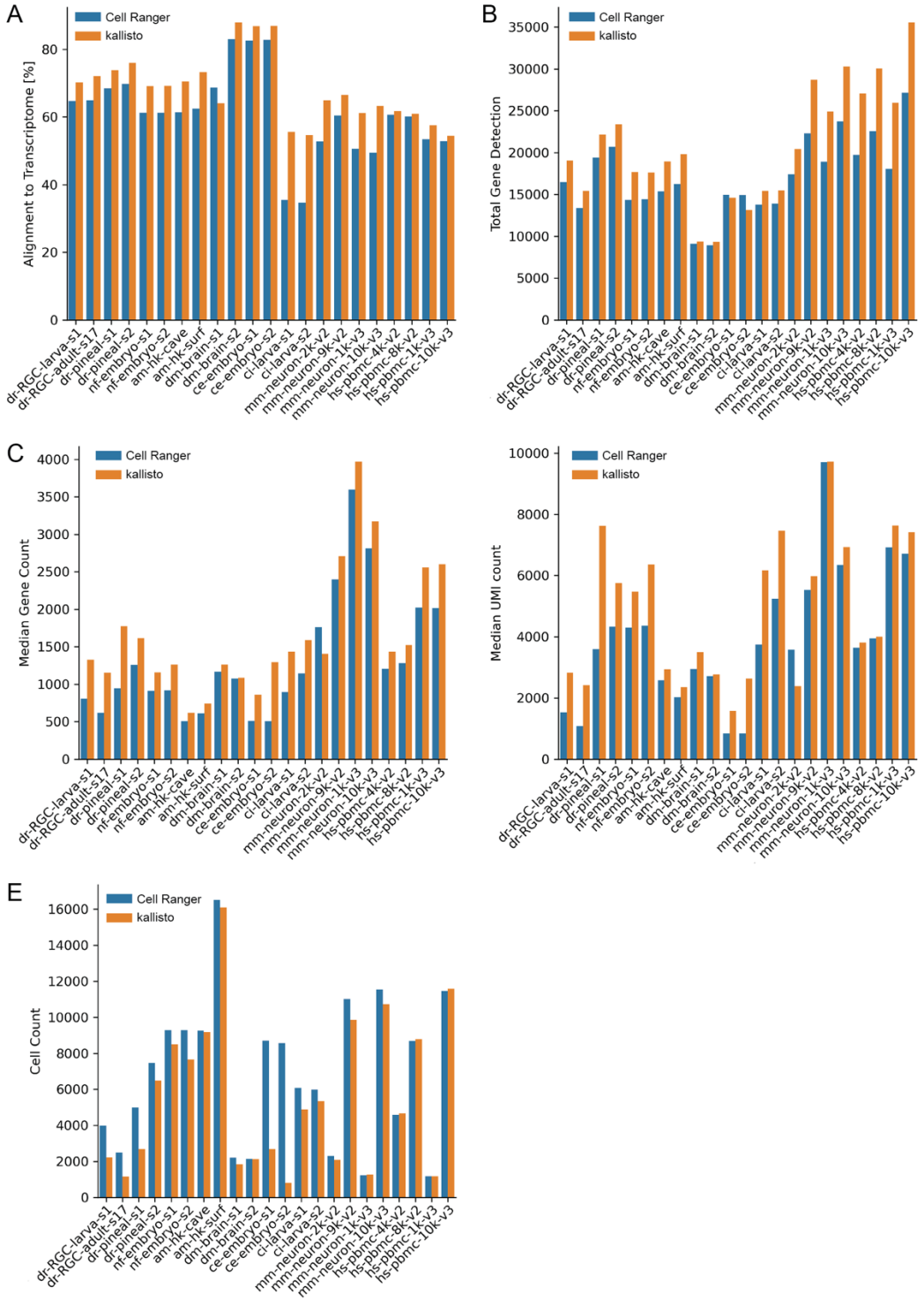

Fig. 1 Alignment results of all datasets (Table 1) run with either Cell Ranger or kallisto against Ensembl reference. **a** Percent alignment rates of reads against the reference transcriptome. **b** Total gene detection. **c** Median gene counts over all cells per dataset. **d** Median UMI counts over all cells per dataset. **e** Total cell counts of each dataset.

Figure 2

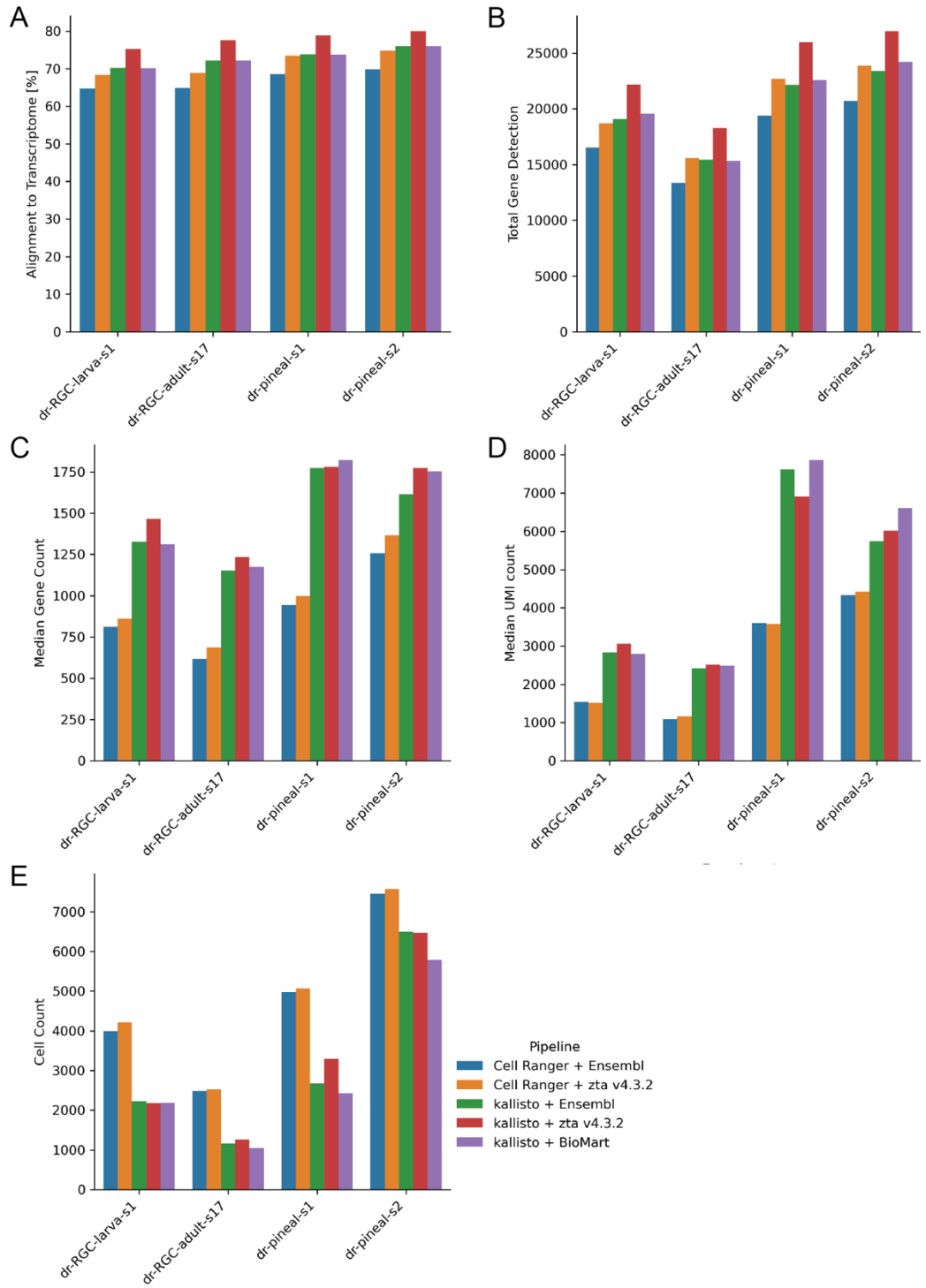

Fig. 2 Impact of reference choice and alignment pipeline on zebrafish scRNA-seq datasets. **a** Percent alignment rates of reads against the reference transcriptome. **b** Total gene detection. **c** Median gene counts over all cells. **d** Median UMI counts over all cells. **e** Total cell counts of each dataset.

Figure 3

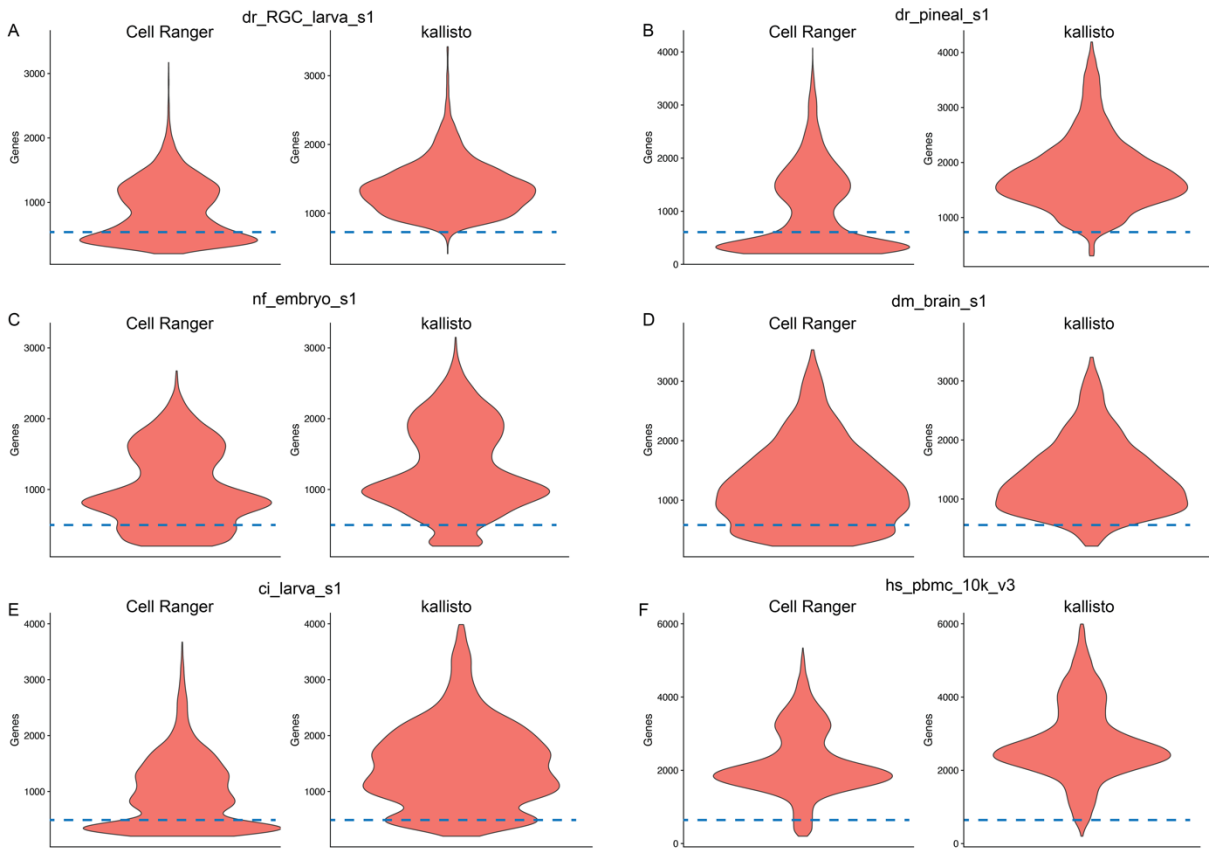

Fig. 3 Violin-plots showing comparison of gene detection per cell between kallisto and Cell Ranger across selected examples. Cell Ranger datasets contain an additional cell population with lower gene detections, not present in the kallisto datasets (at and below dashed lines). This observation was not seen for the human datasets (**f**). The remaining datasets can be found in Additional file 1: Figure S2 and S3.

Figure 4

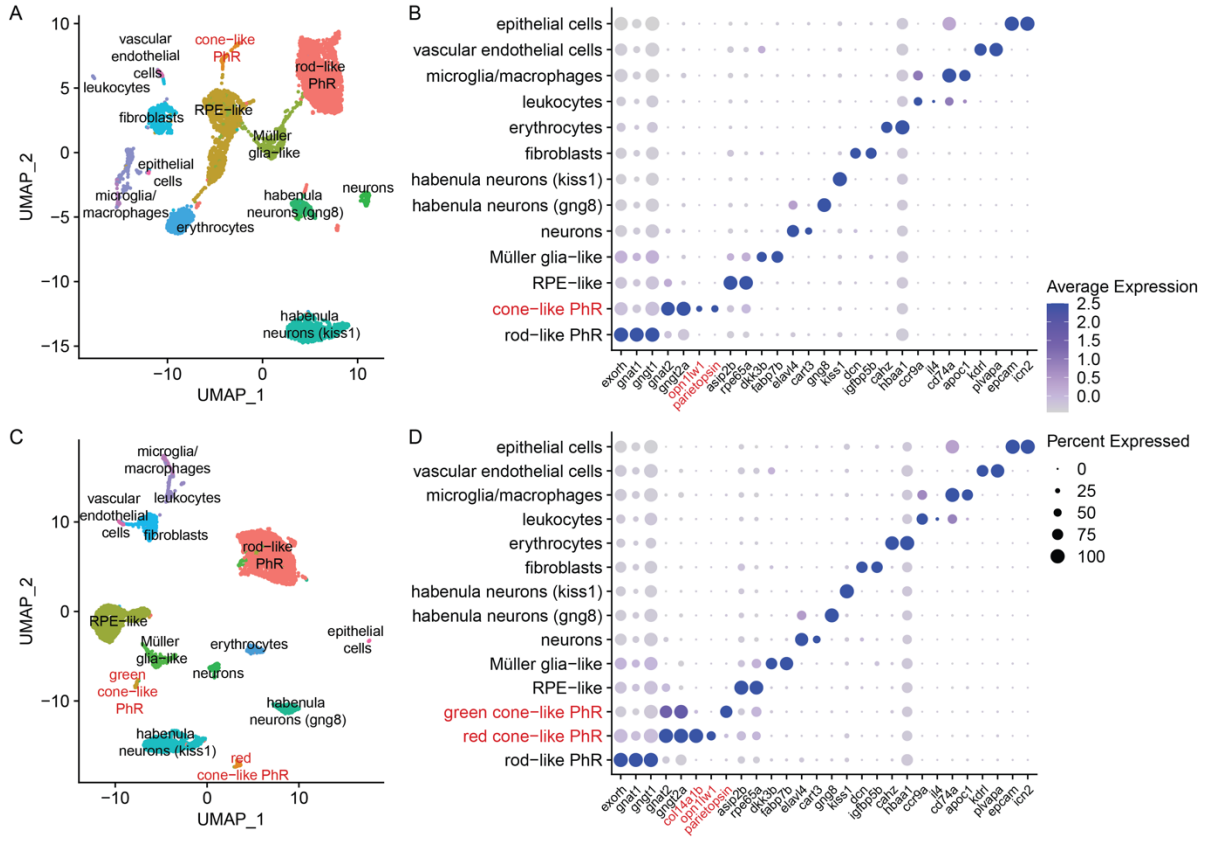

Fig. 4 Downstream analysis of the dr_pineal_s2 dataset. **a** 2D visualization using UMAP of Cell Ranger analyzed clusters. Each point represents a single cell, colored according to cell type. The cells were clustered into 13 distinct types, which were defined according to their unique transcriptomes (Additional file 2: Table S1). **b** Expression profile of marker genes according to cluster (Shainer et al., 2019) of (**a**). Cone- and rod-like PhRs are defined according to their transducin protein subunits (*gnat1* and *gngt1* in rods vs. *gnat2* and *gngt2a* in cones (Lagman et al., 2015)), as well as their unique opsins (*exorh* in rod-like (La Manno et al., 2018) and *opn1lw1* and *parietopsin* in cone-like). **c** 2D visualization using UMAP of kallisto analyzed clusters. Each point represents a single cell, colored according to cell type. The cells were clustered into 14 distinct types, which were defined according to their unique transcriptomes (Additional file 2: Table S1). **d** Expression profile of marker genes according

to cluster (Shainer et al., 2019) of (**c**). Two different populations of cone-like PhRs are detected. Both express the cone unique transducin protein-subunits, but differ in the expressed opsin (*opn1lw1* in red-cones, and *parietopsin* in green-cones), demonstrating the bi-chromatic photoreception characteristic of the pineal gland. *col14a1b* was only detected in the kallisto dataset and is the strongest DE marker within the red-cone cluster (**c, d**).

References

Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527.

Cao, C., Lemaire, L. A., Wang, W., Yoon, P. H., Choi, Y. A., Parsons, L. R., Matese, J. C., Wang, W., Levine, M. and Chen, K. (2019). Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* **571**, 349–354.

Cau, E., Ronsin, B., Bessière, L. and Blader, P. (2019). A Notch-mediated, temporal asymmetry in BMP pathway activation promotes photoreceptor subtype diversification. *PLOS Biol.* **17**, e2006250.

Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, Ł., Aibar, S., Makhzami, S., Christiaens, V., Bravo González-Blas, C., et al. (2018). A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell* **174**, 982-998.e20.

Deschamps-Francoeur, G., Simoneau, J. and Scott, M. S. (2020). Handling multi-mapped reads in RNA-seq. *Comput. Struct. Biotechnol. J.* **18**, 1569–1576.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21.

Du, Y., Huang, Q., Arisdakessian, C. and Garmire, L. X. (2020). Evaluation of STAR and Kallisto on Single Cell RNA-Seq Data Alignment. *G3: Genes|Genomes|Genetics* **10**, g3.401160.2020.

Ekström, P. and Meissl, H. (1997). The pineal organ of teleost fishes. *Rev. Fish Biol. Fish.* **7**, 199–284.

Falcón, J., Besseau, L. and Boeuf, G. (2006). Molecular and Cellular Regulation of Pineal Organ Responses. In *Fish Physiology*, pp. 243–306.

Glasauer, S. M. K. and Neuhauss, S. C. F. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. Genomics* **289**, 1045–1060.

Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. and Marioni, J. C. (2018). Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 2667.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C. and Teichmann, S. A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 1–15.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201.

Kölsch, Y., Hahn, J., Sappington, A., Stemmer, M., António, M., Helmbrecht, T. O., Lele, S., Butrus, S. and Laurell, E. (2020). Molecular classification of zebrafish retinal ganglion cells links genes to cell types to behavior.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber,

K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. *Nature* 206052.

Lagman, D., Callado-Pérez, A., Franzén, I. E., Larhammar, D. and Abalo, X. M. (2015). Transducin Duplicates in the Zebrafish Retina and Pineal Complex: Differential Specialisation after the Teleost Tetraploidisation. *PLoS One* **10**, e0121330.

Lawson, N. D., Li, R., Shin, M., Grosse, A., Yukselen, O., Stone, O. A., Kucukural, A. and Zhu, L. (2020). An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes. *Elife* **9**, 1–76.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323.

Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T. and Marioni, J. C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214.

Melsted, P., Ntranos, V. and Pachter, L. (2019). The barcode, UMI, set format and BUStools. *Bioinformatics* 1–2.

Packer, J. S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., et al. (2019). A lineage-resolved molecular atlas of C. Elegans embryogenesis at single-cell resolution. *Science (80-.).* **365**,.

Pandey, S., Shekhar, K., Regev, A. and Schier, A. F. (2018). Comprehensive Identification and Spatial Mapping of Habenular Neuronal Types Using Single-Cell RNA-Seq. *Curr. Biol.* 1–14.

Patro, R., Mount, S. M. and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419.

Peuß, R., Box, A. C., Chen, S., Wang, Y., Tsuchiya, D., Persons, J. L., Kenzior, A., Maldonado, E., Krishnan, J., Scharsack, J. P., et al. (2020). Adaptation to low parasite abundance affects immune investment and immunopathological responses of cavefish. *Nat. Ecol. Evol.* **4**, 1416–1430.

Shainer, I., Michel, M., Marquart, G. D., Bhandiwad, A. A., Zmora, N., Ben-Moshe Livne, Z., Zohar, Y., Hazak, A., Mazon, Y., Förster, D., et al. (2019). Agouti-Related Protein 2 Is a New Player in the Teleost Stress Response System. *Curr. Biol.* **29**, 2009-2019.e7.

Svensson, V., Vento-Tormo, R. and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604.

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. and Hellmann, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* **10**, 4667.

Wang, W., Hu, C.-K., Zeng, A., Alegre, D., Hu, D., Gotting, K., Ortega Granillo, A., Wang, Y., Robb, S., Schnittker, R., et al. (2020). Changes in regeneration-responsive enhancers shape regenerative capacities in vertebrates. *Science (80-.).* **369**, eaaz3090.

Wu, T. D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**,.