Single cell transcriptome analysis during development in *Dictyostelium*

Vlatka Antolović and Jonathan R. Chubb

UCL Laboratory for Molecular Cell Biology
University College London
Gower Street
London
WC1E 6BT

Lead contact: v.antolovic@ucl.ac.uk

Keywords: single cell transcriptomics, RNA sequencing, transcription bursting, development, bioinformatics, big data, stochastic gene expression

Abstract

Dictyostelium represents a stripped-down model for understanding how cells make decisions during development. The complete life cycle takes around a day and the fully differentiated structure is composed of only two major cell types. With this apparent reduction in "complexity", single cell transcriptomics has proven to be a valuable tool in defining the features of developmental transitions and cell fate separation events, even providing causal information on how mechanisms of gene expression can feed into cell decision making. These scientific outputs have been strongly facilitated by the ease of non-disruptive single cell isolation- allowing access to more physiological measures of transcript levels. In addition, the limited number of cell states during development allows the use of more straightforward analysis tools for handling the ensuing large datasets, which provides enhanced confidence in inferences made from the data. In this chapter we will outline the approaches we have used for handling *Dictyostelium* single cell transcriptomic data, illustrating how these approaches have contributed to our understanding of cell decision making during development.

1. INTRODUCTION

Dictyostelium is an exceptional system for application of single cell transcriptomic analyses to developmental biology. The developmental programme is rapid, with only two major final cell fates, allowing high temporal resolution analysis of cell state transitions and the emergence of cell type specialisation with relatively low cell numbers. In addition, the ease and speed of extraction of high-quality RNA permits very reproducible data, removing the need for complex batch correction procedures and allowing deep insights into how the mechanisms of gene regulation govern cell type heterogeneity. The data from this system are a clear benchmark for high quality single cell transcriptome data, used for the development of reliable quantitative tools for more complex and intractable systems [1].

In this chapter, we will outline the essential approaches for analysis of single cell transcriptome data in *Dictyostelium*, with specific technical and biological examples showing the methods we arrived at to best exploit the data in addition to the gains in knowledge about developmental processes that emerged.

2. MATERIALS

a) Cell preparation.

i) Although cells can be cultured on bacteria, HL5 growth media is used for most experiments (Formedium HLG0102), with cells grown in adherent culture on 10 cm tissue culture dishes.

ii) Development buffer. Any standard lab development buffer could be used, but we use KK2: 20mM KPO_4, pH6.0-6.2.

iii) EDTA. We use a 0.5M stock (pH 8.0) of EDTA, added to the KK2 to a final concentration of 10-20 mM. In practice, we observed little difference in the efficacy of cell disaggregation between 10 and 20 mM.

iv) Syringes, needles. For disaggregation of cells we pass the sample, in KK2, through a 1ml syringe (BD 303172) attached to a 20g needle (BD407).

v) Development substrate. We use 1.5- 2% agar (BD214010) dissolved in KK2 for most applications. This agar is cast in 3.5 cm petri dishes. For more uniform late development, we plate cells on Whatman #50 filter paper rings (Whatman 1450-090). These rings are kept humid by layering them on rings of Whatman #3 filters (1003-090) saturated with KK2 in 10 cm petri dishes.

b) Cell isolation, library preparation and sequencing

Two standard strategies have been used for cell isolation: either using the 10x Genomics Chromium Single Cell Gene Expression platform or Fluidigm (now Standard BioTools Inc.) C1 single-cell mRNA sequencing system. Please refer to company websites for more information. Sequencing has been carried out on a NextSeq500 (Illumina). A detailed description of this standard protocol is beyond the scope of this chapter. Please visit the Illumina website.

c) Data analysis

i) Checking the quality of the sequenced reads: FASTQC (Babraham Bioinformatics).

ii) Mapping the reads to reference genome and quantifying reads per gene: we used Tophat2 and HtSeq for reads retrieved using C1 sequencing system. Cell Ranger (10X Genomics) was used for reads retrieved using 10x Chromium.

iii) Analysis of read count tables (cell transcriptome profiles): MATLAB (MathWorks), Mathematica (Wolfram) and R (R Foundation).

iv) Gene ontology analysis: Panther Classification System (National Science Foundation).

3. METHODS

3.1. Data generation considerations

For completeness, we briefly review the approaches that take us from the biological sample to the sequencing reads. For more details, please visit our recent experimental papers [2-5]. Readers may also wish to consider other studies in which we reused the data based on new scientific questions [6, 7] for insight into the general approaches used.

Single cell isolation in undifferentiated and early developing cells is based on gentle pipetting of cells off their substrates using development buffer. Cells are plated either on 35 mm dishes containing 2-3 ml of 1.5-2% agar made up in KK2 buffer, or developed on Whatman #50 filter paper. Generally, 5 x10^6 -10^7 cells are plated, from a 1 ml cell suspension in KK2 buffer. Cells are allowed to settle for 10-30 minutes, before buffer is removed, then the samples are kept in a humidified chamber. At specific time points after this, differentiating cells are removed from the agar surface by gentle pipetting up and down with ice cold 1 ml KK2 with a standard micropipette (such as a Gilson P1000). Later developmental stages (after 6 hours, when cells have formed aggregates) where cells are more adherent, requires a slightly more aggressive disaggregation, in which cells are repeat pipetted (20 times) through a syringe and 20G needle. For these later stages, the KK2 also contains 10-20 mM EDTA to inhibit cell-cell adhesions. This is a standard approach for studying more developmentally advanced cells, and the dissociated cells can be used for standard physiological assays after this treatment [4, 6]. The resultant cell suspension can be directly used for transcript extraction, but in practice is kept on ice during the transit to the sequencing facility. Gentle pipetting (with Gilson or syringe/needle as appropriate) just prior to loading the suspension ensures single cells are injected into the cell isolation equipment. The absence of a clear stress response in our data, the general concordance of our data with standard population transcriptomics datasets [5] and the overall high predictability of the gene expression behaviours of specific candidate genes arising from our analysis (eg. [3]), suggest this cold disaggregation approach is minimally invasive, at least as far as the transcriptome is concerned.

The commercial pipelines used for going from single cell suspensions to single cell reads are the 10x Genomics Chromium Single Cell Gene Expression platform or Fluidigm (now Standard BioTools Inc.) C1 single-cell mRNA sequencing system. The C1 platform benefitted from high sequencing depth but lower cell counts (around 30-60 cells per run was usual although the theoretical maximum was 96). The 10x platform has been used more recently, which has lower read depth, but many more cells (in our experiments, usually up to around 5000 per run, however, biological replicates can be loaded at the same time, which greatly facilitates quality control). In addition, the 10x platform can take the cells straight from ice cold buffer to the chemistry, whereas the C1 required visual inspection of cell counts/well at room temperature, which might permit stress responses to develop- although in our experience, these were not pronounced- with obvious exception of pre-existing spontaneous stress on cells, such as DNA damage [6]. As the RNA conversion to read counts requires amplification steps, final read counts are not linearly related to the initial single cell content of each transcript. However, the option for exact transcript quantification using unique short sequences (unique molecular identifiers: UMIs) can be applied to both C1 and 10x approaches.

3.2. Generating read count tables

The focus of this section is the analysis of the read count data (number of transcripts belonging to each gene in each cell). For this reason, we provide a short overview of what precedes generating a read count table after receiving sequencing read files from the sequencing facility.

Raw sequenced reads are, by standard, stored as FASTQ files. The quality of the FASTQ files should be checked (with FASTQC tool) and then the reads that pass the quality threshold are aligned to the reference genome (in a FASTA file format, together with its correct annotation in GFF format). Based on these alignments (SAM or BAM file format), each transcript for each cell is counted accordingly. Depending on the sequencing technology used, the examples presented in this chapter used data processed in one of the following ways: reads retrieved with the 10x system were processed with Cell Ranger analysis pipelines, while reads from the C1 system were processed with tuxedo suite protocols (previously the Bowtie and Cufflinks packages, now HISAT2 and StringTie tools).

3.3. Data preprocessing

3.3.1 Basic data formatting

After reads have been mapped and counted, the data is in the form of a count table usually in a format of a genes as rows and cells as columns. Depending on the number of samples and the analysis pipeline we used, some basic formatting steps were carried out first, such as merging replicates or naming cells in an appropriate way (e.g., “rep2_cell1” for a first cell from a second replicate, instead of a barcode “AAACGAACATTGACCA-1”).

3.3.2 Removing outlier samples

We show a simple example where we excluded samples unlikely to be single, whole, properly lysed cells. These were removed to increase the signal to noise ratio in later analysis. We excluded these outliers based on the following properties:

1. Low total number of mapped genes per cell. A low number of mapped genes per cell may indicate cell debris or an ineffectively lysed cell. The best way to decide on the low

threshold value is to visualise the data. The example in Figure 1 is from [5]. We would expect the unfragmented, well lysed cells are less different from each other than from debris. A trough in the plot showing the occurrence of mapped genes per cell (Figure 1A), indicates that the first peak (with few mapped genes) most likely comes from poor samples.

2. Low total number of transcripts. Extremely low numbers of transcripts also indicate cell debris, an ineffectively lysed cell, or the failure in RNA capture step. Again, the best way to decide on the threshold value is to visualise the data (Figure 1B). A lot of these outliers will overlap with those detected in the previous step (Figure 1A).
3. High total number of transcripts. Unexpectedly high numbers of transcripts could mean more than one cell was captured. The best way to detect outliers here is, again, to visualise the data. If there is no obvious threshold value, a Tukey Fences for outliers could be used, where values above the sum of the upper quartile (Q3) and 1.5- or 3-times the interquartile range (IQR) would be considered as outliers or far outliers, respectively.

An example of the outliers detected by the previously mentioned steps is shown in Figure 1C. There are other ways of detecting suspicious samples (e.g., high percentage of mitochondrial transcripts, as used in [2]). These choices will depend on the biology of the sample- for example if it is suspected that there may be rare populations of cells of interest, it would be a good idea to be less stringent in outlier removal.

3.3.3 Normalisation

As the sequencing depth always varies between the samples, there is a need for appropriate normalisation. Depending on the technology used, there might also be a need to normalise for gene length, as in the techniques not using unique barcoding (UMIs) there is a bias towards longer genes being assigned more counts. The initial standard methods used for normalisation were RPM (reads per million mapped reads), where all the gene counts in the cell are divided by the same value (the so called "size factor") and RPKM (reads per kilobase per million mapped reads), where each gene in the cell is divided by a gene specific value. Both these techniques are still used today. The drawback of these methods is that total read counts in cells are often dominated by a few highly expressed genes (e.g., actin), and this can distort the inferred relative level of expression of weakly expressed genes [8].

For deep sequenced samples acquired using the C1 approach [2, 3], we found that using the size factor from 'DESeq' package in R [8] gives a good result. This method normalises single cell read counts based on the geometrical mean over all transcripts (in the case of *Dictyostelium* this means a geometrical mean of up to 14000 different transcript counts). For the more shallow sequencing sample acquired using the 10x method used in [5], where over 4000 cells were sequenced, we used size factors calculated with 'scran' package in R [9]. This method was developed alongside the advancement of single-cell RNA sequencing (scRNA-seq), and assumes a large and heterogenous population. It pre-clusters the cells based on their expression profile, calculating each cluster's size factor value, before deconvolving each individual cell's size factor. For general reviews on normalisation methods, we suggest [10] and [11].

At the end of the analysis, one should always check whether the final observations are also valid in the unprocessed data and not an artefact introduced by preprocessing (see Figure 8 in the Notes section).

3.3.4 Gene expression variability

A major advantage of the development of scRNA-seq is the ability to study the distribution of gene expression in a population of cells and how this leads to the emergence of different cell types. To explore the emergence of cellular heterogeneity during development, we analysed individual transcriptomes of 433 cells from three different development timepoints [2]. Cells were collected at 0h (undifferentiated cells), 3h (starvation) and 6h (aggregation), with a median of 3 million reads and 5600 mapped genes per cell. A typical way of showing gene expression variability in a cell population is plotting square coefficient of variation (CV^2) for each gene against its mean expression level (Figure 2A). The variability in gene expression is negatively correlated with its mean expression level. We define a variable gene as a gene more variable in its expression than expected for a gene with that level of expression. The 'expected' variance is described with a running median of CV^2 values and is depicted with a red curve in Figure 2A. In this example, genes previously known to be variably expressed, contact site A (*csaA*) and discoidin I (*dscA*), and a gene known as homogeneously expressed, actin 5 (*act5*), are plotted on the same figure. We found that overall variability in the cell population increases over time (Figure 2B) and that this increase affects genes at all levels of expression (Figure 2C). This heterogeneity increased without clear separation of the population into different states [2].

Since many genes are downregulated during these initial stages of the development [2], we decided to examine how downregulation and upregulation of transcript levels affect transcript variability. This analysis used the distance from median (DM, [12]) as the measure of transcript variability. The first step for calculating DM is defining each transcript's variance as a distance between their CV^2 value and the value expected for the gene of such level of expression, as visualized by the running median in Figure 2A. If read counts were extracted without UMIs, there is need for another step, which also corrects for the effect of gene length [12]. As a result, we get a transcript variance measure that is independent of gene's length or its expression level (Figure 3A). The DM method is now also a part of the ‘scran’ package in R [9].

Comparing the variance between genes being up- and downregulated from undifferentiated cells (0h) to the start of aggregation (6h), we found that downregulated transcripts become more variable than upregulated transcripts. This effect is consistent across all levels of expression (Figure 3A). In addition, the higher the fold-change expression threshold we used, the difference became even more apparent (Figure 9 in the Notes section). We then tested whether the difference in transcript stability of up- and downregulated genes could be causing this effect on variability, by using the RNA turnover kinetic data from [13] and found no correlation between RNA turnover and gene expression variability (Figure 3B). We then considered how regulation of transcription could influence the variability in transcript abundance between cells. We simulated transcription with a simple two-state model, where a gene is either in OFF or ON state and switches between these states with the rates of k_{on} and k_{off}, respectively (Figure 3C). This results in bursts of transcription, described by the burst size and the frequency of burst occurrence. When in the ON state, the polymerase initiation event happens at the rate λ, and the transcripts have a degradation time τ. We ran the simulations using physiological parameter estimates [13-16]. This suggested that gene expression if regulated primarily by changing the burst frequency, then transcription downregulation will result in a more variable distribution of the transcript abundance in the cell population (Figure 3D), as observed in the data. This is best visualised in Figure 3E: reduction of burst frequency results in a lower mean expression but a higher variance. In the opposite direction, increase in frequency increases the expression, but lowers the variance. Initially, we looked at the difference of variance at what we considered the end point – 6h of differentiation. But from the point of the model, there is no reason why one end of the process would be considered the beginning and the other the end. Following that logic, we would expect exactly the opposite characteristics of up- and downregulated genes at 0h: genes to be developmentally upregulated would be more variable to start with, while the genes to be downregulated would be more homogeneous. This is exactly what we observe at the onset of aggregation (Figure 3F). The summary of this findings is presented in Figure 3G.

3.3.5 Linear Dimensionality Reduction

As each *Dictyostelium* cell is described with around 14000 genes, it is impossible to visualise the data in its original form – each cell would be a point in the 14000-dimensional space. In order to visualise, investigate and describe the data we need to use techniques of dimensionality reduction.

To be able to directly interpret the data in this reduced space, the first step is usually Principal Component Analysis (PCA), a linear dimensionality reduction method. To remove experimental noise, we only used the genes with mean expression above the certain threshold. The choice of this threshold was based on the convention at the time of the study: we used a value where the coefficient of variation stops reaching a maximal value. For example in Figure 2A the mean expression threshold value was 10 read counts. Another standard approach (used for example in the Cell Atlas projects [17]) is to use only the most variable genes, selecting all the genes more variable than average or only the 10% most variable genes (see Figure 10 in Notes).

A useful example considers the mound phase of development [3]. For this phase, we captured 116 single cell transcriptomes with around 3 million reads and 5400 mapped genes per cell. PCA linearly transforms the data into new components, sorted in the way that the first component will capture most of the variance in the set, with each further component capturing less and less variance. The top left panel in Figure 4A shows the cells positioned based on their transcriptome profiles in principal component 1 (PC1) and PC2 space. To interpret this plot, we use loadings analysis. As each component is a linear combination of original dimensions (i.e., genes), we can interpret each component based on the contributions of genes to its total variance. The higher the loading's absolute value an individual gene has in a principal component, the larger its contribution is.

We carried out the loadings analysis in the following way. For each component, we sorted the genes based on their loadings, and took the top genes whose contribution summed to the 25% of the component's variance. We then used these genes for gene ontology enrichment analysis, using the PANTHER Classification System on the Gene Ontology web site [18]. The resulting enrichments in biological processes are shown in the tables below the PCA panel for the PC1, and right of the panel for the PC2 in Figure 4A. It is apparent from these results that PC1 and PC2 capture different aspects of development: PC1 captures translation being downregulated and development initiated, while PC2 captures cytokinesis, signalling, adhesion, and cAMP chemotaxis. If we overlay the known biological information on this plot, such as expression of developmentally upregulated and downregulated genes (Figure 4B left and right panel, respectively; [19]) we can see the direction of development going from left to right (across PC1), with the tilt towards the positive PC2 value. Looking at the PC3, while keeping PC1 as a proxy for development, we can observe the separation of cells in two separate clusters at the high values of PC1 (Figure 4C, left panel). Loading analysis of PC3 gives easy interpretable information for the top cluster (spore wall assembly), but a more complex one for the bottom cluster (right panel in Figure 4C). We then used the same set of genes to look at their enrichment in prestalk or prespore cells, using the information from [20] on the dictyExpress platform [21]. The result of this analysis is shown in Figure 4D. PC3 captures the cell fate information, and we observe the obvious separation of prespore and prestalk cells across PC3. Overlaying the expression of known cell fate markers, *pspA* for spore and *ecmA* for stalk cells, illustrates this inference (Figure 4E).

Following the same principle, we identified another cluster of cells along the PC4 axis [3]. This population appears as an intermediate between the early and differentiated cells, with a slight enrichment in prestalk markers. All identified clusters are shown in Figure 5A.

3.3.6 Hierarchical Clustering

A simple and useful approach to explore a single cell transcriptomics dataset is two-way hierarchical clustering, where cells and genes are simultaneously ordered and clustered in an unbiased way based on their similarities (Figure 5B). For this analysis we selected genes based on the following conditions: their mean normalised read count was >10 and are correlated with at least 10 other genes with Pearson's correlation >0.5. Over 5000 genes passed this criterion (almost half of the genome), which is a huge advantage of the C1 deep sequencing methods. With the same criteria for data collected using 10x [5] we retrieved 957 genes.

The result of the hierarchical clustering is shown in Figure 5B. Expression level of each gene is normalised, so each gene will have a mean of zero and a standard deviation of one. The colours in the heatmap represents the z-score for each gene across the cells, which reflect the up or down regulation of the specific gene in the specific cell. The coloured boxes next to the cells' hierarchical tree mark cell clusters as in Figure 5A. We observe a sudden shift in the transcriptome profiles as cells go through development within the initial multicellular structure (mound), with 80% of the genes being downregulated. Two intermediate populations are also clearly visible and seem to be captured at the moment of transition, with the 'early' genes not yet being downregulated and 'late' genes already showing activation (cell clusters marked with 1 and 2). Within the genes upregulated during this transition, there are clear clusters of cell fate specific genes (marked with 3 and 4 for spore and stalk markers, respectively).

Overall, this analysis revealed key features of a developmental transition. Firstly, cells occupy discrete states during developmental progression. Secondly, these analyses confirm the existence of intermediates states that have been previously inferred from single gene analysis [22], showing that these states show mixed early and late gene expression, in addition to signatures of the ultimate fate. Thirdly, the major gene expression transition in the mound stage is dominated by repression, not activation, which contrasts the standard developmental narrative of cell differentiation being determined by what genes are "on" (rather than what genes are "off").

3.3.7 Nonlinear Dimensionality Reduction

More recently, we used the 10x method to extract individual transcriptomes from 4743 cells [5]. The sequencing depth here is much lower, with around 35000 reads and 2300 mapped genes per cell. Although shallow sequencing is less useful for extracting detailed mechanistic information on gene expression, it enabled us to profile 10 times more cells than in the previously mentioned experiments, which has provided very clear data on the transitions occurring from the undifferentiated state through to the mound [5].

Cells were captured while feeding on bacteria on agar plate, by taking a continuous streak of cells from the edge of the multicellular zone and into the layers of bacteria. This way we captured the cells in a mimic of their physiological environment, in all moments of development up to the tight mound stage, where differentiation of spore and stalk cells begins within the multicellular structure. This enabled us to capture a continuous time series of development in a very complex signalling environment, rather than the more standard timepoint sampled experiments where all cells are synchronously starved in homogeneous environment at uniform density. The early developmental process captured is shown in Figure 6A.

4743 cells described with 9698 genes (mean expression > 0.01) are plotted in the space of PC1 and 2. Overlaying the biological knowledge we have on the system (expression of genes active before the development begins, during the cell aggregation and after the cell aggregate), we conclude that again, PC1 represents the axis of developmental time - the direction of development captured in this plot is schematically shown with the overlayed arrow in the last panel (Figure 6B). Distribution of cell density across these two dimensions is shown in Figure 6C. Using the higher order principal components, a sudden change in transcriptome profiles becomes apparent which appears to correlate with the start of aggregation (Figure 6D). PC5 captures the separation of the population towards two different cell fates at the end of development (Figure 6E).

Although we are dealing with a relatively “simple” model organism, the information we are interested in spreads across at least five dimensions. To reduce dimensionality even further, we use nonlinear dimensionality reduction techniques. Many of these methods are often used in singe cell transcriptomes analysis (e.g., t-SNE, UMAP, diffusion maps etc.). We decided on elastic embedding (EE), which seems to preserve both local and global structures of the data [23]. For Figure 7, we used EE on the first 11 PCs from the example in Figure 6. Here, the visualization reveals more direct information: as cells transition from individual cells to streams (from North-West to South-East), there is a sudden substantial shift in the transcriptome profiles (the ‘jump’). Secondly, during tight mound formation, cell transcriptomes seem to converge to a similar expression profile (the ‘bottleneck’), before separating towards two different fates, stalks and spores (Figure 7A,B). The developmental context of the jump is validated by overlaying expression of gene sets marking undifferentiated cells and streams from a separate study [24] (Figure 7C), which also shows the sudden shift. The cell density plot in Figure 7D shows the distribution of cells across these two dimensions and shows a considerable heterogeneity just before the jump, with only a few cells caught in the moment of transition and an accumulation of cells at the bottleneck, before separating in two different cell types. Using again the data from [24], this bottleneck corresponds to the loose to tight mound transition (Figure 7E).

3.3.8 Trajectory inference

There are many trajectory inference methods developed over the last decade, aiming to determine the exact/approximate order and direction of differentiation across the transcriptome space, and to predict the trajectories by which individual cells might explore the possible cell states. We tried several of these, but often found them inadequate for tracking Dictyostelium development for different reasons. For instance, using the scRNA-seq data from onset of development up to aggregation stage [2] and mound stage data [3] we found both Monocle [25] and Wishbone [26] mapped 3h starved cells branching into two clusters: aggregating and mound cells. When trying to infer the developmental progression trajectory on the continuously sampled development up to the mound stage [5] with DensityPath [23] and StemID [27], the heterogeneous region before the 'jump' was interpreted as containing at least three clusters: two dead ends and only one continuing across the jump into the development that skips aggregation phase as another dead-end cluster. These observations seem unrealistic based on several decades of literature on *Dictyostelium* development.

For trajectory analysis, we also used RNA velocity [28], an algorithm which uses the information on spliced versus unspliced introns to predict the future state of each cell. We were concerned that Dictyostelium genes have too few introns for the algorithm to work, with other considerations that the introns are highly AT rich and therefore unlikely to be mapped effectively and also that RNA processing may be too efficient to leave substantial intron signatures in the data. Indeed, our analysis revealed very few introns in our reads, and these appeared limited to highly expressed genes with long introns, such as ribosome protein genes. However, we nevertheless attempted to apply RNA velocity to our data. Our initial analysis suggested cells in the early mound state "circulate" within this state, rather than progress forward. However, our velocity fields were weak and unconvincing. The initial data may warrant further study, perhaps using gene expression imaging [14] or more recent transcriptome inference methods [29]. We also suggest [30] for a detailed overview of different pseudotime methods. The possibilities for DNA barcode-based lineage tracing [31] may be limited for following cell histories during *Dictyostelium* development. Cells divide no more than once during development [32, 33], so the likelihood of sampling cells from the same division (to enable the construction of lineage trees) using the sequencing technologies we have used to date is expected to be small.

4. NOTES

In this section we briefly reflect on some key examples of how our exploratory data analysis works. As mentioned above, it is good practice to always check whether the final observations are also valid in the unprocessed data or in the data processed in a slightly different manner. For instance, normalisation is used to dampen the noise in a data. Different normalisation methods could make a less clear signal but should not give a completely different conclusion.

Example 1 – comparing normalisation approaches

In Figure 8 we show an example from [5]. The top panels show EE plots of cells' transcriptome profiles, while the bottom plots show distribution of cell density. Figure 8A shows data normalised by pre-clustering and deconvolution method ('scran' package in R, [9]. In Figure 8B the data were not normalized, but only variable genes were included to reduce the noise. Figure 8C shows data normalised by the geometrical mean size factor from (DESeq2 package in R, [8]. The major observations are consistent in all three examples: high heterogeneity just before the jump, the jump, the decrease of heterogeneity in the bottleneck, and the separation of cell fates immediately after.

Example 2 – binning, threshold, and variance

In the next example (Figure 9), we will refer to the example from gene expression variability study ([2] and Figures 2 and 3 above). As in Figure 3, the DM values for all genes are shown in grey. The transcript variability of genes up- and downregulated during development are marked in black and purple, respectively. The top left panel shows the variance distribution for all genes with at least two-fold change in expression, in any direction ($|\log_2 FC| > 1$). If we increase this threshold to at least 10-fold (right panel), the difference in variance becomes even more apparent. In both these analyses we compared the genes of the similar expression level by binning so that each bin contains 500 genes. Changing bin size (bottom left panel), the overall difference between up and down regulated genes is still clear. Finally, the bottom right panel shows that overall conclusion is also retained if we use CV^2 values of genes rather than DM.

Example 3 – deciding which genes to include in the analysis

This example is linked with example in Figure 6, with PCA using 9698 genes (criteria being the mean expression > 0.01; [5]). We can additionally simplify the data by selecting only variable genes, which is actually the most common trend in single cell transcriptomics analysis, notably in the studies focused on building cell atlases [17].

In Figure 10A genes are plotted based on their variance (CV^2) and mean expression. The vertical line marks the gene expression threshold (>0.01). The genes marked yellow, black, and red are the top 50% variable genes. The top 25% most variable genes are marked black and red, while top 10% are marked only in red. Figures 10C-D show the result of PCA on the same sample as in Figure 6 but using only 50%, 25% or 10% most variable genes. We can see that the sudden transcriptome shift is already visible in PC1/PC2 space, when using only variable genes. In addition, fate specification already appears in PC3, rather than PC5 when using only top 10% variable genes. Finally, more variance is explained by single PCs: the variance captured by PC1 goes from 8% to 20%, when reducing the set of genes from 9698 genes with mean expression > 0.01 to the 969 top 10% variable genes.

REFERENCES

1. Eling, N., et al., *Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data.* Cell Syst, 2018. **7**(3): p. 284-294 e12.
2. Antolovic, V., et al., *Generation of Single-Cell Transcript Variability by Repression.* Curr Biol, 2017. **27**(12): p. 1811-1817 e3.
3. Antolovic, V., et al., *Transition state dynamics during a stochastic fate choice.* Development, 2019. **146**(12).
4. Nichols, J.M., et al., *Cell and molecular transitions during efficient dedifferentiation.* Elife, 2020. **9**.
5. Westbrook, E.R., et al., *Collective signalling drives rapid jumping between cell states.* bioRxiv, 2023: p. 2023.05.03.539233.
6. Miermont, A., et al., *The fate of cells undergoing spontaneous DNA damage during development.* Development, 2019. **146**(12).
7. Tunnacliffe, E., A.M. Corrigan, and J.R. Chubb, *Promoter-mediated diversification of transcriptional bursting dynamics following gene duplication.* Proc Natl Acad Sci U S A, 2018. **115**(33): p. 8364-8369.
8. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
9. Lun, A.T., K. Bach, and J.C. Marioni, *Pooling across cells to normalize single-cell RNA sequencing data with many zero counts.* Genome Biol, 2016. **17**: p. 75.
10. Vallejos, C.A., et al., *Normalizing single-cell RNA sequencing data: challenges and opportunities.* Nat Methods, 2017. **14**(6): p. 565-571.
11. Lytal, N., D. Ran, and L. An, *Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey.* Front Genet, 2020. **11**: p. 41.
12. Kolodziejczyk, A.A., et al., *The technology and biology of single-cell RNA sequencing.* Mol Cell, 2015. **58**(4): p. 610-20.
13. Muramoto, T., et al., *Live imaging of nascent RNA dynamics reveals distinct types of transcriptional pulse regulation.* Proc Natl Acad Sci U S A, 2012. **109**(19): p. 7350-5.
14. Chubb, J.R., et al., *Developmental timing in Dictyostelium is regulated by the Set1 histone methyltransferase.* Dev Biol, 2006. **292**(2): p. 519-32.
15. Stevense, M., et al., *Digital nature of the immediate-early transcriptional response.* Development, 2010. **137**(4): p. 579-84.
16. Corrigan, A.M., et al., *A continuum model of transcriptional bursting.* eLife, 2016. **5**.
17. Hu, B.C., *The human body at cellular resolution: the NIH Human Biomolecular Atlas Program.* Nature, 2019. **574**(7777): p. 187-192.
18. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
19. Rosengarten, R.D., et al., *Leaps and lulls in the developmental transcriptome of Dictyostelium discoideum.* BMC Genomics, 2015. **16**(1): p. 294.
20. Parikh, A., et al., *Conserved developmental transcriptomes in evolutionarily divergent species.* Genome Biol, 2010. **11**(3): p. R35.
21. Stajdohar, M., et al., *dictyExpress: a web-based platform for sequence data management and analytics in Dictyostelium and beyond.* BMC Bioinformatics, 2017. **18**(1): p. 291.
22. Clay, J.L., R.R. Ammann, and R.H. Gomer, *Initial cell-type choice in a simple eukaryote: cell-autonomous or morphogen-gradient dependent?* Dev Biol, 1995. **172**(2): p. 665-74.
23. Chen, Z., et al., *DensityPath: an algorithm to visualize and reconstruct cell state-transition path on density landscape for single-cell RNA sequencing data.* Bioinformatics, 2019. **35**(15): p. 2593-2601.
24. Katoh-Kurasawa, M., et al., *Transcriptional milestones in Dictyostelium development.* Genome Res, 2021. **31**(8): p. 1498-1511.

25. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nat Biotechnol, 2014. **32**(4): p. 381-386.
26. Setty, M., et al., *Wishbone identifies bifurcating developmental trajectories from single-cell data.* Nat Biotechnol, 2016. **34**(6): p. 637-45.
27. Grun, D., et al., *De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data.* Cell Stem Cell, 2016. **19**(2): p. 266-277.
28. La Manno, G., et al., *RNA velocity of single cells.* Nature, 2018. **560**(7719): p. 494-498.
29. Maizels, R.J., D.M. Snell, and J. Briscoe, *Deep dynamical modelling of developmental trajectories with temporal transcriptomics.* bioRxiv, 2023: p. 2023.07.06.547989.
30. Saelens, W., et al., *A comparison of single-cell trajectory inference methods.* Nat Biotechnol, 2019. **37**(5): p. 547-554.
31. Weinreb, C., et al., *Lineage tracing on transcriptional landscapes links state to fate during differentiation.* Science, 2020. **367**(6479).
32. Zimmerman, W. and C.J. Weijer, *Analysis of cell cycle progression during the development of Dictyostelium and its relationship to differentiation.* Dev Biol, 1993. **160**(1): p. 178-85.
33. Muramoto, T. and J.R. Chubb, *Live imaging of the Dictyostelium cell cycle reveals widespread S phase during development, a G2 bias in spore differentiation and a premitotic checkpoint.* Development, 2008. **135**(9): p. 1647-57.

FIGURE LEGENDS

Figure 1. Removing outlier cells.

A) Distribution of the number of detected genes per cell. The vertical line marks the chosen threshold. All samples with total number of genes lower than threshold are removed from further analysis. **B)** Distribution of the total number of UMI counts per cell. The vertical line marks the chosen threshold. All samples with total UMI count lower than threshold are discarded from further analysis. **C)** Total UMI count per cell. Purple dots mark cells excluded based on their total UMI count (too high or too low). The threshold for high outliers was Q3 + 1.5 · IQR. Black dots mark cells excluded based on the low total number of genes detected per cell. Horizontal grey line marks the overall sample median.

Figure 2. Dynamics of gene expression heterogeneity during early development.

A) Transcript variability (expressed as CV^2) of each gene with mean expression >10 read counts plotted against the mean expression. The red line marks the running median. *csaA* and *dscA* are shown as examples of known variably expressed genes and *act5* as an example of a gene with more uniform expression in the population. **B)** Overall transcript heterogeneity within a cell population increases during early development. Distribution of $\log_{10}(CV^2)$ values of genes are shown for populations at 0, 3 and 6h of development. **C)** The increase in variance occurs across all levels of expression. The running medians are plotted for all three timepoints. Figure reproduced from [2].

Figure 3. Origins of gene expression heterogeneity.

A) Genes downregulated during development become more variable than upregulated genes. Transcript variability (expressed as DM) of each gene with mean expression >10 read counts in the population of aggregating cells (6h) is plotted against the mean expression. Distribution of DM values for each bin is shown in purple and black, for down- and upregulated genes, respectively. Each bin contains 500 genes. The threshold for fold change (FC) is 5-fold in both directions, i.e., $|\log_2 FC| > 2.32$. **B)** DM of each gene plotted against its RNA turnover. RNA turnover is measured as a ratio of the expression before and after 1h actinomycin D treatment [13]**.** **C)** Schematic of a two-state model of transcription. **D)** Results of a transcription simulation presented in the same format as A. **E)** Schematic of the influence of changing burst size and burst frequency on mean expression and variance. **F)** Genes upregulated during development are more variable at the onset of starvation than downregulated genes. **G)** Schematic of the findings. Figure reproduced from [2].

Figure 4. Extracting biological information from PCA.

A) Cell positioned in PC1 and PC2 space based on their transcriptional profiles. Lower left panel: GO terms (biological processes) enriched in the negative loadings of PC1. Lower right panel: GO terms enriched in the positive loadings of PC1. Right panel: GO terms enriched in the positive loadings of PC2. **B)** Same plot as in A, but overlayed with the average expression of genes upregulated and downregulated during the mound stage. The arrow on the right shows inferred direction of developmental progression. **C)** Cells positioned in PC1 and PC3 space based on their transcriptional profiles. PC1 here serves as a proxy for developmental time. The orange, red and green ellipses mark early prespore, prespore and prestalk populations, respectively, based on observations in C, D and E. Upper right panel: GO terms enriched in the positive loadings of PC3. Lower right panel: GO terms enriched in the negative loadings of PC3. **D** and **E)** PC3 describes the separation of cell population in the spore and stalk cells in the later mound stages. **D)** Volcano plots of positive PC3 loadings (left panel) and negative PC3 loadings (right panel) show almost exclusive enrichment of spore and stalk markers, respectively. Log_2FC is a measure of fold change in prespore cells vs. prestalk cells; $-log_{10}FDR$ is a measure of false discovery rate. Plots downloaded from dictyExpress platform based on data from [20]. **E)** The same as plot in C, overlayed with the expression of *pspA* and *ecmA* as a spore and stalk marker, respectively. Figure partially reproduced from [3].

Figure 5. Visualising cell states.

A) Cells from the example in Fig.4, positioned based on their transcriptome profiles in PC1 and PC2 (left), PC3 (middle) and PC4 (right panel). Colour legend on the right describes all inferred cell states. **B)** Two-way hierarchical clustering of cells and genes from the same experiment. Columns are genes and rows are cells. Coloured boxes next to the cells' hierarchical tree represent same subpopulations marked in the PCA plots in A. Boxes marked 1 and 2 show early prespore and intermediate/early prestalk subpopulation, respectively. Boxes marked 3 and 4 show induced spore and stalk markers, respectively. Figure adapted from [3].

Figure 6. Continuous sampling of developmental progression.

A) Image of cells plated on bacteria, with feeding front on the left and development progression to the right. The white box approximates the region sampled. Right panel: schematic of the developmental stages sampled. **B)** Cell positioned in PC1 and PC2 space based on their transcriptional profiles. Overlayed is expression of sets of genes expressed during early development, aggregation, and aggregate. The arrow in the right panel shows approximate direction of development. **C)** Same as B but overlayed with relative cell density values. **D)** Cells positioned in PC1 and PC3 space (left) and PC1 and PC4 space (right), overlayed with expression of aggregation genes. **E)** Cells positioned in PC1 and PC5 space overlayed with expression of *pspA* and *tps3*, as spore and stalk markers, respectively. Figure adapted from [5].

Figure 7. 2D visualisation of cell state transitions during development.

A) Cells positioned in 2D space, after EE was performed on the first 11 PCs. From left to right: overlayed expression of genes expressed during early development, aggregation, and mound. Right panel: overlayed expression of *tps3* (stalk) and *pspA* (spore) markers. **B)** Jump, bottleneck, stalk and spore regions marked on the same plot as in A. The arrow shows the inferred direction of development. **C)** Zoom-in of the lower-right region of the plot in A, overlayed with the average expression of genes down and up-regulated during the morphological transition from single cells to streams [24] **D)** 3D cell density plot, with the same regions marked as in B **E)** The same as in A, overlayed with the expression of genes down and up-regulated during the morphological transition from loose to tight mounds [24]. Figure reproduced from [5].

Figure 8. Comparing different normalisation approaches.

A-C) Cell positioned in 2D space, based on their transcriptome profiles. **A**: The same as in Fig.7. **B**: Same as in A but performed on the raw data (not normalised) and using only variable genes (4849, instead of 9698 genes). **C**: Same as in A but normalised using the size factor from DESeq2 package in R. Lower panels show the same data as in upper panels but overlayed with relative cell density information. Jump, bottleneck, stalk, and spore regions are marked. Data used are from [5].

Figure 9. Comparing different thresholds, bin sizes and variance measures.

Genes downregulated during development become more variable than upregulated genes. Transcript variability (expressed as DM) of each gene with mean expression >10 read counts in the population of aggregating cells (6h) is plotted against the mean expression. Distribution of mean DM for each bin is shown in purple and black, for down and upregulated genes, respectively. Each bin contains 500 genes. Same data as in Fig.3, except for gene expression fold change threshold is 2-fold in each direction ($|\log_2 FC| > 1$). Bottom panel: result of decreasing the bin size to 100 genes. Right panel: result of increasing the FC threshold to 10-fold in each direction. Lower right: result of using CV^2 as a measure of variance (CV^2 does not correct for expression level or gene length). Figure reproduced from [2].

Figure 10. Exploring effects of different numbers of genes on the analysis.

A) Transcript variability (CV^2) of each gene plotted against mean expression. The vertical line marks the gene expression threshold. Marked yellow, black, and red are genes more variable than average (50%; 4849 genes). Black and red are the top 25% variable genes (2424). Red only are the 10% top variable genes (969). **B)** Same as in Figure 6, but cells are described only with variable genes (top 50%). Left panel: cells in PC1 and PC2 space. Overlayed is average expression of aggregation genes. Middle: cells in PC1 and PC5 space. Overlayed is the expression of *pspA* (spore marker). Right: same as left panel but overlayed with relative cell density values. **C)** The same as in B, but cells described only with top 25% variable genes. Middle panel is PC4 vs. PC1. **D)** The same as in B, but cells described only with top 10% variable genes. Middle panel is PC3 vs. PC1. Data from [5].

Figure 1

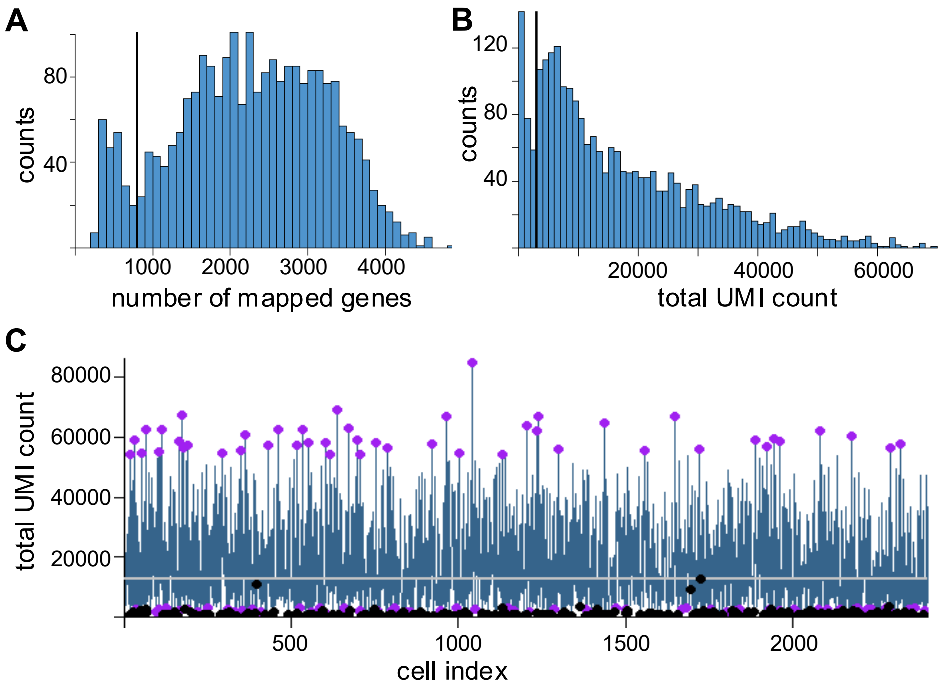

A
counts
80
40
1000
2000
3000
4000
number of mapped genes
B
counts
120
80
40
20000
40000
60000
total UMI count
C
total UMI count
80000
60000
40000
20000
500
1000
1500
2000
cell index

Figure 2

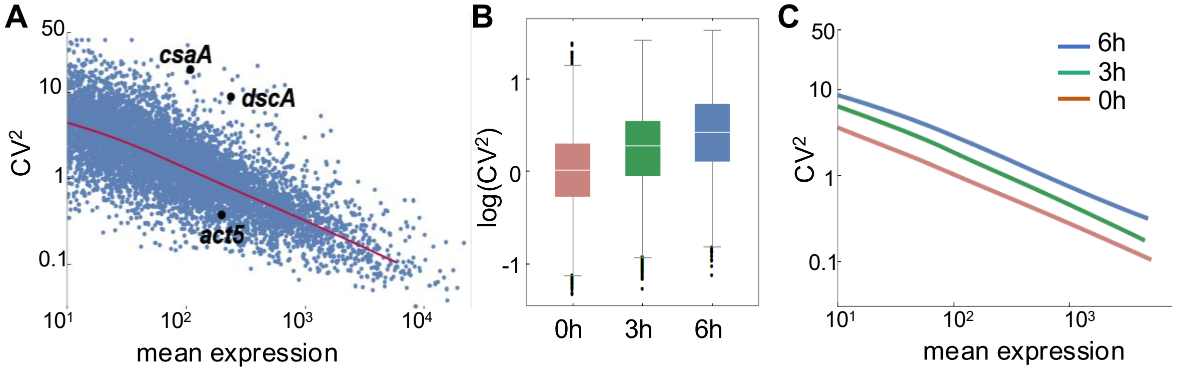

A
50
10
1
0.1
CV²
csaA
dscA
act5
10¹
10²
10³
10⁴
mean expression
B
1
0
-1
log(CV²)
0h
3h
6h
C
50
10
1
0.1
CV²
6h
3h
0h
10¹
10²
10³
mean expression

Figure 3

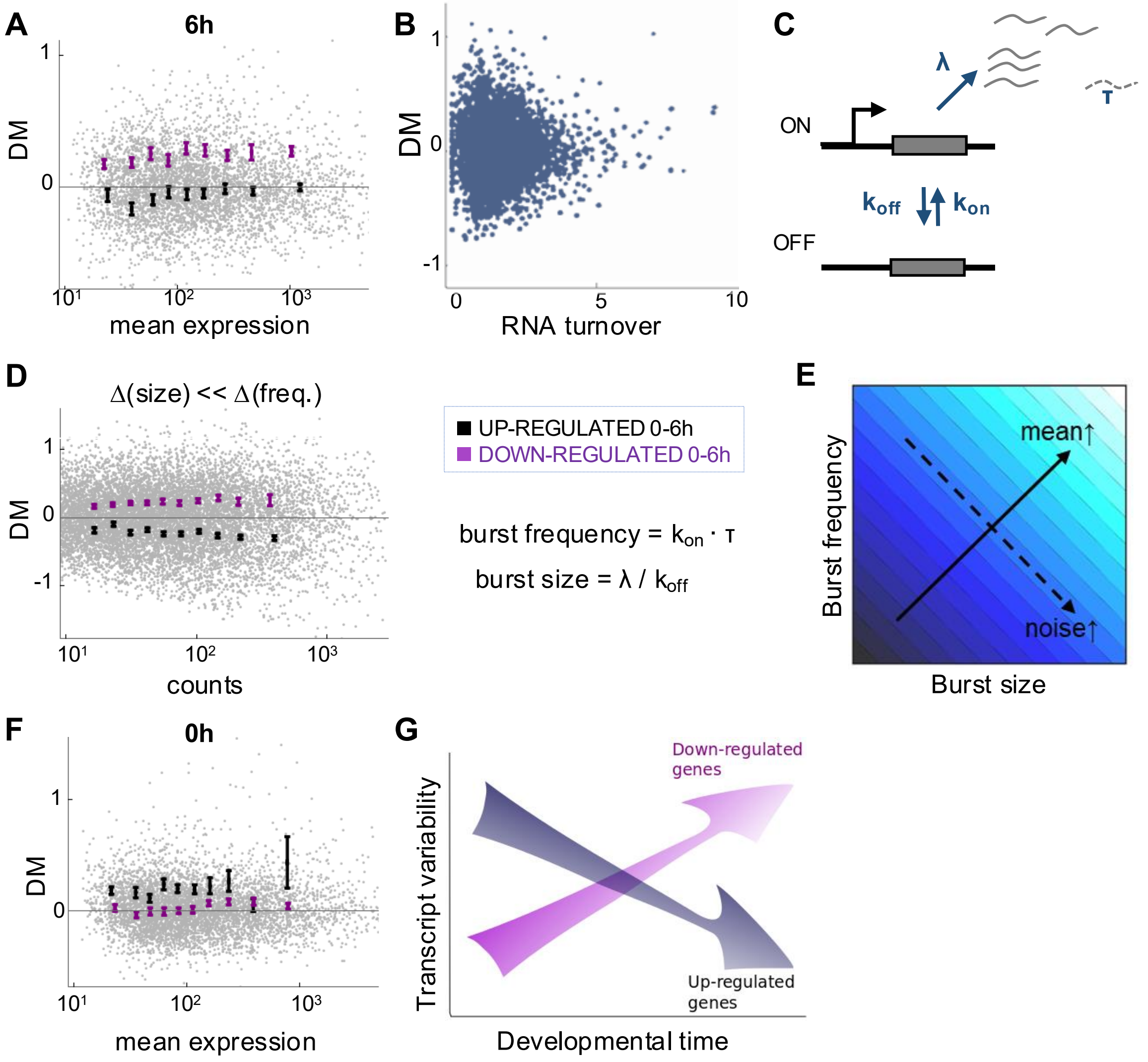
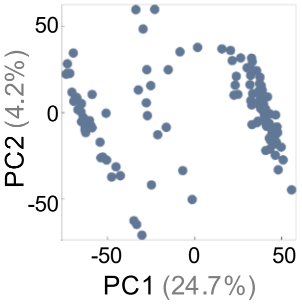

A
6h
DM
1
0
10^1
10^2
10^3
mean expression
B
DM
1
0
-1
0
5
10
RNA turnover
C
λ
τ
ON
k_{off}
k_{on}
OFF
D
Δ(size) << Δ(freq.)
DM
1
0
-1
10^1
10^2
10^3
counts
■ UP-REGULATED 0-6h
■ DOWN-REGULATED 0-6h
burst frequency = $k_{on} \cdot \tau$
burst size = λ / k_{off}
E
Burst frequency
mean↑
noise↑
Burst size
F
0h
DM
1
0
10^1
10^2
10^3
mean expression
G
Transcript variability
Down-regulated genes
Up-regulated genes
Developmental time

Figure 4

A



(+) PC2 loadings

GO biological process	FC	P value
aggregation involved in development	4.32	1.6x10^{-9}
chemotaxis to cAMP	4.10	0.00007
cell adhesion	3.60	0.00161
mitotic cytokinesis	3.57	0.00066
cell differentiation	3.10	0.00065
small GTPase mediated signal transduction	2.93	0.00265

(-) PC1 loadings ←→ (+) PC1 loadings

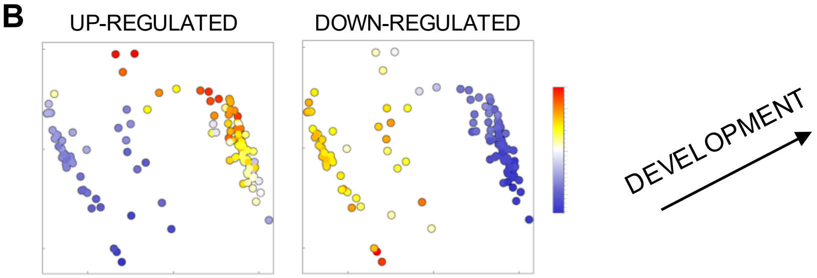
(-) PC1 loadings

GO biological process	FC	P value
cytoplasmic translation	5.40	0.00246
ribosome assembly	4.83	0.00368
RNA processing	1.90	0.00604

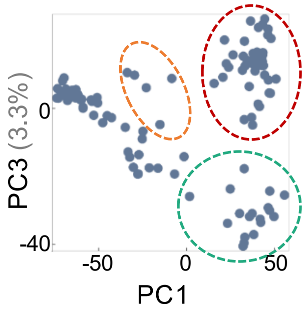
(+) PC1 loadings

GO biological process	FC	P value
multicellular organism development	9.42	0.02190
response to osmotic stress	8.32	0.04650
sorocarp development	3.77	0.01900
signal transduction	3.42	0.03070

B



C



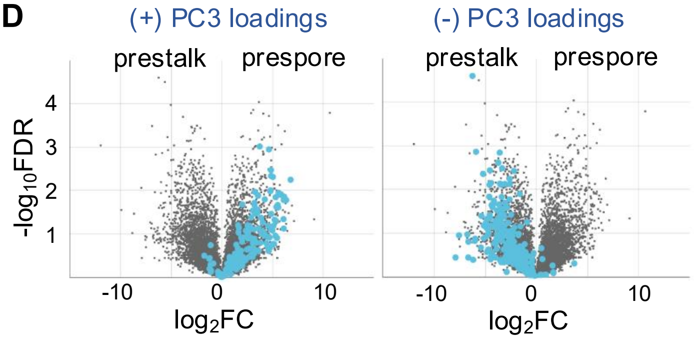
(+) PC3 loadings

GO biological process	FC	P value
spore wall assembly	20.61	0.0266

(-) PC3 loadings

GO biological process	FC	P value
response to organic cyclic compound	5.59	0.00245
regulation of cellular component size	5.43	0.02
actin cytoskeleton organization	4.96	0.00001
locomotion	3.30	0.0279
multi-organism process	2.23	0.0286

D



E

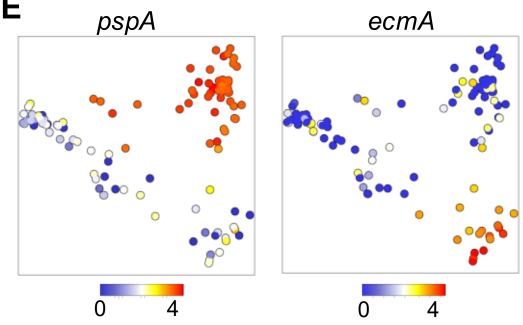


Figure 5

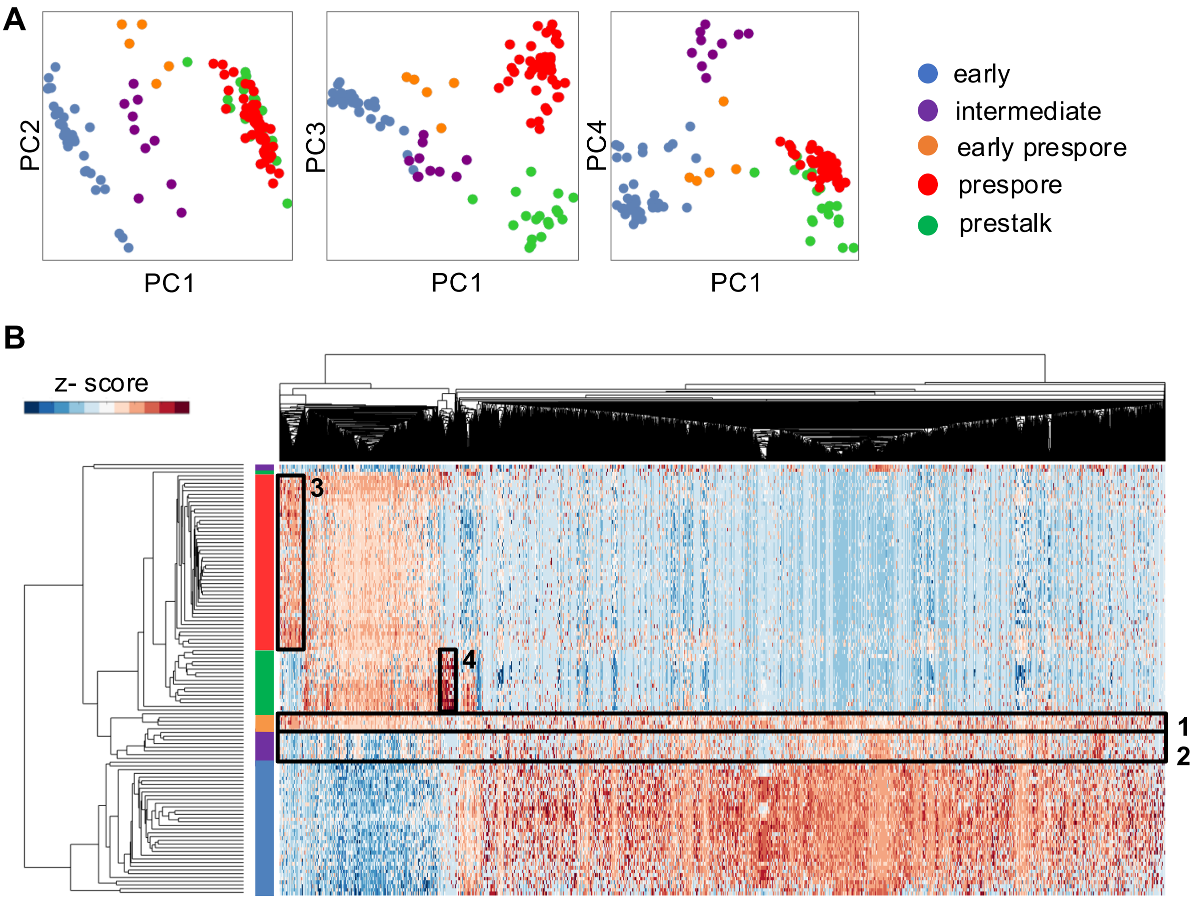
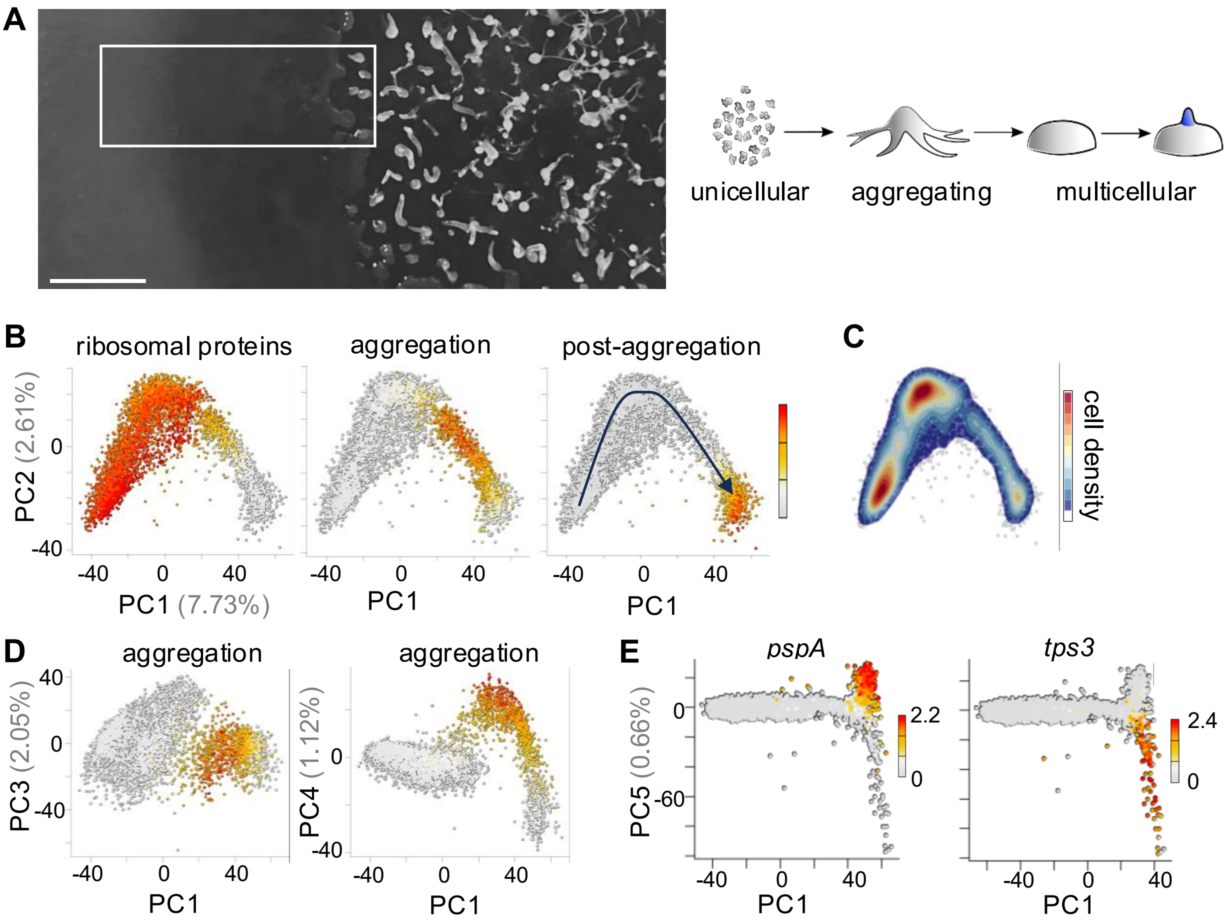

A
PC2
PC1
PC3
PC1
PC4
PC1
early
intermediate
early prespore
prespore
prestalk
B
z- score
3
4
1
2

Figure 6



A
unicellular
aggregating
multicellular
B
ribosomal proteins
aggregation
post-aggregation
PC2 (2.61%)
0
-40
-40
0
40
PC1 (7.73%)
-40
0
40
PC1
-40
0
40
PC1
C
cell density
D
aggregation
aggregation
PC3 (2.05%)
40
0
-40
-40
0
40
PC1
PC4 (1.12%)
0
-40
-40
0
40
PC1
E
pspA
tps3
PC5 (0.66%)
0
-60
2.2
0
-40
0
40
PC1
2.4
0
-40
0
40
PC1

Figure 7

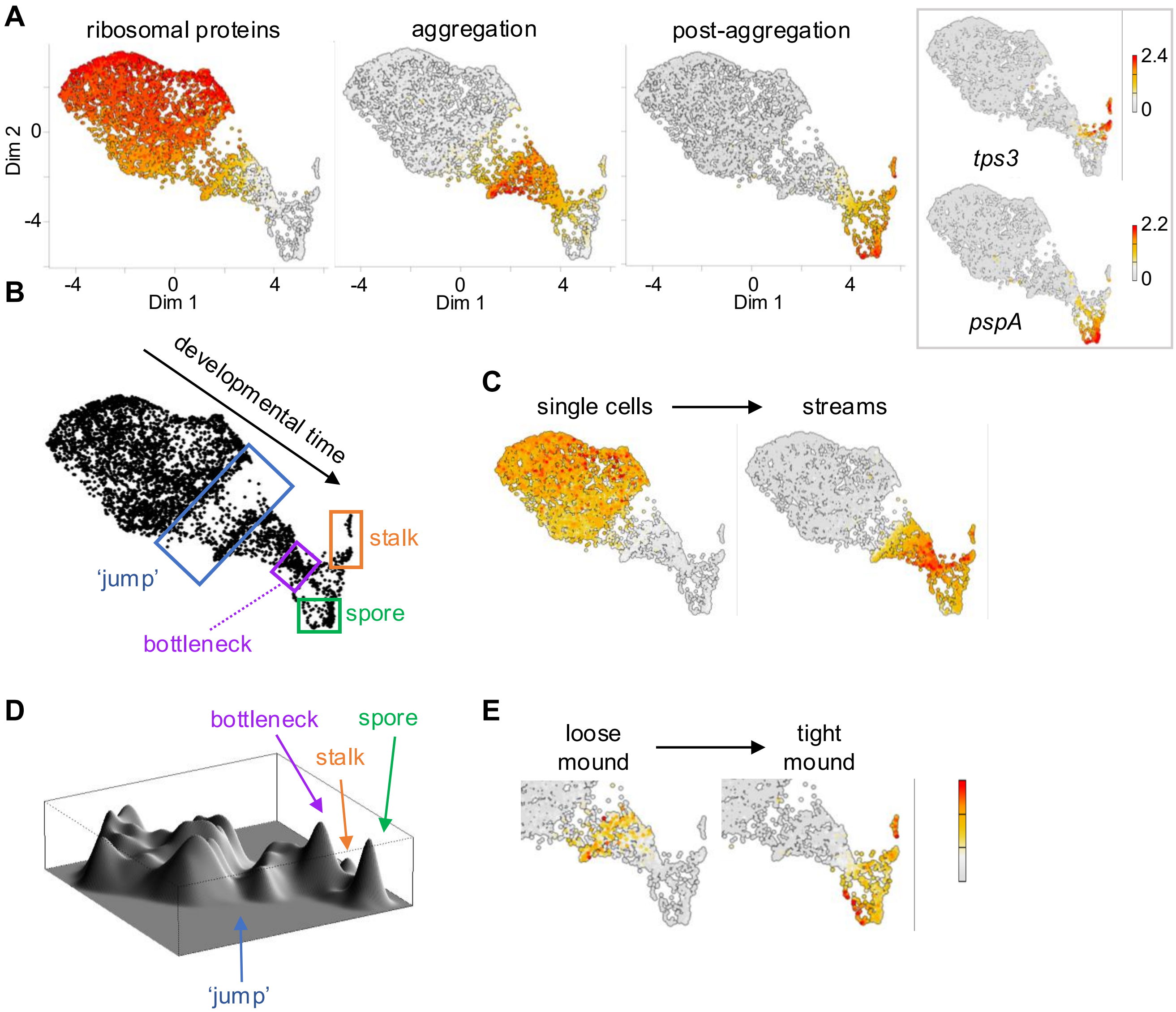
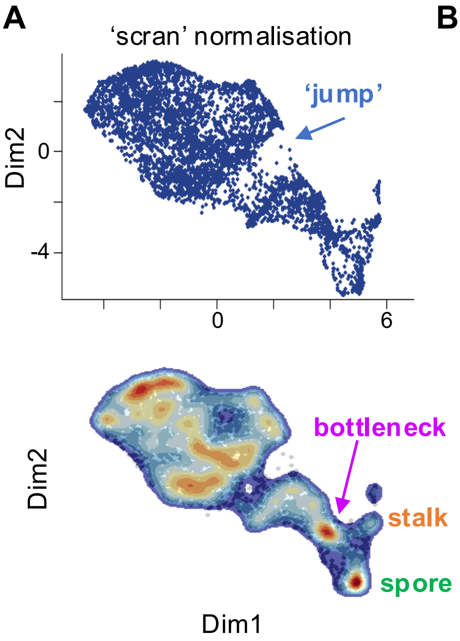
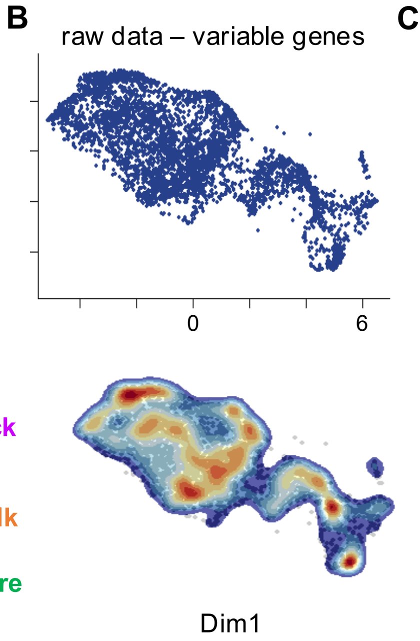

A
ribosomal proteins
aggregation
post-aggregation
Dim 2
0
-4
-4
0
4
Dim 1
-4
0
4
Dim 1
-4
0
4
Dim 1
2.4
0
tps3
2.2
0
pspA
B
developmental time
stalk
'jump'
spore
bottleneck
C
single cells
streams
D
bottleneck
spore
stalk
'jump'
E
loose
mound
tight
mound

Figure 8 - Notes





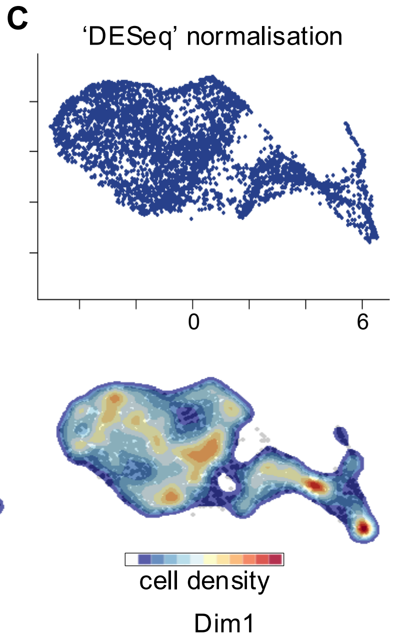
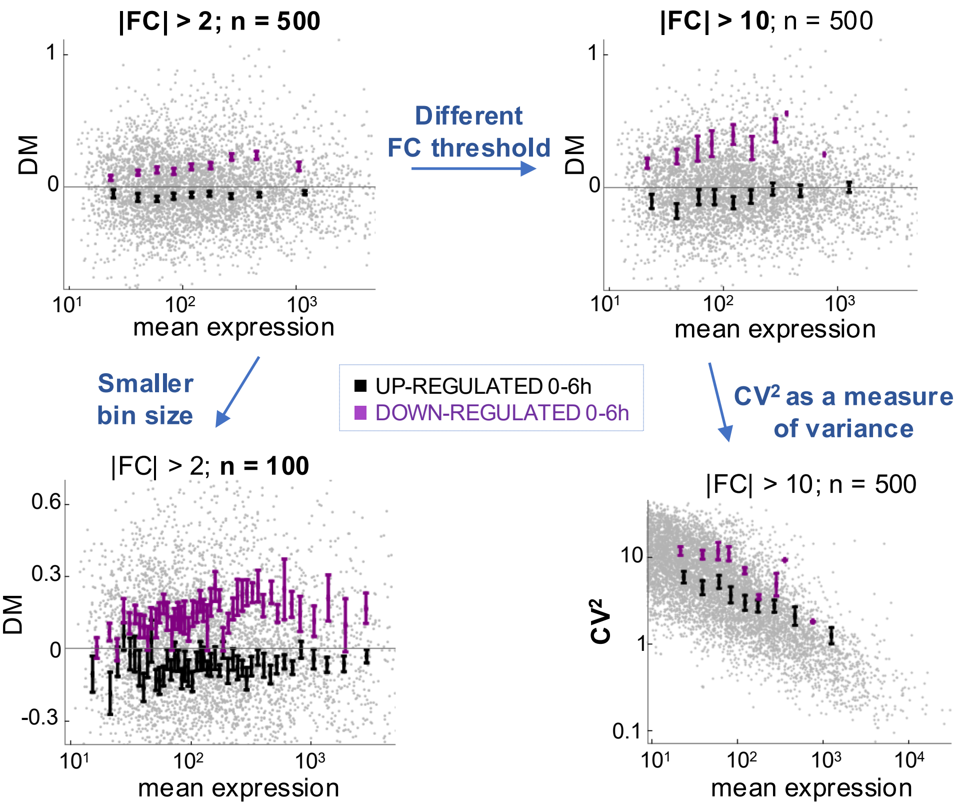


Figure 9 - Notes

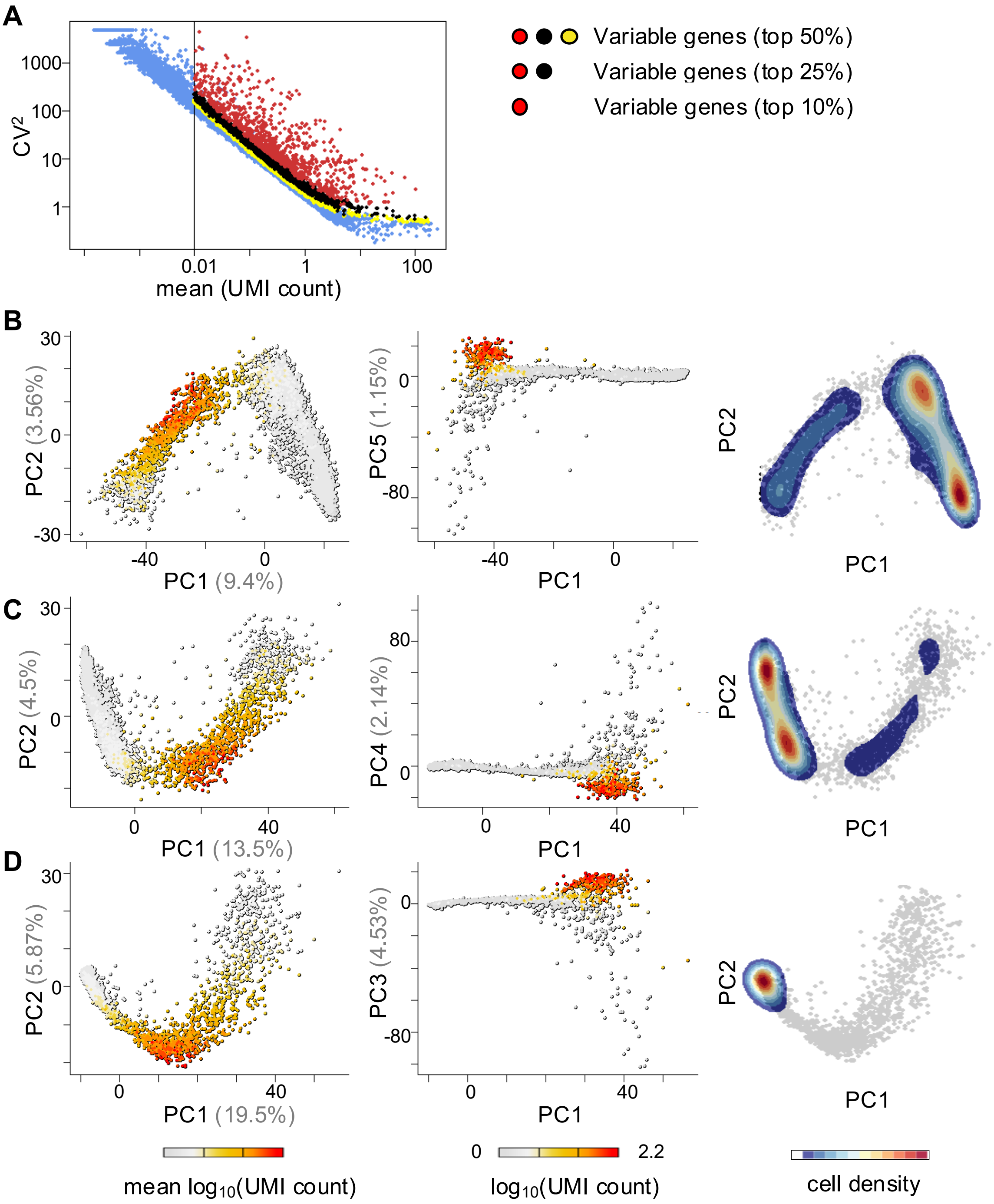


Different FC threshold

Smaller bin size

CV² as a measure of variance

Figure 10 - Notes


A
CV^2
mean (UMI count)
● ● ● Variable genes (top 50%)
● ● Variable genes (top 25%)
● Variable genes (top 10%)
B
PC2 (3.56%)
PC1 (9.4%)
PC5 (1.15%)
PC1
PC2
PC1
C
PC2 (4.5%)
PC1 (13.5%)
PC4 (2.14%)
PC1
PC2
PC1
D
PC2 (5.87%)
PC1 (19.5%)
PC3 (4.53%)
PC1
PC2
PC1
mean log$_{10}$(UMI count)
log$_{10}$(UMI count)
cell density