Accounting for technical noise in single-cell RNA-seq experiments

Philip Brennecke[1,6], Simon Anders[1,6], Jong Kyoung Kim[2,6], Aleksandra A Kołodziejczyk[2,3], Xiuwei Zhang[2], Valentina Proserpio[4], Bianka Baying[1], Vladimir Benes[1], Sarah A Teichmann[2,3], John C Marioni[2] & Marcus G Heisler[1,5]

Single-cell RNA-seq can yield valuable insights about the variability within a population of seemingly homogeneous cells. We developed a quantitative statistical method to distinguish true biological variability from the high levels of technical noise in single-cell experiments. Our approach quantifies the statistical significance of observed cell-to-cell variability in expression strength on a gene-by-gene basis. We validate our approach using two independent data sets from *Arabidopsis thaliana* and *Mus musculus*.

Progress in gene expression analysis using minute amounts of starting material has made single-cell transcriptomics accessible[1–5]. So far, the main goal has been discovery-driven research on gene expression in rare cells, and analysis has focused on global properties of the data. However, another promising application is the study of transcriptional heterogeneity within supposedly homogeneous cell types, a phenomenon of physiological importance[6–9], which can now be studied in a transcriptome-wide manner in single cells[10]. In such analyses, which should be distinguished from the more common two-group comparison setting (**Supplementary Note 1**), it is necessary to account for strong technical noise. Technical noise is unavoidable owing to the low amount of starting material, and it must be quantified in order to avoid mistaking it for genuine differences in biological expression levels. We present a statistical method that allows the user to assess, separately for each gene, whether the observed variation provides evidence of high biological variability, i.e., whether one can rule out that variation is merely a consequence of technical noise. Our approach is based on the observation that the strength of technical noise of a given gene depends mainly on the gene's average read count[3] and that this dependence can be inferred from a sufficiently rich set of spike-ins.

The low amount of RNA present in a single cell represents the main challenge in single-cell RNA-seq experiments. We demonstrated the relationship between technical noise and the amount of starting material with a dilution series, using technical replicates of decreasing amounts of total RNA taken from the same pool of total RNA (**Fig. 1** and **Supplementary Fig. 1**). Throughout our analysis we generated libraries using the protocol of Tang *et al.*[4]; we obtained very similar levels of technical noise when we compared it to results with another protocol[3] (**Supplementary Note 2** and **Supplementary Fig. 2**).

The dilution series allowed us to assess the accuracy of measurements of relative concentration for each starting amount[3] (**Fig. 1**). For 5,000 pg of input material, the noise pattern was comparable to that of technical replicates from bulk RNA-seq experiments, in which the spread can be accounted for by the Poisson distribution[11]. However, the number of genes affected by high levels of technical noise increased notably at lower amounts of starting material (for example, a transcript could have 100–1,000 read counts in one technical replicate but 0 counts in another), as also noted previously[3]. Nevertheless, genes with a high read count showed very good agreement between replicates even for the 10-pg data point—meaning that low–read count genes show strong noise and high–read count genes show weak noise; what changes across differing amounts of starting material is the

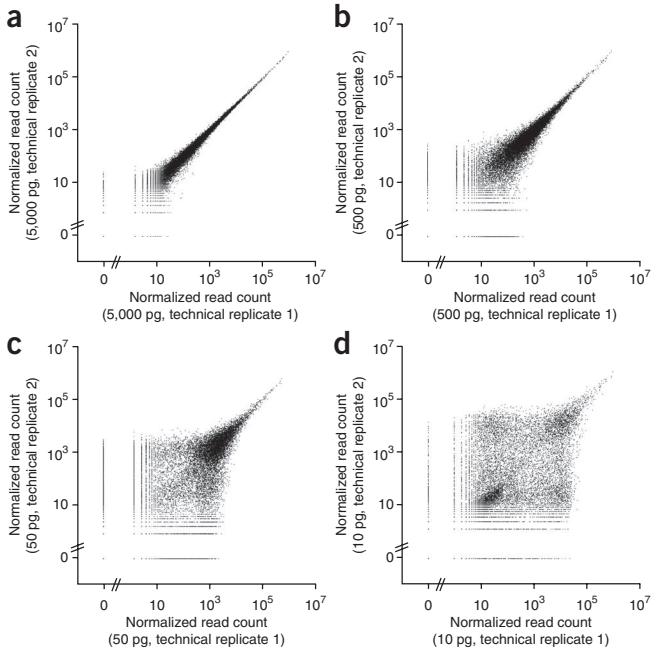

Figure 1 | Dilution series of total *A. thaliana* RNA. (**a**–**d**) Experiments with 5,000 pg (**a**), 500 pg (**b**), 50 pg (**c**) and 10 pg (**d**) of total RNA. For a scatter plot of the full dilution series, see **Supplementary Figure 1**.

[1]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [2]EMBL, European Bioinformatics Institute (EBI), Hinxton, UK. [3]Wellcome Trust Sanger Institute, Hinxton, UK. [4]Medical Research Council Laboratory of Molecular Biology, Cambridge, UK. [5]University of Sydney, Sydney, Australia. [6]These authors contributed equally to this work. Correspondence should be addressed to J.C.M. (marioni@ebi.ac.uk) or M.G.H. (heisler@embl.de).

RECEIVED 3 APRIL; ACCEPTED 8 AUGUST; PUBLISHED ONLINE 22 SEPTEMBER 2013; CORRECTED ONLINE 11 OCTOBER 2013 (DETAILS ONLINE); DOI:10.1038/NMETH.2645

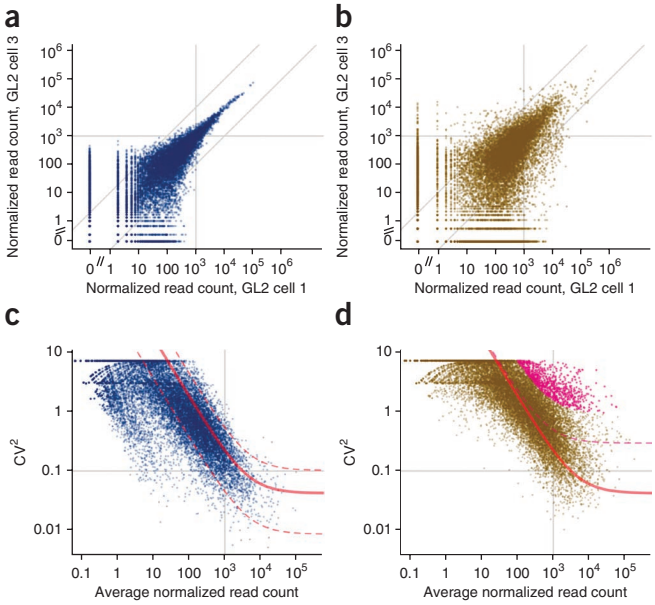

Figure 2 | Technical noise fit and inference of highly variable genes using total HeLa RNA as a spike-in in GL2 cells. (**a**) Scatter plot for normalized read counts for the HeLa total RNA spike-in. (**b**) Scatter plot for normalized read counts for the plant genes. (For five more GL2 cells, see **Supplementary Fig. 5**; for the QC cells, see **Supplementary Fig. 6**). (**c**) Technical noise fit: squared coefficients of variation are plotted against the means of normalized read counts for each HeLa gene using data from all seven GL2 cells. The solid red curve represents the fitted variance-mean dependence; the dashed lines indicate a 95% interval for the expected residual distribution (Online Methods). (**d**) Identification of highly variable genes across all seven GL2 cells. For the genes highlighted in magenta, the coefficient of biological variation significantly exceeds 50% according to our test (with the false discovery rate controlled at 10%). The dashed line marks the expected position of genes with 50% biological CV; however, owing to the statistical uncertainty of CV estimation, statistical significance is achieved only for CV^2 values well above this line.

read-count range in which noise strength transitions from weak to strong. Our approach uses technical spike-ins to quantify this dependence across the whole dynamic range of expression before exploiting this information for subsequent inferences on biological cell-to-cell differences.

To illustrate our method, we used data from a single-cell RNA-seq experiment performed on *A. thaliana* cells. Cells marked by the expression of GFP driven by the *GL2* (ref. 12) or *WOX5* promoters were collected from the roots of *A. thaliana* seedlings. The former marks non-hair cells in the root epidermis ('GL2 cells'), and the latter marks cells from the quiescent center of the root ('QC cells'). We used total RNA from HeLa cells as spike-ins because it covers the whole dynamic range at sufficient density and behaves similarly to plant RNA (**Supplementary Fig. 3**). Moreover, it is easily distinguishable at the sequence level. For comparison, we also added the set of 92 spike-ins developed by the External RNA Control Consortium (ERCC)[13].

For the GL2 cells, on average 57.6% and 22.9% of reads mapped back to the *A. thaliana* and *H. sapiens* genomes, respectively, suggesting that a typical GL2 cell contains ~60 pg of total RNA (Online Methods and **Supplementary Table 1**). QC cells contained on average only ~10 pg (**Supplementary Table 1**) and therefore represent a technically more challenging sample type. Nevertheless, we could map all 13 single-cell transcriptomes to the correct cell types in a previously published *A. thaliana* root atlas[12], a result that provides confidence in the quality of our data (**Supplementary Note 3** and **Supplementary Fig. 4**).

A comparison of two spiked GL2 cells illustrated the difference between the correlation of read counts from the HeLa cell genes (**Fig. 2a**), which are affected by technical noise only, and those from the plant genes (**Fig. 2b** and **Supplementary Figs. 5** and **6**), which are subject to both technical noise and biological variability. We saw that for genes with a read count up to ~100, the technical noise was 'maximal': the same HeLa gene could have ~100 read counts in one sample and no read counts at all in the other. It was therefore impossible for a plant gene of similar expression strength to show stronger variability even if it were subject to strong biological variability in addition to technical noise.

To quantify the relationship between technical noise and mean expression strength, we first normalized the counts to account for sequencing depth and cellular RNA content (Online Methods and **Supplementary Note 4**), though not for transcript length (**Supplementary Note 5** and **Supplementary Fig. 7**). Then we calculated, for each plant gene and each spike-in, the squared coefficient of variation (CV^2) of the normalized read counts across samples (**Fig. 2c**,**d**). To capture the dependence of the CV^2 of the spike-ins on their average normalized count μ, we fit a curve to the observed data, using the parameterization $CV^2 = a_1/\mu + \alpha_0$ (Online Methods).

We developed a statistical approach to test, for each gene, the null hypothesis that the biological coefficient of variation is less than a level chosen by the user. All genes will display some biological variability in expression from cell to cell, but a high level of variance (exceeding the specified threshold) will indicate genes important in explaining heterogeneity within the cell population under study (Online Methods and **Supplementary Note 6**). The Supplementary Software provides extensively commented code to illustrate how to perform the analysis in the statistical programming language R.

Using our approach—and correcting for multiple testing[14]—we found, at a false discovery rate of 10%, 876 genes across the seven GL2 cells that showed statistically significant evidence against the null hypothesis that their biological coefficient of variation was less than our chosen minimum CV of 50% (i.e., $CV^2 < 0.25$). We therefore considered these genes to be highly variable (**Fig. 2d** and **Supplementary Table 2**). As the QC cells are much smaller than the GL2 cells, with only about one-sixth of the starting amount of total RNA (**Supplementary Table 1**), the technical noise was noticeably stronger in the QC cells (**Supplementary Figs. 6** and **8**), which considerably reduced the statistical power

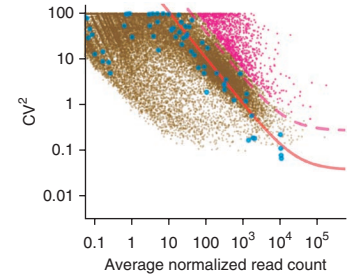

Figure 3 | Technical noise fit and inference of highly variable genes using ERCC spike-ins. Our statistical method was applied to a data set comprising 91 mouse cells spiked with only the ERCC spike-in set. Blue dots correspond to ERCC data points, brown dots to mouse genes, and magenta dots to significantly highly variable mouse genes (at a 10% false discovery rate). The solid red line represents the technical noise fit, and the dashed magenta line marks the expected position of genes with 50% biological CV.

to infer highly variable genes across cells. As a result, only 64 genes were identified as being highly variable in this cell type (**Supplementary Fig. 8** and **Supplementary Table 3**).

Across the highly variable genes, we found clear enrichments for Gene Ontology (GO) categories such as "Nucleosome Assembly" ($P = 2.5 \times 10^{-24}$), "Cell Proliferation" ($P = 6.0 \times 10^{-6}$), "Anaphase" ($P = 5.4 \times 10^{-7}$) and "Cell Wall" ($P = 3.8 \times 10^{-6}$), which are expected to vary across cells because they are indicative of distinct growth states for GL2 and QC cells (for a full list, see **Supplementary Table 4**). Additionally, individual GO categories tended to be upregulated in a coordinated fashion in individual cells, a result suggesting that these GO categories reflect different cellular states and possible instances of co-regulation[10] (**Supplementary Fig. 9**). However, the analysis of a larger number of cells would be needed to further substantiate these claims.

As in any hypothesis test, our results did not imply that none of the remaining genes was highly variable. In fact, for all genes in the GL2 cells with normalized counts below ~100 (weakest significant gene in **Fig. 2d**), even the strongest biological variation could not be detected because technical noise was maximal (**Supplementary Note 7**). This is not a limitation of our statistical approach; rather, it is a direct consequence of the limited sensitivity of current single-cell RNA-seq protocols. (**Supplementary Note 4**).

The precision of our approach requires an accurate characterization of the dependence of technical noise strength on average read count. Although using total RNA as a spike-in is a pragmatic solution with certain advantages (**Supplementary Note 8** and **Supplementary Fig. 10**), it has the disadvantage that approximately half of the reads per sample are allocated to modeling technical variability. An alternative approach would be to use a smaller spike-in set such as the 92 ERCC spike-ins[13], which were also added to our *A. thaliana* single-cell samples. For our small-scale experiments with only six or seven cells, the CV^2 estimate provided by each individual spike-in had a large sampling variance, and the 92 ERCC spike-ins did not provide sufficient information to obtain a stable fit (**Supplementary Fig. 11**). However, in an experiment with more cells, each individual spike-in provides a more precise estimate of the CV^2 at the respective read-count value.

To illustrate this, we applied our method to identify highly variable genes across a set of 91 single-cell transcriptomes, obtained from a single cell type of the murine immune system, wherein each cell had been spiked with the ERCC spike-in set only (Online Methods). These are data from a very recent experiment, and the biology is still being explored; here we used these data (an anonymized version of which is available in **Supplementary Table 5**) to demonstrate the efficacy of our approach using only ERCC spike-ins. The relationship between technical variability and expression strength showed a robust fit (**Fig. 3**). We identified highly variable genes across the 91 cells and found 1,198 at a 10% false discovery rate. This set of genes was strongly enriched for several GO categories including "Cytokine Activity" ($P = 6.9 \times 10^{-8}$), as expected. This suggested that the set of genes identified are likely to be physiologically relevant. We also note that the sequencing coverage of these data was lower than that used in the *A. thaliana* experiments, thereby illustrating that sequencing deeply is typically unnecessary for drawing biological conclusions from single-cell transcriptomes (**Supplementary Note 9** and **Supplementary Fig. 12**).

METHODS

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

Accession codes. ArrayExpress: raw sequencing data are available at accession E-MTAB-1499.

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

ACKNOWLEDGMENTS

We thank E. Furlong and W. Huber for helpful discussions. We also acknowledge K. Birnbaum (New York University) for kindly providing pWOX5::GFP and pGl2::GFP seed. S.A. acknowledges partial funding from the European Union (FP7-Health, project Radiant); M.G.H. acknowledges the Australian Research Council for present funding. The EMBL Genomics Core Facility provided technical support for this work. We acknowledge A. Surani for the use of the C1 Single-Cell Auto Prep System in his lab and B. Jones for performing the experiment. We also acknowledge A. McKenzie (Medical Research Council Laboratory of Molecular Biology) for the *Il13-GFP* reporter mice and the Sanger-EBI Single Cell Centre for technical support. We acknowledge the support of European Research Council Starting Grant no. 260507, ThSWITCH.

AUTHOR CONTRIBUTIONS

P.B. designed plant cell experiments, carried out experiments, interpreted results and wrote the paper; S.A. developed the statistical method, performed bioinformatics analyses and wrote the paper; J.K.K. performed bioinformatics analyses and helped write the paper; A.A.K. designed and carried out mouse cell experiments and helped write the paper; X.Z. designed and analyzed mouse cell experiments and helped write the paper; V.P. designed and carried out mouse cell experiments and helped write the paper; B.B. adapted an Illumina sequencing library preparation protocol; V.B. contributed to adapting the Illumina sequencing library preparation protocol and gave advice; S.A.T. designed mouse cell experiments and helped write the paper; J.C.M. contributed to the development of the statistical method, performed bioinformatics analyses, supervised the project and wrote the paper; M.G.H. initiated the project, designed plant cell experiments, interpreted results, supervised the project and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep.* **2**, 666–673 (2012).
2. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
3. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
4. Tang, F. *et al. Nat. Protoc.* **5**, 516–535 (2010).
5. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
6. Chambers, I. *et al. Nature* **450**, 1230–1234 (2007).
7. Reynolds, N. *et al. Cell Stem Cell* **10**, 583–594 (2012).
8. Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E. & Huang, S. *Nature* **453**, 544–547 (2008).
9. Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K. & Niwa, H. *Development* **135**, 909–918 (2008).
10. Shalek, A.K. *et al. Nature* **498**, 236–240 (2013).
11. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. *Genome Res.* **18**, 1509–1517 (2008).
12. Brady, S.M. *et al. Science* **318**, 801–806 (2007).
13. Jiang, L. *et al. Genome Res.* **21**, 1543–1551 (2011).
14. Benjamini, Y. & Hochberg, Y. *Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

ONLINE METHODS

Plant growth conditions. The pWOX5::GFP and pGL2::GPF seeds were sterilized using 4 h of standard bleach vapor sterilization[15]. After sterilization, seeds were stratified in the dark at 4 °C for 72 h, and plants were subsequently grown vertically for 4 d in constant-light conditions at 22 °C and 45% humidity. Standard root plates were used for growing plants (i.e., 1× Murashige and Skoog basal medium, 0.5% sucrose, 2.6 mM 2-(*N*-Morpholino) ethanesulfonic acid (pH 5.7) and 1% agarose).

Isolation of single plant cells. QC and GL2 cells were isolated as previously described[16]. In brief, root tips of fluorescent marker lines were cut off using a scalpel and transferred to solution B for protoplasting (0.6 M Mannitol, 10 mM KCl, 10 mM $MgCl_2$, 10 mM $CaCl_2$, 1 mg/ml BSA, 0.39 mg/ml MES, pH 5.5, 1.5% Cellulase R10 (Yakult), 0.1% Pectolyase Y-23 (Yakult)). Root tips were protoplasted on a platform shaker for 30 min for the GL2 cells and 60 min for the QC cells. Release of QC cells was facilitated by gently streaking root tips on a 75-µm cell strainer every 15 min, whereas GL2 cells were released by gently pipetting pieces of tissue up and down using a mouth pipette. GL2 cells that are present along the whole root were picked from a small region of the root to minimize variability due to position. Single cells of interest were identified by GFP signal and washed three times using individual drops of washing solution A (0.6 M Mannitol, 10 mM KCl, 10 mM $MgCl_2$, 10 mM $CaCl_2$, 1 mg/ml BSA, 0.39 mg/ml MES, pH 5.5). After three washes, cells were transferred within a volume of <0.5 µl of solution A to PCR tubes containing the lysis buffer.

Preparation of cDNA libraries. Single-cell or technical-replicate cDNA libraries were prepared as described previously[4,5] with the following modifications: PBS had to be replaced with solution A for plant cells because *A. thaliana* cells rapidly die in PBS (data not shown). We confirmed that solution A does not negatively affect the performance of the protocol (**Supplementary Fig. 13**). Furthermore, we included a third RNase inhibitor from Qiagen (cat# 129916), used 1 µl of the reverse-transcription master mix, and performed 24 cycles of the initial PCR. HeLa total RNA and ERCC spike-ins were diluted using RNase-free water. 1 µl of a 1:1,000,000 dilution of the ERCC spike-ins and 50 pg of HeLa total RNA were included in the lysis buffer per reaction where applicable. Final cDNA libraries (200–2,000 ng depending on the amount of input material) were checked for known marker genes using qPCR (**Supplementary Fig. 14**), and after passing quality control, libraries were fragmented with the Covaris S2 system as reported previously[4,5] using Covaris mircoTUBEs (cat# 520045) and a volume of 130 µl for shearing (libraries were diluted to that volume using RNase-free water). After fragmentation, the volume was reduced to 85 µl using a SpeedVac concentrator, and samples were subjected to standard Illumina library preparation using the NEBNext DNA Sample Prep Master Mix Set 1 kit according to the manufacturer's instructions. Modified Illumina PE adapters including custom-made multiplexing barcodes (**Supplementary Table 6**) were ligated (amount of adapters was adjusted according to amount of input material), and Illumina PE primers (PE PCR Primer 1.0 and PE PCR Primer 2.0) were used for the PCR enrichment step (ten cycles) of the NEBNext protocol. The final purification step was performed using AMPure XP beads (Beckman Coulter) rather than columns, and clusters were generated by following the standard Illumina protocol. Samples were sequenced on an Illumina HiSeq 2000 machine. Single-end 50-bp reads were used, and sequencing yielded approximately 20–90 million reads per sample, dependent on the extent of multiplexing. All primers used for quality-control qPCRs are listed in **Supplementary Table 7**.

Preparation of cDNA libraries using the SMARTer Ultra Low RNA Kit (Clontech). This protocol was used only for comparison to the protocol developed by Tang and coworkers[4,5]. *A. thaliana* total RNA was diluted using RNase-free water, and 50 pg was used for each cDNA library preparation using the SMARTer Ultra Low RNA Kit for Illumina Sequencing from Clontech according to manufacturer's instructions. 24 cycles of PCR were used for library amplification, and after passing quality control, finalized cDNA libraries (6–20 ng) were fragmented with the Covaris S2 system as reported previously[4,5] using Covaris mircoTUBEs (cat# 520045) and a volume of 130 µl for shearing (libraries were diluted to that volume using RNase-free water). After fragmentation, the volume was reduced to 44 µl using a SpeedVac concentrator, and samples were subjected to standard Illumina library preparation using the NEBNext ChIP-Seq Sample Prep Master Mix Set 1 kit according to the manufacturer's instructions. Illumina PE adapters including custom-made multiplexing barcodes (**Supplementary Table 6**) were ligated (the amount of adapters was adjusted according to the amount of input material), and Illumina PE primers (PE PCR Primer 1.0 and PE PCR Primer 2.0) were used for the PCR enrichment step (15 cycles) of the NEBNext protocol. The final purification step was performed using AMPure XP beads (Beckman Coulter) rather than columns, and clusters were generated by following the standard Illumina protocol. Samples were sequenced on an Illumina HiSeq 2000 machine. Single-end 50-bp reads were used, and sequencing yielded approximately 35,000,000 reads per sample.

Alignment of reads. Reads were aligned to the *A. thaliana* genome (TAIR10) and to a set of known splice sites from the GTF file for TAIR10 provided by Ensembl Plants (release 15) (TAIR10, release 15) using GSNAP (version 2012-07-20) with default options[17]. For samples with the HeLa total RNA spike-in, reads were mapped simultaneously to the *Homo sapiens* (GRCh37) and *A. thaliana* genomes (TAIR10). For the *H. sapiens* genome, the set of known splice sites was taken from the GTF file for GRCh37 provided by Ensembl (release 69). We considered only reads uniquely mapped to the genomes. Owing to the large evolutionary distance between *H. sapiens* and *A. thaliana*, cross-mapping of reads is not an issue: less than 0.0015% of the reads could not clearly be assigned to one of the two species by the aligner and hence have been excluded from the analysis. From the mapped reads and the GTF files, we counted reads for each gene using htseq-count (http://www-huber.embl.de/users/anders/HTSeq/). The read count table is available in **Supplementary Table 8**.

Mapping of cells to a spatiotemporal atlas of the root. QC and GL2 cells were mapped to a spatiotemporal atlas of the *A. thaliana* root that has been described elsewhere[12]. The atlas consists of 19 GFP-marked cell populations representing 14 nonoverlapping cell types, as well as 13 sections along the longitudinal axis (which also represents a temporal axis) of the *A. thaliana* root. All Affymetrix

doi:10.1038/nmeth.2645

CEL files generated on an ATH1 microarray, from the AREX LITE repository (http://www.arexdb.org/), were normalized by using the RMA method implemented in the "affy" package of Bioconductor[18]. We removed all probe sets containing genes that did not appear in Ensembl Plants (TAIR10, release 15) and merged intensity values of multiple probe sets mapping to the same gene by taking the maximum value. For mapping, we constructed a *k*–nearest neighbor classifier that assigns the class label of a query sample or cell based on its closest training sample in the spatiotemporal atlas. We set *k* to 1 because we have only two or three samples for each class in the training data, and we used Spearman's rank correlation coefficient as a similarity measure between samples. To remove nonvariable genes among samples in the spatiotemporal atlas, we chose the top 3% genes according to the coefficient of variation of normalized intensities across samples and computed Spearman's rank correlation coefficient using these genes. This cutoff was chosen because it maximized classification accuracy as adjudged by a leave-one-out cross-validation analysis performed using only the spatial atlas data. To visualize the mapping of QC and GL2 cells to the spatiotemporal atlas, we performed principal-component analysis on the normalized data matrix using the princomp function in Matlab, wherein each row represents one of the top 3% genes and each column represents one of the samples (53 samples for the spatial atlas and 25 samples for the temporal atlas together with 13 samples for the QC and GL2 cells). The relationship between columns (samples or cells) was visualized by loadings (or principal-component coefficients) of the first three principal components.

Estimating the amount of total RNA of single cells. The amount of total RNA obtained from single cells, which were spiked with the HeLa total RNA spike-in, was estimated through the proportion of reads mapped to the *A. thaliana* genome. For the three technical replicates of 50 pg *A. thaliana* total RNA with 50 pg total HeLa RNA (shown in **Supplementary Fig. 3**), this proportion had a mean value of 0.5808. Assuming a linear relationship between the amount of *A. thaliana* total RNA and the proportion of reads mapped to the *A. thaliana* genome, the amount of total mRNA of a single cell can be estimated by 50 pg × *x*/0.5808, where *x* is the proportion of reads of the sample mapped to the *A. thaliana* genome.

Normalization. This section and the next two describe analysis steps that have been carried out using the statistical programming environment R. The complete code used, with extensive commenting, is available in the **Supplementary Software** to facilitate reuse of our method.

To normalize the read counts, we used the method that we developed for DESeq[19] (see also **Supplementary Note 6**). Briefly, for each gene, *i*, we calculate the geometric mean

$$k_i^{\mathrm{M}} = \left(\prod_{j=1}^{m} k_{ij} \right)^{1/m}$$

over the counts k_{ij} across all the samples $j = 1,\ldots, m$ and then use, for each sample, the median of the ratio of the sample's counts to these means as a 'size factor': $s_j = \mathrm{median}_i(k_{ij}/k_i^{\mathrm{M}})$.

Notably, we calculate two sets of size factors: (i) by running the median over only the 'technical genes' (i.e., the spiked-in HeLa genes), we obtain the 'technical size factors' s_j; and (ii) by running the median over only the 'biological genes' (i.e., the plant genes), we obtain the 'biological size factors' s_j^{B}. Each set of size factors is used to normalize the expression measures for technical or biological genes by dividing the read counts by the appropriate size factor in order to obtain normalized read counts. As discussed in more detail in **Supplementary Note 4**, the two normalizations have different effects: whereas the technical normalization accounts for only sequencing depth, the biological normalization accounts also for differences in the amount of biological starting material obtained from each cell.

Optionally, one may divide the count values not only by size factors but also by transcript length (**Supplementary Table 9**). For the analysis presented here, we did not account for length in this way; see **Supplementary Note 5** for a discussion.

Estimating and fitting technical noise. This and the next section briefly outline the method used for inference; for its justification, see **Supplementary Note 6**.

For each technical gene *i*, we estimate the sample mean and sample variance of its normalized counts, i.e., the respective quantities

$$\hat{\mu}_i = \frac{1}{m} \sum_{j=1}^{m} \frac{k_{ij}}{s_j}$$

and

$$\hat{W}_i = \frac{1}{m-1} \sum_{j=1}^{m} \left(\frac{k_{ij}}{s_j} - \hat{\mu}_i \right)^2$$

Figure 2c is a plot of the squared coefficients of variation (CV^2), $\hat{w}_i = \hat{W}_i / \hat{\mu}_i^2$, against $\hat{\mu}_i$. We fit a generalized linear model (GLM) of the gamma family with identity link and parameterization $w = \tilde{a}_1/\mu + \alpha_0$ to this plot using the GLM fitter provided in the R package statmod. We fit the model using only those value pairs $(\hat{\mu}_i, \hat{w}_i)$ that surpass a certain threshold $\hat{\mu}_i > \mu_{\mathrm{th}}$; this serves to exclude genes with very high CV^2 and hence high uncertainty in the estimate for μ. The threshold is chosen such that only 5% of the genes with $\hat{w}_i > 0.3$ have $\hat{\mu}_i > \mu_{\mathrm{th}}$. (For the mouse data, we allowed 20% rather than only 5% of the ERCC transcripts to surpass the threshold, as we have many fewer ERCC spike-ins than HeLa genes.) The coefficients α_0 and $\tilde{a}_1$ obtained from the fit (solid red line in **Fig. 2c**) characterize the technical noise and are used in the following.

Testing for high variance. Again, we calculate the sample moments, now for biological gene *i*

$$\hat{\mu}_i^{\mathrm{B}} = \frac{1}{m} \sum_{j=1}^{m} \frac{k_{ij}^{\mathrm{B}}}{s_j^{\mathrm{B}}}$$

and

$$\hat{W}_i^{\mathrm{B}} = \frac{1}{m-1} \sum_{j=1}^{m} \left(\frac{k_{ij}^{\mathrm{B}}}{s_j^{\mathrm{B}}} - \hat{\mu}_i^{\mathrm{B}} \right)^2$$

Slightly simplifying, the expected value of $\hat{W}_i^{\mathrm{B}}$ should be the sum of the technical and the biological components of the variance.

Up to corrections (see below), the technical variance takes the value predicted by the technical noise fit, $\tilde{a}_1 \hat{\mu}_i^B + \alpha_0(\hat{\mu}_i^B)^2$. For the total variance, we add to this $\alpha_i^B(\hat{\mu}_i^B)^2$, where α_i^B is the squared true coefficient of biological variation.

A more careful calculation, given in **Supplementary Note 6**, gives rise to a number of extra correction terms, yielding the result $E\hat{W}_i^B = \Omega(\alpha_i^B, \hat{\mu}_i^B)$, with

$$\Omega(\alpha,\mu) = \frac{\mu(\Psi + a_1\Theta) + \mu^2\alpha_F}{1 + \frac{\alpha_F}{m}}$$

where $\alpha_F = \alpha_0 + \alpha + \alpha_0\alpha$, $a_1 = \tilde{a}_1 - \Xi$ and

$$\Xi = \frac{1}{m}\sum_j \frac{1}{s_j}, \quad \Psi = \frac{1}{m}\sum_j \frac{1}{s_j^B} \text{ and } \Theta = \frac{1}{m}\sum_j \frac{s_j}{s_j^B}$$

Now, in order to test the null hypothesis that a gene's biological CV^2 does not significantly surpass a chosen minimal value α_{th}, we proceed as follows. As $\hat{W}_i^B$ is a sampling variance calculated from m observations, we assume its sampling distribution to approximately be that of a χ^2 distribution with $m - 1$ degrees of freedom, scaled to the expected mean. If the gene's biological CV^2 were exactly α_{th}, this mean would be $\Omega(\alpha_{th}, \hat{\mu}_i^B)$. Hence, the right tail probability of this distribution provides P values p_i for a one-sided test of the null hypothesis $\alpha_i^B \le \alpha_{th}$.

$$p_i = 1 - F_{\chi^2_{m-1}}\left(\frac{(m-1)\hat{W}_i^B}{\Omega\left(\alpha_{th}, \hat{\mu}_i^B\right)}\right)$$

where

$$F_{\chi^2_{m-1}}$$

is the cumulative distribution function of the χ^2 distribution with $m - 1$ degrees of freedom.

GO analysis of high variance genes. We used TopGO[20] to find enriched Gene Ontology (GO) categories. In order to safeguard against confounding due to expression strength, we needed to exclude genes from the analysis universe for which inferential power to detect high variability was insufficient owing to high technical noise. Therefore, we included only genes with an average normalized read count of at least 200 for the GL2 cells and 600 for the QC cells. We ran TopGO in its "elimination mode" with Fisher's exact test and considered categories with an unadjusted P value below 10^{-5} as significant.

Cell capture and library preparation for mouse cells using the Fluidigm C1 system. 2,000 cells were loaded onto a 10- to 17-µm C1 Single-Cell Auto Prep IFC (Fluidigm), and cell capture was performed according to the manufacturer's instructions. The capture efficiency was inspected using a microscope, and there were single cells in 93 positions and two cells in three positions. These three positions were noted, and the data from these cells were subsequently removed from analysis.

Upon capture, reverse transcription and cDNA preamplification were performed in the 10- to 17-µm C1 Single-Cell Auto Prep IFC using the SMARTer PCR cDNA Synthesis kit (Clontech) and the Advantage 2 PCR kit (Clontech). 1 µl of the ERCC Spike-In Control Mix (Ambion) in a 1:400 dilution in C1 Loading Reagent was added to the lysis mix.

cDNA was harvested and diluted to a range of 0.1–0.3 ng/µl, and Nextera libraries were prepared using the Nextera DNA Sample Preparation Kit and the Nextera Index Kit (Illumina) by following the instructions in the Fluidigm manual "Using the C1 Single-Cell Auto Prep System to Generate mRNA from Single Cells and Libraries for Sequencing." Libraries were pooled, and paired-end 75-bp sequencing was performed on eight lanes of an Illumina HiSeq. All experiments involving mice were approved by the local ethical review committee, and a certificate of designation from the UK Home Office (the national authority for animal experimentation) was obtained.

Mapping of reads and normalization for the mouse data set (91 cells). Paired-end reads were mapped simultaneously to the *M. musculus* genome (Ensembl version 38.70) and the ERCC sequences using GSNAP (version 2013-02-05)[17] with default parameters. Two cells were removed at this stage owing to very low numbers of reads mapping to these libraries, which left 91 cells in total. From here we proceeded as described for the *A. thaliana* data.

15. Clough, S.J. & Bent, A.F. *Plant J.* **16**, 735–743 (1998).
16. Birnbaum, K. *et al. Nat. Methods* **2**, 615–619 (2005).
17. Wu, T.D. & Nacu, S. *Bioinformatics* **26**, 873–881 (2010).
18. Irizarry, R.A. *et al. Biostatistics* **4**, 249–264 (2003).
19. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
20. Alexa, A., Rahnenfuhrer, J. & Lengauer, T. *Bioinformatics* **22**, 1600–1607 (2006).

doi:10.1038/nmeth.2645