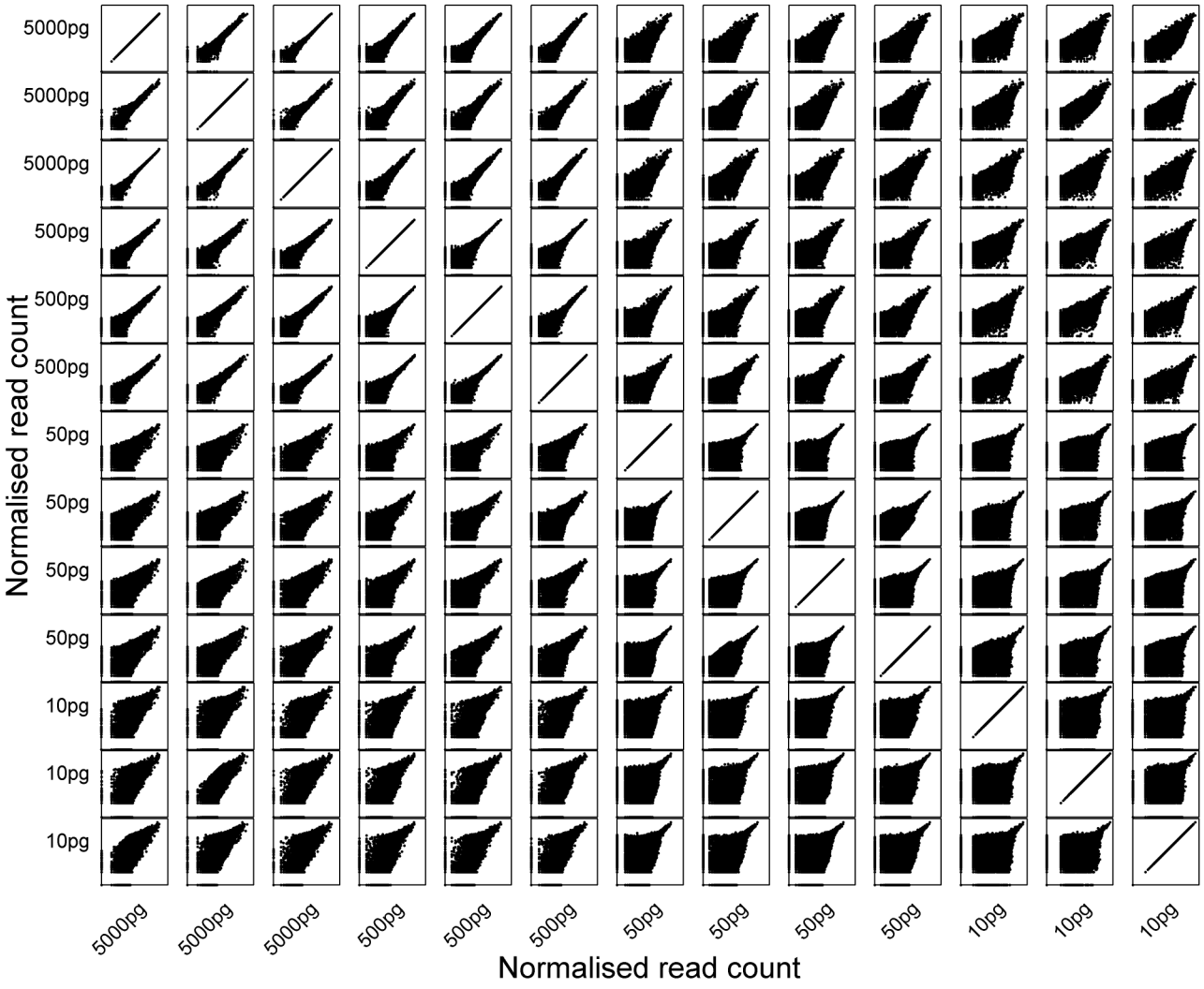
Supplementary Figures and Supplementary Notes for:

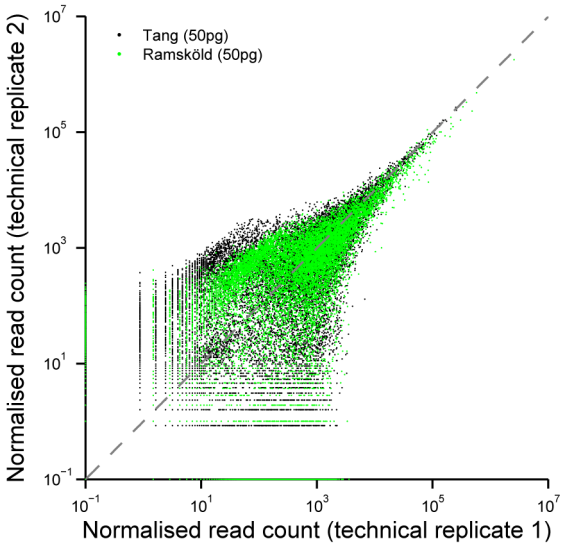
Accounting for technical noise in single-cell RNA-seq experiments

Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, and Marcus G Heisler

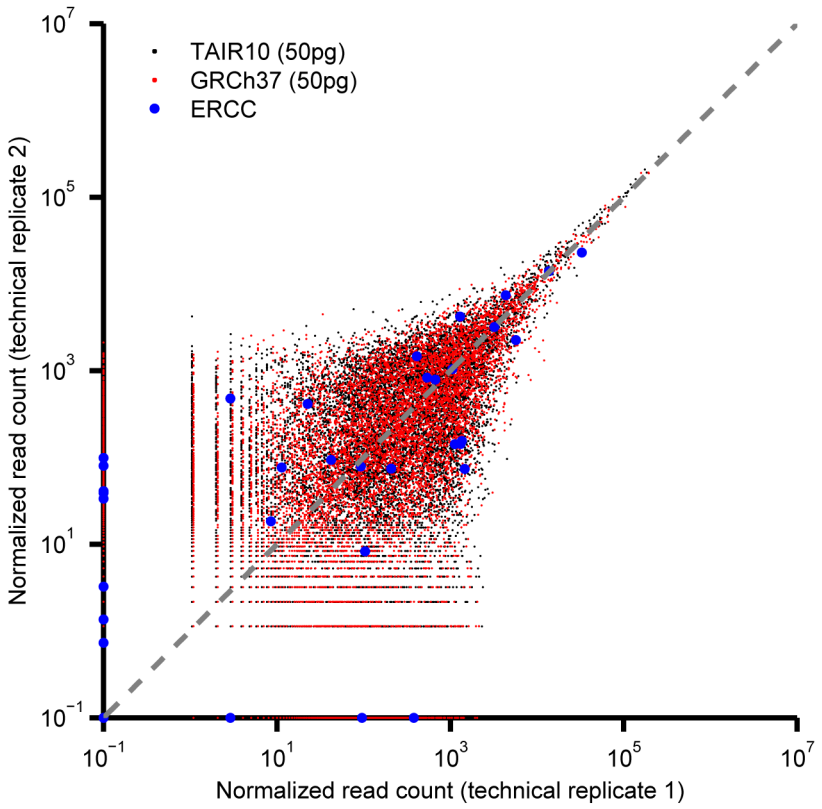
Nature Methods: doi:10.1038/nmeth.2645



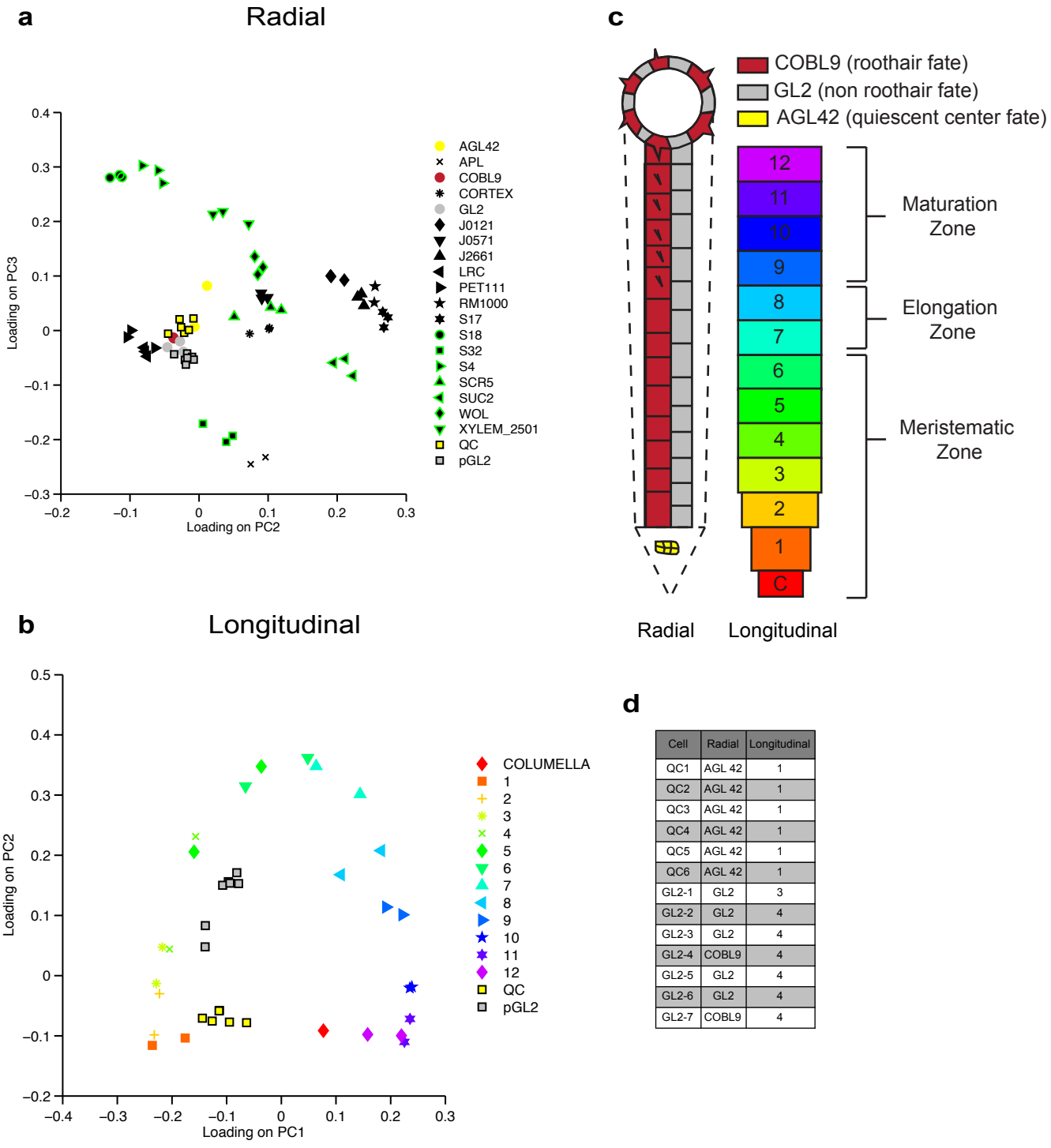
Supplementary Figure 1| Full data set of the *A. thaliana* total RNA dilution series. For details refer to legend of **Fig. 1**.



Supplementary Figure 2| Comparison of two single-cell RNA-seq protocols (Tang et al.'s versus Ramsköld et al.'s). Superimposed scatter plots of the two protocols using 50 pg *A. thaliana* total RNA as starting material. Both protocols show very similar noise profiles.



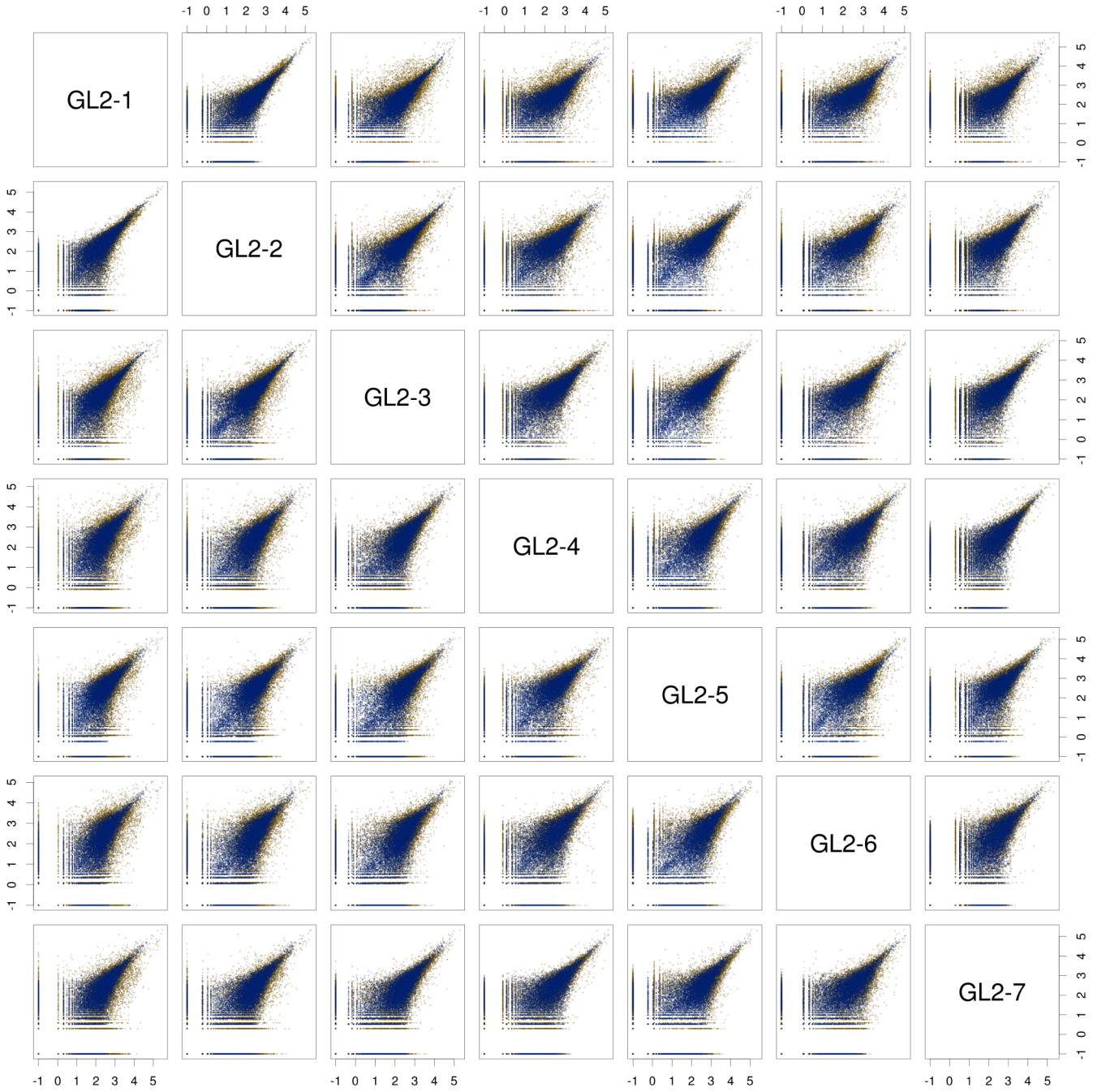
Supplementary Figure 3| Good agreement between *A. thaliana* and *H. sapiens* technical noise profiles. Technical replicates containing 50 pg of total HeLa RNA, 50 pg of total *A. thaliana* RNA and ERCC spike-ins at a 1: 1,000,000 dilution were sequenced. Reads were mapped to the TAIR10 and GRCh37 genomes and the ERCC sequences simultaneously. ERCC spike-ins are represented as blue dots. The axes are logarithmic with 10^{-1} representing zero reads.



d

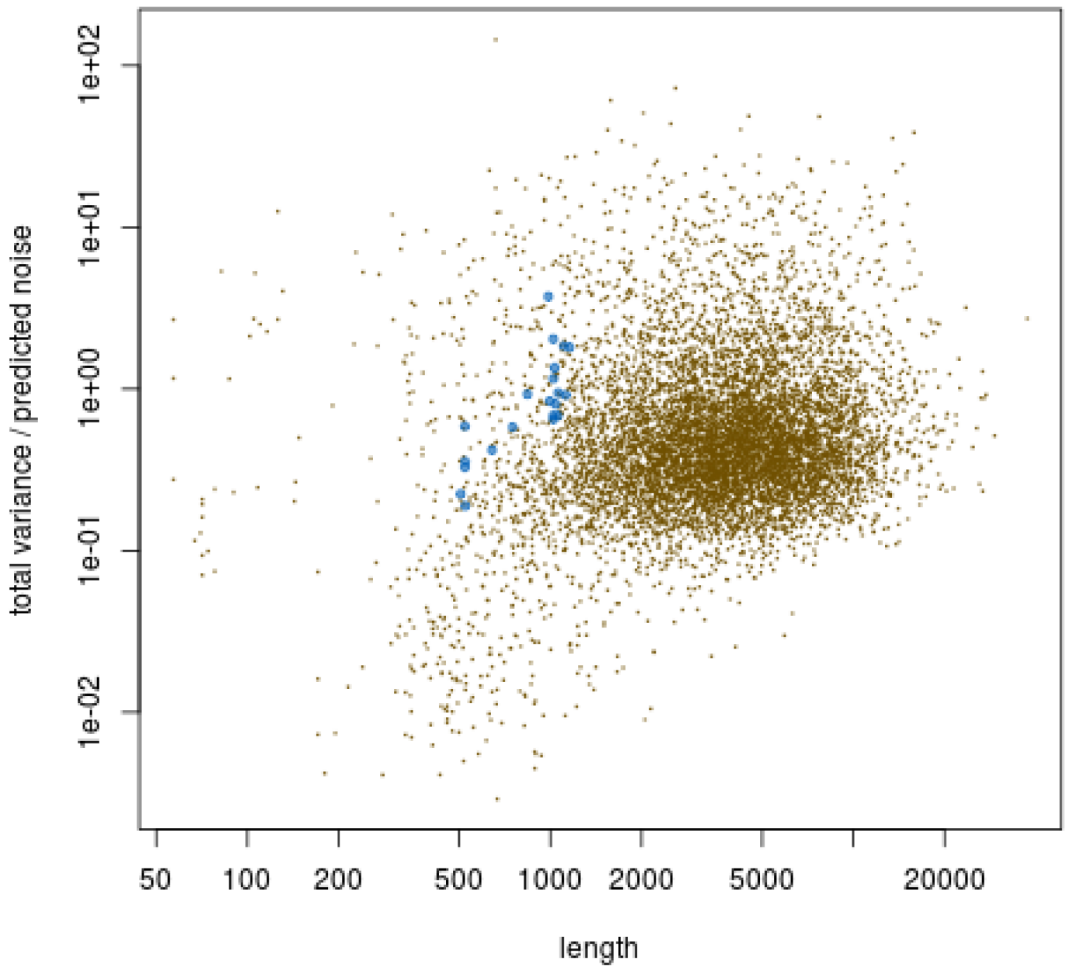
Cell	Radial	Longitudinal
QC1	AGL 42	1
QC2	AGL 42	1
QC3	AGL 42	1
QC4	AGL 42	1
QC5	AGL 42	1
QC6	AGL 42	1
GL2-1	GL2	3
GL2-2	GL2	4
GL2-3	GL2	4
GL2-4	COBL9	4
GL2-5	GL2	4
GL2-6	GL2	4
GL2-7	COBL9	4

Supplementary Figure 4| Cell-type mapping to Brady et al.'s atlas of the *A. thaliana* root. Brady et al. produced an atlas of the *A. thaliana* root by sorting cells using markers that were specific for certain radial and longitudinal positions. Here, we show a principal components analysis (PCA) of their expression micro-array data, combined with our single-cell data, for the radial **(a)** and longitudinal **(b)** markers. We assigned each of our cells a marker type in the atlas using k-nearest-neighbors clustering. The colored symbols represent Brady et al.'s data, the gray boxes our individual GL2 cells and the yellow boxes our individual QC cells. Note that in the atlas of Brady et al. each marker type is represented by three biological replicates. Panel **(c)** illustrates the spatial position implied by the mappings and panel **(d)** shows the mappings we inferred in this manner. For details regarding the mapping refer to the **Online Methods**.

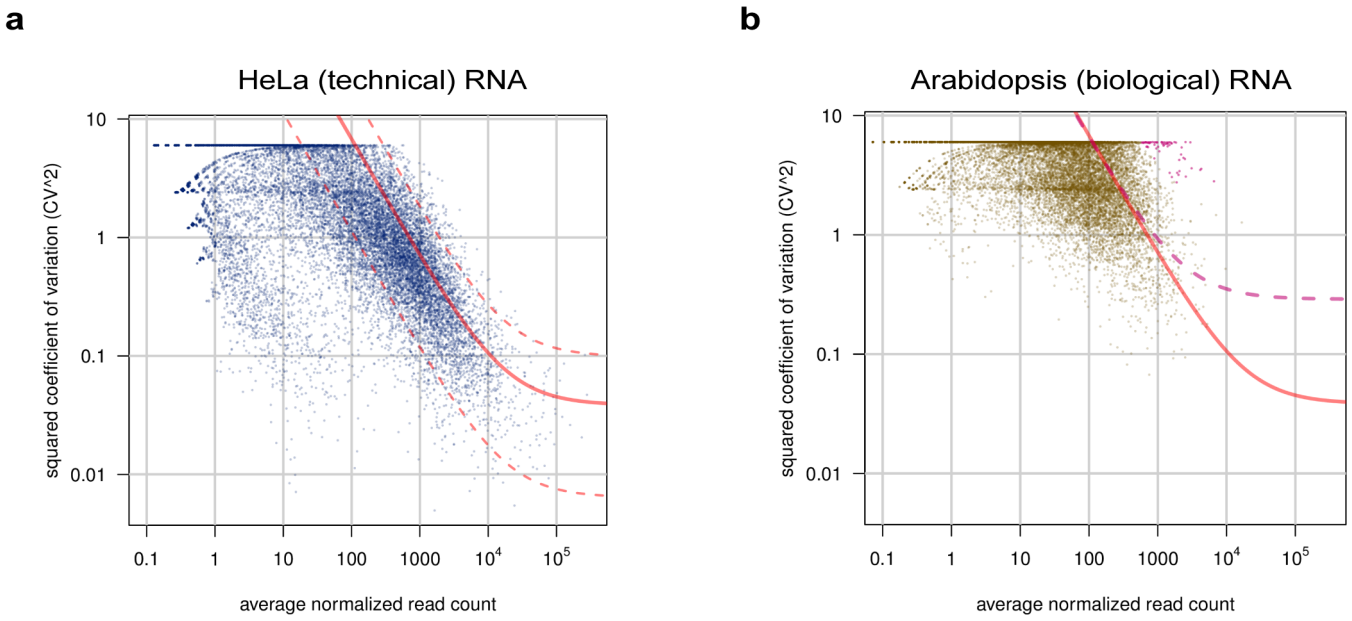


Supplementary Figure 5| Scatter plots for all seven GL2 cells. For detailed description refer to figure legend of **Fig. 2 a and b**. Gold and blue points represent plant and human genes, respectively. The axes are logarithmic (base 10) with -1 representing zero reads.

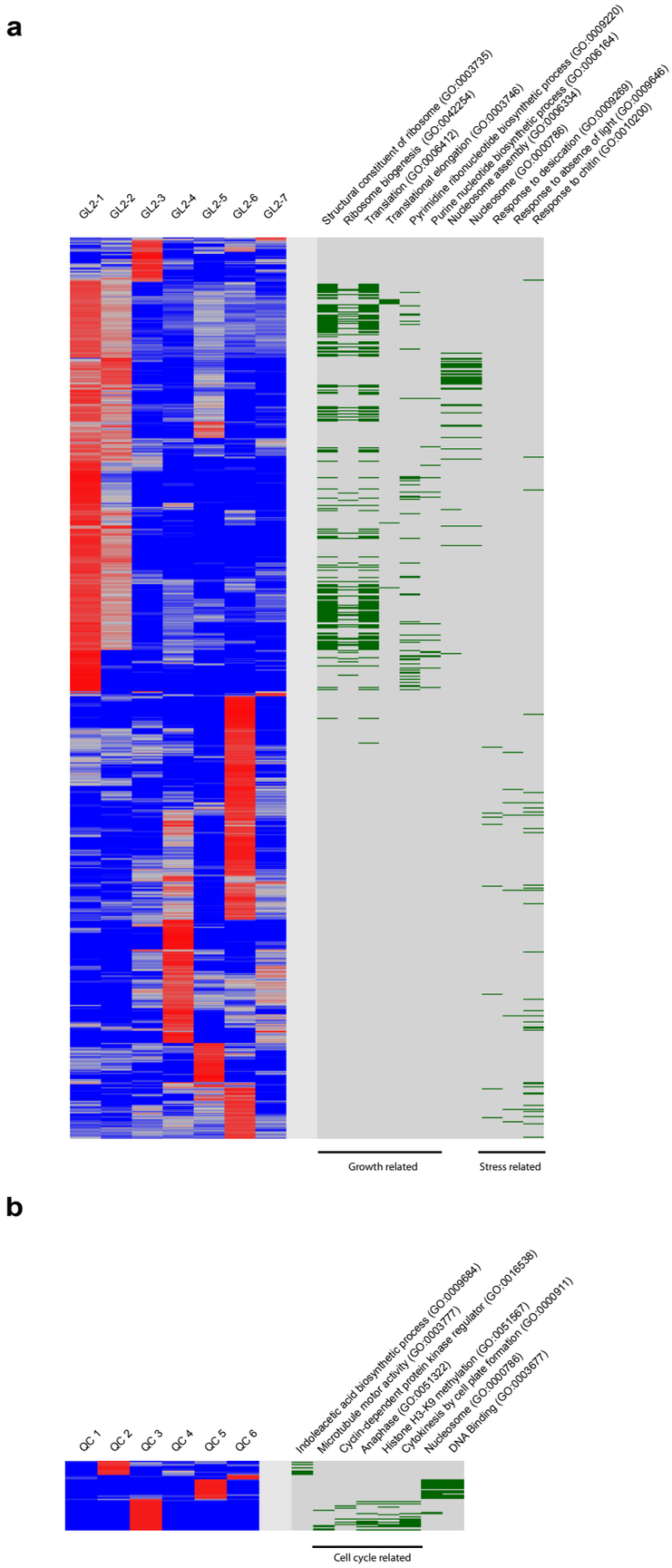
Supplementary Figure 6| Scatter plots for all six QC cells. For detailed description refer to figure legend of **Fig. 2a and b**. Gold and blue points represent plant and human genes, respectively. The axes are logarithmic (base 10) with -1 representing zero reads.



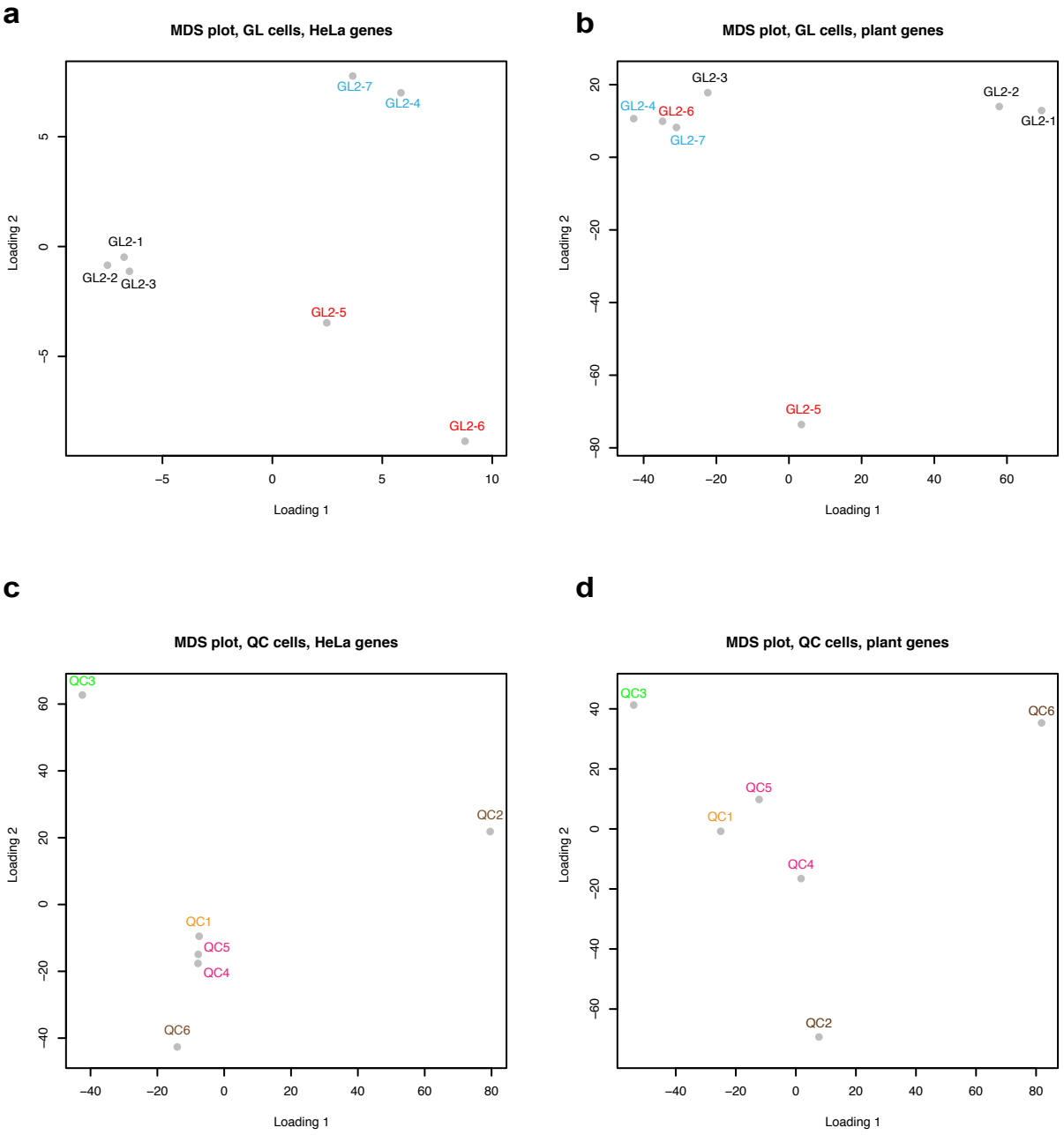
Supplementary Figure 7| Ratio of observed total variance to predicted technical noise, plotted against transcript length, for the mouse cell data. Brown dots are mouse genes, blue circles are ERCC spike-ins. Genes and spike-ins with very low read count (average normalized counts < 81) are excluded from the plot. See **Supplementary Note 5** for a discussion.



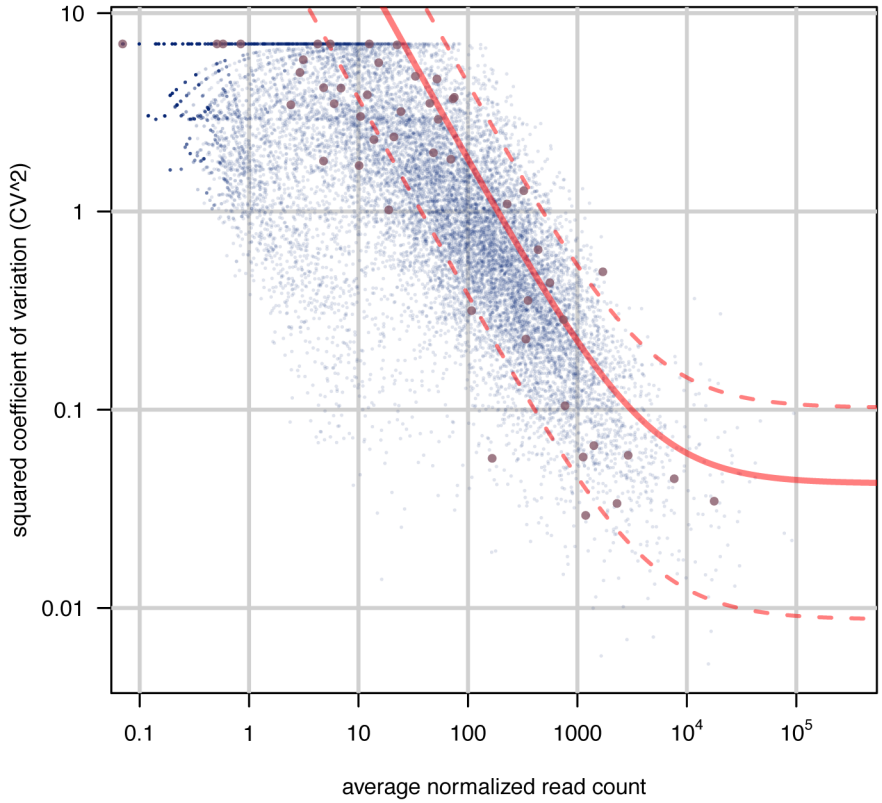
Supplementary Figure 8| Technical noise fit and identification of highly variable genes for the QC cells (for details refer to figure legend of **Fig. 2c-d**).



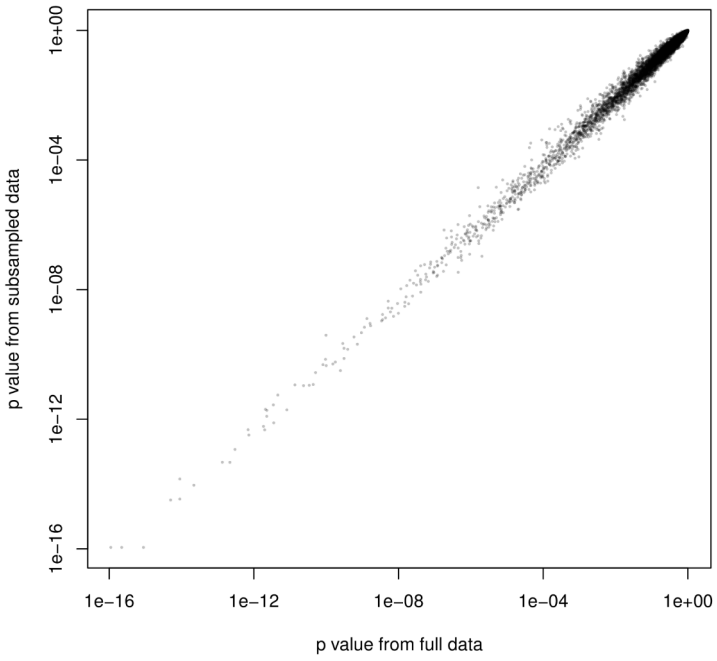
Supplementary Figure 9| Individual cellular expression profiles and co-varying genes. Panels **(a)** and **(b)** show the expression of highly variable genes in individual GL2 and QC cells, respectively. Each blue/red column represents a cell, each row a gene. The color indicates the log ratio of the gene's expression in a given cell to the average over all cells. Blue represents lower-than-average, gray represents average and red represents higher-than-average expression. Rows are ordered by clustering according to the expression ratios relative to the averages. The green-on-gray columns show genes in green that are part of the indicated GO categories.



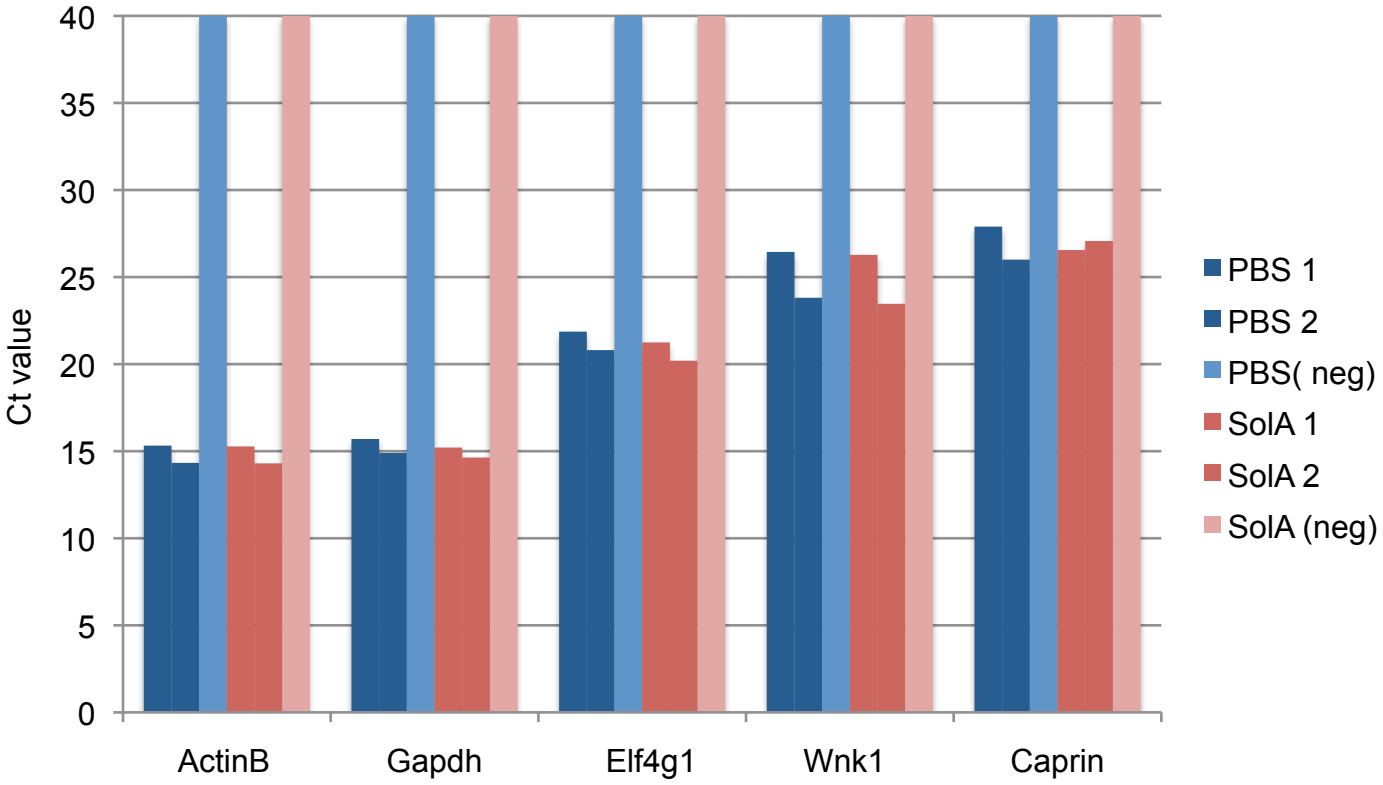
Supplementary Figure 10| Multidimensional scaling (MDS) plots of the normalized read counts from the GL2 (upper panels) and the QC cells (lower panels), showing the technical spike-ins (HeLa genes, left panels) and the biological material (plant genes, right panels). Different batches are represented as different colors. For a discussion see **Supplementary Note 8**.



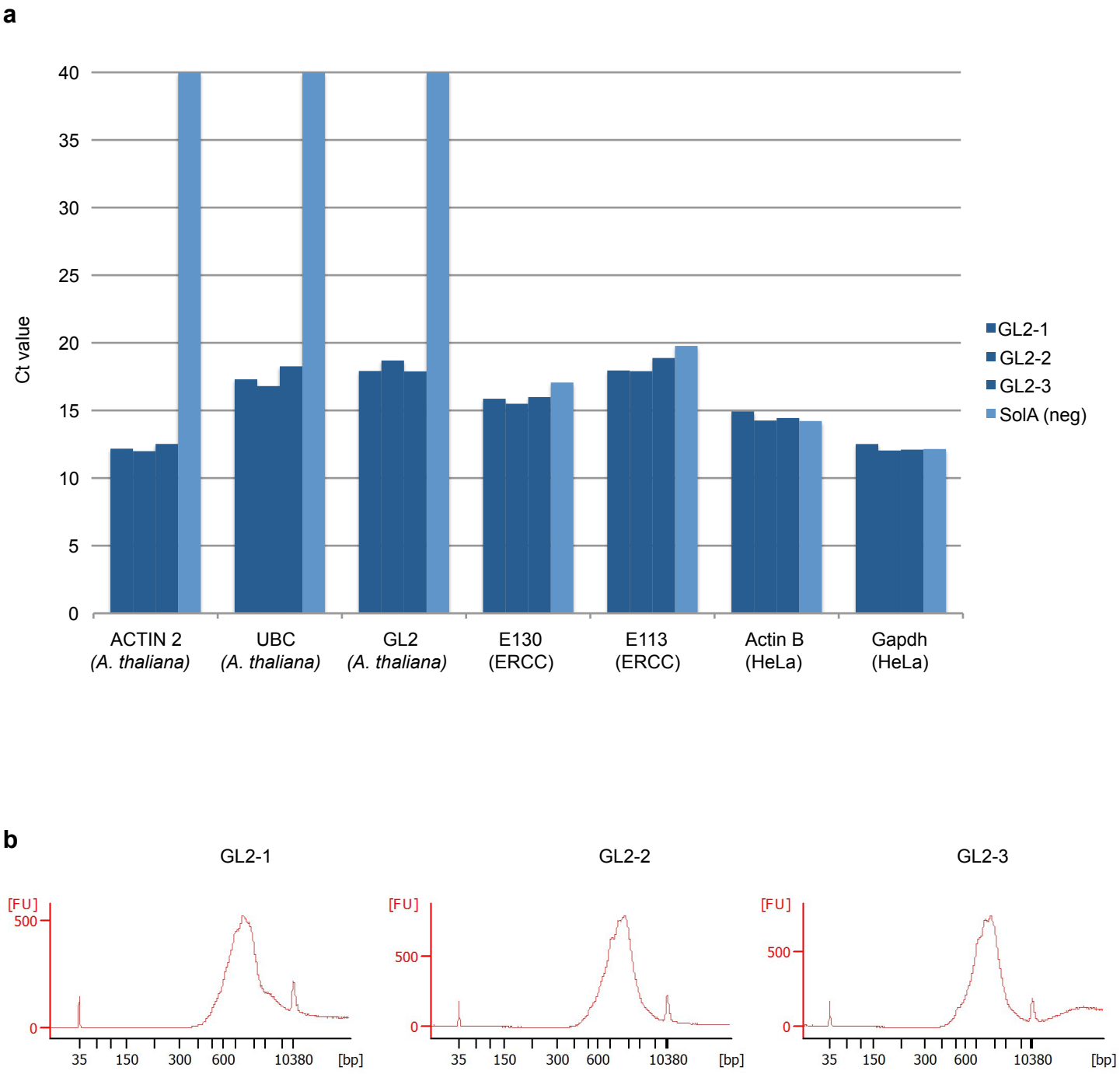
Supplementary Figure 11| The same plot as shown in **Fig. 2c**, with the ERCC spike-ins superimposed. While the many HeLa genes provide sufficient data for a good fit (solid line), the sampling distribution due to the modest number of cells is too wide (dashed lines) to obtain sufficient precision from only the limited number of ERCC spike-ins.



Supplementary Figure 12| Effect of sub-sampling on the *P* values from the test for high biological variability. *P* values from the full GL2 data set are on the x axis, those from the reduced data using only 20 % of the reads on the y axis. The good agreement demonstrates that a reduction of the sequencing depth to 20 % causes hardly any loss of inferential power.



Supplementary Figure 13| Solution A (SolA) does not affect performance of the Tang protocol. 50 pg of total HeLa RNA were amplified in duplicates using the Tang protocol. The RNA was either provided in 0.5 µl PBS or 0.5 µl SolA. A qPCR reaction was performed after the initial 24 cycles of library amplification PCR using several sets of qPCR primers as indicated in the figure. Negative controls failed to produce PCR products and Ct values were set to 40 for clarity. SolA shows the same performance as PBS and is not inhibiting the reaction.



Supplementary Figure 14| Control of cDNA library quality. **(a)** GL2 cell single-cell cDNA libraries were screened for known marker genes via qPCR after the first round of PCR amplification. *A. thaliana* genes, ERCC spike-ins belonging to different expression level groups and genes of the total HeLa RNA spike-in were checked. GL2 is a GL2 cell specific marker. ACTIN 2 is a highly and UBC a moderately expressed *A. thaliana* gene. Since the total HeLa RNA spike-in and the ERCC spike-ins were added to the master mix of the cell lysis buffer their PCR products also show up in the negative control whereas *A. thaliana* specific genes are absent in the negative control as expected. A Ct value of 40 corresponds to a failed amplification (Ct value set to 40 for clarity). **(b)** Bioanalyzer electropherogram of the final GL2 cell cDNA libraries after the second round of PCR amplification prior to Covaris shearing.

SUPPLEMENTARY NOTES

Supplementary Note 1 – Differences between groups versus variability within a group

The purpose of our method is to identify genes whose expression levels vary across single cells within a single population of cells. These cells are supposedly similar or, at least, they are not *a priori* known to come from two distinct cell populations. This scenario is significantly different from the more common experimental setup of finding genes that are differentially expressed between two or more discrete groups of cells.

In the latter setting, one calculates the difference between the average expression of a gene in one group and the average in the other group and wishes to show that this difference is large compared to the variability within a group. Such a comparison can be performed with any method suitable for differential expression analysis of ordinary, non-single-cell, RNA-Seq data, e.g., edgeR[21] or DESeq[19] or even simply a *t* test. Such a test will account for the total within-group variability, which comprises both the contributions from technical noise and from biological cell-to-cell variability, and there is no need to assess the extent to which either part contributes to the total variability. Consequently, such a test does not need technical noise estimates derived from technical spike-in data.

In our setting, however, we seek to find genes that are variable within a single population of cells. In other words, the biological variability, which was part of the nuisance parameter in the two-group comparison setting, now becomes the parameter of interest. Hence, distinguishing biological noise from technical noise is critical, and only in this situation does it become necessary to resort to spike-in data to characterize the strength of technical noise.

Supplementary Note 2 – Comparison of two single-cell RNA-seq protocols

The presence of strong technical noise is not a shortcoming of our experimental work. Rather, technical noise affects all work published on single-cell RNA-seq to date, and all but the earliest publications examined its strength to some extent and found it to be considerably high[1-3]. To confirm that the protocol we applied to the *A. thaliana* cells, a

slightly modified version of the one by Tang et al.[4,5], is comparable to other protocols in its noise strength we also adapted the protocol of Ramsköld et al.[3] and used it to prepare two more technical replicate libraries from 50 pg aliquots of total RNA (**Online Methods**).

Focusing only on gene expression measurements, the correlation coefficients between our replicates generated using the Ramsköld protocol are comparable to those reported by Ramsköld et al. for their technical replicates, indicating that our modifications to their protocol did not diminish its performance. Importantly, the overall strength of technical noise in the data generated using the Tang and Ramsköld protocols and its dependence on count levels is very similar (**Supplementary Fig. 2**). We note that one major aim of Ramsköld et al.'s protocol was to improve uniformity of coverage across the whole length of transcripts to allow for better characterization of isoform usage, a feature that we are not trying to implement in our specific application. Therefore, we used the protocol by Tang and coworkers, which is substantially cheaper, for all subsequent experiments.

Supplementary Note 3 – Mapping of single-cell transcriptomes

We mapped each single-root-cell transcriptome to the gene expression atlas of the *A. thaliana* root generated by Brady et al.[12] using a clustering approach (see **Online Methods**) in order to demonstrate that our biological data has the same quality as that in other studies, which also demonstrated that the cell type of single cells could be determined using single-cell RNA-seq[1-3].

We found that all six QC cells mapped to the predicted regions in both dimensions of the root atlas, namely the AGL42 region for the radial dimension and Zone 1 for the longitudinal dimension (**Supplementary Fig. 4**). Five of the seven GL2 cells mapped, as expected, to the GL2 region, while the two remaining cells mapped to the nearby COBL9 region that specifies hair cell fate in the radial dimension of the atlas (**Supplementary Fig. 4**). Upon closer inspection, however, we found GL2 to be expressed and not COBL9 in both these cells suggesting that the mis-mapping is due to the very similar molecular fingerprints of these cell types. In the longitudinal dimension of the atlas the GL2 cells behaved as expected. All seven GL2 cells mapped to the

restricted area of the root they were picked from, namely Zones 3 and 4. Taken together the mapping results suggest that our data are of high quality.

Supplementary Note 4 – On the need to account for cell-to-cell differences in RNA yield

A core starting point of the present paper is the claim that despite recent progress in single-cell RNA-seq protocols, reliable quantification of expression strength, and especially reliable identification of cell-to-cell variability in expression levels is possible only for genes that are expressed with sufficient strength. This may seem to contradict previous work, which described a single-cell RNA-seq experiment where biological variance appeared to always exceed technical noise[2]. In the following, we show that this observation can be explained by a careful consideration of a subtle but important point concerning normalization.

For each cell j, we calculated two normalization constants ("size factors"), one for the technical (i.e., HeLa) genes, denoted s_j, and one for the biological (i.e., plant) genes, denoted s_j^{B} (see **Online Methods**). These size factors are meant to provide a scaling normalization, i.e., dividing the counts by the appropriate size factors brings them onto a common scale that allows comparison across samples. Since total HeLa RNA (representing "technical" genes) is spiked in at the same amount in each sample, the normalization constant for the technical genes, s_j, is simply an estimate of relative sequencing depth. The amount of biological RNA, however, differs from sample to sample, due to possible differences in the efficiency of cell lysis and also simply due to differences in cell size and total RNA content of each cell. Now consider an ideal housekeeping gene, i.e., a gene that has the same expression (relative to the other genes) in all cells: If its average count over all samples is expected to be μ, the expected read count in a specific sample, j, is $\zeta_j s_j \mu$, where the factor ζ_j is proportional to the total amount of biological RNA in sample j. The factor ζ_j can be estimated as the ratio s_j^{B} / s_j of the biological to the technical size factor for sample j. In our approach, we always compute the sample variance of the *normalized* counts, i.e., the read count divided by the appropriate size factor, i.e., s_j for technical genes (e.g. for **Fig. 2c**) and s_j^{B} for biological genes (e.g., for **Fig. 2d**).

By contrast, Islam and coworkers used eight calibrated spike-ins to calculate how many read counts were obtained from a single mRNA molecule and this was then used to calculate for each gene a "copy number", i.e., an estimate of the number of initial mRNA molecules for each gene that were present in the sample[2]. This is akin to a normalization with what we termed the technical size factor: Even for an ideal housekeeping gene, whose mRNA concentration is exactly the same in all cells, the absolute number of mRNA molecules will be proportional to the cell's mRNA content and cell lysis efficiency, while this will not be the case for the spike-ins. Hence, the observation that even very weakly expressed genes from the single cells show stronger variance than technical spike-ins of the same abundance is not necessarily biologically meaningful – it could also be explained as merely the effect of variation in cell size or cell lysis efficiency. In contrast, our method of estimating two size factors allows the separation of these global per-cell properties from the true variation of individual genes.

Supplementary Note 5 – Influence of transcript length

One challenge in interpreting single-cell RNA-sequencing data, or RNA-sequencing data more generally, is to account for transcript length. The basic problem is very straightforward – if two genes are expressed at the same level but one is twice as long as the other, it is expected that twice as many reads would come from the longer gene as compared to the shorter one. An intuitive way of correcting for this is to divide the number of counts obtained for a certain gene by a suitable estimate of the transcript length before performing the analysis.

Such a normalization is based on the assumption that reads are sampled uniformly from across the whole length of the transcript, because only then, read count can be expected to be proportional to transcript length. However, even though significant progress has been made recently, coverage in single-cell RNA-Seq is still not as uniform as it is in bulk RNA-Seq methods, and it is therefore unclear whether a division by transcript length is helpful or not. In this note, we therefore compare these two possible approaches.

In our experiments, to generate the plant RNA-seq data, we used the well-established protocol of Tang and coworkers[4,5]. Whilst this protocol provides a robust

method for quantifying gene expression levels (the principle focus of our study) it has been reported to have a pronounced 3' bias[3]

The mouse data has been generated using the Fluidigm C1 machine, which is currently only compatible with the SMARTer protocol[3] This protocol provides a much better coverage across the full length of the transcripts and so allowed the analysis of alternative splicing and full length-transcripts in a substantial part of the multi-exon genes[3] Nevertheless, there still is a noticeable peak in the number of reads within the first 200–500 bases of each transcript, which levels off only after approximately 1000 bases (Ramsköld et al.[3], their Figure 1a), i.e., coverage is still notably stronger at the 3'end then across the rest especially of longer transcripts

For data from both protocols, we investigated the effect of transcript length (see **Supplementary Tables 5** and **9**) by dividing gene counts by transcript length before performing the downstream analysis, and then comparing the results thus found with that of the analyses presented in the main text, where we did not divide by transcript length. We first examined how much of the variance of the logarithms of the CV^2 values for the technical genes could be explained by the fit, once without and once with division of counts by length.

For the GL2 cell data, regressing CV^2 values of the "technical HeLa genes" on counts divided by length gave a worse fit than regressing on counts without accounting for length. For the mouse cell data, where the ERCC spike-ins were used as "technical genes", the division by length, however, improved the quality of the technical noise fit.

More specifically, for the GL2 cells, the fit explained 58% of the explainable variance of the log CV^2 values of the technical genes. (By "explainable variance", we mean the fraction of the total variance not due to the estimators sampling variance; see **Supplementary Note 6**.) When dividing counts by length prior to performing the regression, only 34% of the variance was explained. By contrast, for the mouse data, the fit without accounting for length explains 79% of the variance of the log CV^2 values of the ERCC spikes. Once we divide by length, however, the fit improves somewhat, now explaining 89% of the variance.

Part of this observed difference can be attributed to differences in the protocols noted above. Given the pronounced 3' bias of the Tang protocol there is no reason to

expect the number of reads derived from a typical gene to be proportional to the transcript length. By contrast, for the SMARTer protocol, this 3' bias is less pronounced and hence correcting for transcript length could be expected to improve the fit.

However, surprisingly, when we attempted to identify highly variable genes using the length corrected mouse data, the number of significant genes fell from 1198 to 523 – a very substantial decrease. This counter-intuitive result can be explained by observing **Supplementary Fig. 7**.

In this figure we plot gene length on the x-axis against the ratio of observed variance to predicted technical noise (on the log-scale) on the y-axis, where the predicted technical noise is obtained from the fit without division by length. The brown points correspond to mouse genes, the blue ones to the ERCC spike-ins. Only data from genes and spike-ins with a mean normalized count above 81.3

We observe that, in general, the observed and predicted values are roughly similar. However, for genes < ~1000 bp in length there is a systematic tendency to overestimate the dispersion. Importantly, a number of the ERCC spike-in genes used in the fit fall into this region – consequently, correcting for this improves the fit.

However, it is equally apparent that for genes > 1000 bp in length there is no relationship between gene length and the distance between the observed and predicted noise. Consequently, for genes > 1000 bp in length, normalizing by length will have a detrimental effect: it will effectively under-estimate the number of reads mapped to these genes, thus leading to a reduction in power for identifying highly variable genes. The likely reason for this is that, although the SMARTer protocol is superior to the Tang protocol regarding coverage across the transcript, it still has a relatively strong 3' bias.

Supporting this observation, when we adjust the counts for genes < 1500 bp for length and leave the counts for the remaining genes unaltered before identifying highly variable genes, we identified 1269 significant genes as being highly variable. Importantly, these genes had > 80% overlap with the 1198 genes called highly variable without using a length correction. This suggests that the benefits of using a length correction for quantifying gene expression measurements using single-cell RNA-seq data might be of somewhat questionable benefit at present but will become important once full-length coverage increased markedly.

However, as protocols keep improving it is likely that correcting for length will significantly improve results in the near future and this should be kept in mind by the practitioner.

Nature Methods: doi:10.1038/nmeth.2645

Supplementary Note 6 - Derivation of the statistical method

Here, we describe in detail the statistical method used to identify genes that show more biological variability across cells than is expected by chance.

The basic goal is to quantify the amount of variation present from the "technical RNA" (i.e., the RNA spiked in from the bulk mix) and to compare this to the variation present in the "biological RNA" where this refers to the material collected from each of the single cells assayed. In what follows we let K_{ij} denote the number of reads mapped to technical gene i from sample j, and K^{B}_{ij} is the read count for biological gene i.

Modelling the technical data

The key assumptions about the "technical RNA" data are that:

1. The expected ratios of transcript concentrations are the same in all samples since the aliquots come from the same bulk RNA sample

2. The absolute amount of "technical RNA" present in each sample is the same since a fixed aliquot volume is pipetted into each sample

We denote by μ_i a measure of abundance of transcripts from gene i, in suitable units, in the initial bulk mix of technical RNA. Due to sampling effects in diluting and aliquoting and due to variations in sequence-specific efficiencies of the steps of preparing library j, the actual concentration Q_{ij} of sequenceable transcript fragments from gene i in the specific library j scatters around its expectation, $\mathrm{E}\,Q_{ij} = \mu_i$. (This implies that all sequence-specific biases affecting the number of reads gained from a transcript molecule that do not vary from sample to sample are considered absorbed in μ_i.)

The variance of Q_{ij} is hence a measure of the strength of technical noise affecting gene i. We postulate that this variance is mainly determined by the mean μ_i, and that the variance-mean dependence can be parametrized as

$$\mathrm{Var}\,Q_{ij} = \tilde{a}_1\mu_i + \alpha_0\mu_i^2. \tag{1}$$

We justify this assumption later by examining the goodness of this fit.

To account for differences in sequencing depth, we introduce "size factors" s_j and model the sequencing of the prepared library j as a Poisson process (as justified by the observation of

Ref. 11 that sequencing noise shows only negligible overdispersion):

$$K_{ij}|Q_{ij} \sim \mathrm{Pois}(s_j Q_{ij}).$$

Marginalizing over Q_{ij}, we get

$$\mathrm{E}\, K_{ij} = s_j \mu_i \tag{2}$$

$$\mathrm{Var}\, K_{ij} = s_j(1 + s_j \tilde{a}_1)\mu_i + s_j^2 \alpha_0 \mu_i^2. \tag{3}$$

Fitting the model for the technical data

To estimate the size factors s_j, we use the estimator from the DESeq method, i.e., we set

$$s_j = \mathrm{median}_i \frac{K_{ij}}{\left(\prod_{j'=1}^{m} K_{ij'}\right)^{1/m}} \tag{4}$$

See Ref. 19 and Supplementary Note 1 of Ref. 22 for further explanations. As the size factor estimator pools information from many genes, we consider its sampling variance negligible and treat the size factors s_j as known constants in the following.

To fit the model parameters $\tilde{a}_1$ and α_0 of the technical mean-variance relation, we first calculate for each gene the sample mean and variance of the normalized counts K_{ij}/s_j, i.e., the quantities

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

and

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

Some algebra shows that, with Equations (2) and (3), this estimator has expectation

$$\mathrm{E}\, \hat{W}_i = (\Xi + \tilde{a}_1)\mu_i + \alpha_0 \mu_i^2 \qquad \text{with } \Xi = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{s_j}. \tag{5}$$

As we intent to regress $\hat{W}_i$ on $\hat{\mu}_i$ rather than μ_i, it is helpful to express $\mathrm{E}\, \hat{W}_i$ in terms of $\hat{\mu}_i$.

Of course, $\mathrm{E}\,\hat{\mu}_i = \mu_i$, and with Equation (3), one finds that

$$\mathrm{E}\,\hat{\mu}_i^2 = \mu_i^2 \left(1 + \frac{\alpha_0}{m}\right) + \mu_i \frac{\Xi + \tilde{a}_1}{m}.$$

With this, Equation (5) becomes

$$\mathrm{E}\,\hat{W}_i = \frac{1}{1 + \frac{\alpha_0}{m}} \mathrm{E}\left[(\Xi + \tilde{a}_1)\hat{\mu}_i + \alpha_0 \hat{\mu}_i^2\right] \tag{6}$$

As typically $\alpha_0 \ll 1$, the corrective factor in front of the square brackets will usually be quite close to 1, and we will hence neglect it.

We use $\hat{w}_i := \hat{W}_i / \hat{\mu}_i^2$ as a plug-in estimator for the squared coefficient of variation (CV^2). If $\hat{w}_i \ll 1$, the true CV^2 will be small, too, and so will be the estimation error of $\hat{\mu}_i$. Hence, $\hat{w}_i$ can be considered reasonably unbiased as long as it is sufficiently small, with $\mathrm{E}\,\hat{w}_i \approx (\Xi + \tilde{a}_1)/\hat{\mu}_i + \alpha_0$.

Therefore, we regress $\hat{w}_i$ on $1/\hat{\mu}_i$ to obtain estimates for the parameters of the variance-mean relation. We use the intercept returned by the regression as an estimator for α_0 and the coefficient for $1/\hat{\mu}_i$ as estimator for $a_1 = \Xi + \tilde{a}_1$. We denote these fitted coefficients as $\hat{\alpha}_0$ and $\hat{a}_1$.

As $\hat{w}_i$ is derived from a variance-like quantity ($\hat{W}_i$ is a sum of squares of random variables), its sampling distribution may be expected to be approximately χ^2. This is why we use a generalized linear model of the gamma family to perform the regression.

We expect the residuals $\hat{w}_i / (\hat{a}_1/\hat{\mu}_i + \hat{\alpha}_0)$ to follow approximately a $\chi^2_{m-1}/(m-1)$ distribution. In **Fig. 2c**, we indicate the 2.5- and 97.5-percentile of the χ^2_{m-1} distribution, scaled to $(\hat{a}_1/\hat{\mu}_i + \hat{\alpha}_0)$, with dashed lines, to show that this expectation is met reasonably well. At least in the right half of the plot, where the CV^2 stays well below its maximum, most points are in fact within this 95% interval. This confirms that our parametrization is suitable and that our model offers a good fit.

It is important to verify the quality of the fit, because a lack of fit, causing uncertain prediction of tecnical noise for the biological genes, can compromise type-I error control. We recommend to check how much of the variance of the $\log \hat{w}_i$ values is explained by the fit and how much by the sampling variance of the estimator (using the fact that $\mathrm{Var} \log \hat{w} \approx 2/(m-1)$), and ensure that the residual fraction is reasonably small.

Modelling the biological data

We denote by R_{ij} the concentration of transcripts from biological gene i in the mRNA extracted from the cell assayed in sample j and write the first two moments of this quantity as $\mathrm{E}\, R_{ij} = \mu_i^{\mathrm{B}}$ and $\mathrm{Var}\, R_{ij} = \alpha_i^{\mathrm{B}}(\mu_i^{\mathrm{B}})^2$, i.e., α_i^{B} is the squared coefficient of *biological* variation for gene i.

The amount of extracted mRNA differs from cell to cell, due to variation in the cell's total mRNA content and the efficiency of mRNA extraction. We denote the ratio of extracted biological mRNA to the amount of spiked in technical mRNA by ζ_j. Thus, the concentration Q_{ij}^B of transcripts from biological gene i in the prepared library j has expectation $\mathrm{E}\, Q_{ij}^{\mathrm{B}} = \zeta_j \mu_i^{\mathrm{B}}$. The variation of Q_{ij}^B due to *technical* noise (i.e., conditioned on the value R_{ij} that contains the biological variation) is found by substituting $\mathrm{E}\, Q_{ij}^{\mathrm{B}} = \zeta_j \mu_i^{\mathrm{B}}$ for $\mathrm{E}\, Q_{ij} = \mu_i$ in Equation (1):

$$\mathrm{Var}\, Q_{ij}^{\mathrm{B}} | R_{ij} = \tilde{a}_1 \zeta_j \mu_i^{\mathrm{B}} + \alpha_0 \zeta_j^2 (\mu_i^{\mathrm{B}})^2.$$

Marginalizing out gives the expression including both biological and technical variation:

$$\mathrm{Var}\, Q_{ij}^{\mathrm{B}} = \tilde{a}_1 \zeta_j \mu_i^{\mathrm{B}} + \zeta_j^2 \left(\alpha_0 + \alpha_i^{\mathrm{B}} + \alpha_0 \alpha_i^{\mathrm{B}}\right) (\mu_i^{\mathrm{B}})^2.$$

The actual counts values are modelled, as before, as $K_{ij}^{\mathrm{B}} | Q_{ij}^{\mathrm{B}} \sim \mathrm{Pois}(s_j Q_{ij}^{\mathrm{B}})$.

Testing for high variance

Applying Equation (4) to the read counts K_{ij}^{B} from the biological genes yields the biological size factors s_j^{B}, which account for the combined effect of starting amount of biological mRNA and of sequencing depth, i.e., via $s_j^{\mathrm{B}} = \zeta_j s_j$, they give an estimate of ζ_j.

We again consider the sample variance of the normalized counts, $K_{ij}^{\mathrm{B}}/s_j^{\mathrm{B}}$, i.e., $\hat{W}_i^{\mathrm{B}} = \frac{1}{m-1} \sum_{j=1}^{m} \left(\frac{K_{ij}}{s_j^{\mathrm{B}}} - \hat{\mu}_i^{\mathrm{B}} \right)^2$ with $\hat{\mu}_i^{\mathrm{B}} = \frac{1}{m} \sum_{j=1}^{m} \frac{K_{ij}}{s_j^{\mathrm{B}}}$.

A straight-forward calculation establishes that

$$\mathrm{E}\, \hat{W}_i^{\mathrm{B}} = \mu_i^{\mathrm{B}} \left(\Psi + a_1 \Theta\right) + (\mu_i^{\mathrm{B}})^2 \left(\alpha_0 + \alpha_i^{\mathrm{B}} + \alpha_0 \alpha_i^{\mathrm{B}}\right),$$

where

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

This expression is similar to Equation (5). This time, however, we cannot neglect the pre-

factor in Equation (6) any more, because α_i^{B} may be large.

Hence, we introduce the function

$$\Omega(\alpha, \mu) = \frac{\mu(\Psi + a_1\Theta) + \mu^2\alpha_{\mathrm{F}}}{1 + \frac{\alpha_{\mathrm{F}}}{m}} \text{ with } \alpha_{\mathrm{F}} = \alpha_0 + \alpha + \alpha_0\alpha,$$

which describes the expectation of the full sample variance,

$$\mathrm{E}\,\hat{W}_i^{\mathrm{B}} = \mathrm{E}\,\Omega\left(\alpha_i^{\mathrm{B}}, \hat{\mu}_i^{\mathrm{B}}\right),$$

as can be shown by a calculation analogous to the one for $\mathrm{E}\,\hat{W}_i$.

We now wish to test the null hypothesis that the biological CV does not exceed a certain threshold,

$$H_0 : \alpha_i^{\mathrm{B}} \leq \alpha_{\mathrm{th}}.$$

To construct a one-sided test, we determine the sampling distribution of $\hat{W}_i^{\mathrm{B}}$ for $\alpha_i^{\mathrm{B}} = \alpha_{\mathrm{th}}$. The expectation of this statistic is then, of course, $\Omega(\alpha_{\mathrm{th}}, \hat{\mu}_i^{\mathrm{B}})$. For the higher moments, we assume that $\hat{W}_i^{\mathrm{B}}$ takes the shape of a χ^2 distribution with $m-1$ degrees of freedom, scaled to the appropriate mean. This is reasonable because $\hat{W}_i^{\mathrm{B}}$ is the sum of the squares of m random variables, which should not be too far from normal. Therefore, a p value can be obtained from the CDF of the χ^2_{m-1} distribution, denoted here as $p_{\chi^2_{m-1}}$ by

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

Supplementary Note 7 – On "maximal" noise

In **Fig. 2c**, the CV^2 estimates for weak genes often hit a "ceiling", which the fit does not recapitulate and which hence needs to be briefly explained. For a given gene with average normalized read count μ, the maximum CV^2 is reached when all but one of the m cells have zero counts and one sample has a normalized count of $m\mu$, resulting in a variance of $m\mu^2$ and hence a CV^2 of m. A CV^2 estimate for m non-negative numbers hence will always be a value between 0 and m. This results in the hard upper boundary at a CV^2 of m=7 in the CV^2-mean plots in **Fig. 2c-d**. As **Fig, 2c** clearly shows, the technical noise reaches this maximum very frequently for genes with an average count below a certain threshold (e.g., ~100 reads for the GL2 cells). Consequently, it is impossible to attribute biological meaning to observed variation for *A. thaliana* genes with an average expression count less than this value.

Supplementary Note 8 – Large spike-in sets help to avoid false positives due to batch effects

We performed the experiments with the GL2 cells in three batches, the first batch comprising three cells, the other two batches two cells each. Since we did not perform all the cell lysis and library preparations in parallel, one might worry that small differences in the way the protocol was executed in each batch add additional variance, a very common issue in transcriptomics studies[23]. **Supplementary Fig. 10** shows a multi-dimensional scaling (MDS) plot for our experiments, using different colors for the batches. While the use of MDS to check for batch effects is a standard technique, the availability of very many technical spikes gives us the novel opportunity to perform the MDS analysis separately for the biological and the technical (i.e., total HeLa RNA spike-in) genes. In fact, the technical genes do cluster according to batches for the GL2 cells (**Supplementary Fig. 10a**), showing that we could not avoid batch effects completely. However, the much larger distances in the MDS plot for biological genes (**Supplementary Fig. 10b**; note that the axis scaling is different to **Supplementary Fig. 10a**) reassure us that the biological differences dominate the technical ones in this global view. For the QC cells, however, the sample-to-sample distances are similar in the technical (**Supplementary Fig. 10c**) and the biological (**Supplementary Fig. 10d**) genes,

consistent with our earlier observation that in these very small cells only a few very strongly expressed genes significantly exceeded the technical noise. On the other hand, we see less clustering by batches, showing that batch effects are not unavoidable.

In either case, it is crucial to notice that, in clear difference to the cases discussed by Leek et al.[23] the presence even of strong batch effects does not invalidate results obtained with our analysis method: Precisely because both the technical and biological genes are subject to the same batch effects their presence drives up the technical noise estimates, which is thus incorporated into the test for highly variable genes. Thus, batch effects do not cause false positives but merely reduce inferential power. In other words, our scheme offers an intrinsic safeguard against false conclusions due to batch effects.

Finally we note that in our GL2 experiment, all cells processed in the same batch were extracted from one plant, but different plants were used between batches. Consequently, one may wonder whether the highly variable genes really vary between different cells in the same plant or rather only between different plants. The fact that in **Supplementary Fig. 10b** the cells from the same batch, and hence from the same plant, do not cluster together is an important reassurance that the variability observed is really occurring within and not between plants. Nevertheless, future experiments using many more cells should use a design that introduces a hierarchy - several plants, and several cells from each plant - and further work is needed to extend our analysis scheme to fully distinguish variability between cells from the same sample from variability between samples.

Supplementary Note 9 - Sequencing deeply is not necessary

Our dilution series data suggest that the available starting material limits the sensitivity of single-cell assays and that sequencing depth is hence not necessarily a limiting factor. We used two full HiSeq lanes for our seven GL2 cells, obtaining 11 to 47 million reads per sample. To assess whether this sequencing depth was necessary we down-sampled our read count table by using only one fifth of all the reads (selected at random), and then reran our analysis pipeline on the reduced data. Using this reduced set of data we identified about the same number genes as being significantly highly variable (the exact number changed from run to run due to the random sampling, but was typically slightly higher, around 950 genes). Importantly, the list always had at least 90% overlap with the

876 genes found in the analysis of the full data. The *P* values changed only slightly (**Supplementary Fig. 12**). Moreover, the subsequent GO enrichment analyses found the same terms using both the full and sub-sampled data sets. As we would have reached essentially the same conclusions from only a fifth of all the reads, we conclude that there is currently little benefit in sequencing single-cell data to the same depth as is usually done for bulk RNA-seq assays. This is in agreement with previous work that has come to a very similar conclusion[3]. This observation is explained by the fact that only low-read count genes benefit from the reduction of Poisson noise achieved by deeper sequencing.

ADDITIONAL REFERENCES

21. McCarthy, D.J., Chen, Y. & Smyth, G.K. *Nucleic Acids Res* **40**, 4288-4297 (2012).

22. Anders, S., Reyes, A. & Huber, W. *Genome Res* **22**, 2008-2017 (2012).

23. Leek, J.T. et al. *Nature Reviews Genetics* **11**, 733-739 (2010).