BMC Bioinformatics

RESEARCH

Open Access

Leveraging gene correlations in single cell transcriptomic data

Kai Silkwood[1,2], Emmanuel Dollinger[1,2,3], Joshua Gervin[1,2], Scott Atwood[1,2], Qing Nie[1,2,3] and Arthur D. Lander[1,2*]

*Correspondence:
adlander@uci.edu

[1] Center for Complex Biological Systems, University of California, Irvine, Irvine, CA, USA
[2] Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA, USA
[3] Department of Mathematics, University of California, Irvine, Irvine, CA, USA

Abstract

Background: Many approaches have been developed to overcome technical noise in single cell RNA-sequencing (scRNAseq). As researchers dig deeper into data—looking for rare cell types, subtleties of cell states, and details of gene regulatory networks—there is a growing need for algorithms with controllable accuracy and fewer ad hoc parameters and thresholds. Impeding this goal is the fact that an appropriate null distribution for scRNAseq cannot simply be extracted from data in which ground truth about biological variation is unknown (i.e., usually).

Results: We approach this problem analytically, assuming that scRNAseq data reflect only cell heterogeneity (what we seek to characterize), transcriptional noise (temporal fluctuations randomly distributed across cells), and sampling error (i.e., Poisson noise). We analyze scRNAseq data without normalization—a step that skews distributions, particularly for sparse data—and calculate *p* values associated with key statistics. We develop an improved method for selecting features for cell clustering and identifying gene–gene correlations, both positive and negative. Using simulated data, we show that this method, which we call BigSur (Basic Informatics and Gene Statistics from Unnormalized Reads), captures even weak yet significant correlation structures in scRNAseq data. Applying BigSur to data from a clonal human melanoma cell line, we identify thousands of correlations that, when clustered without supervision into gene communities, align with known cellular components and biological processes, and highlight potentially novel cell biological relationships.

Conclusions: New insights into functionally relevant gene regulatory networks can be obtained using a statistically grounded approach to the identification of gene–gene correlations.

Keywords: Single cell RNA sequencing, Gene–gene correlation, Gene regulatory network, Gene co-expression network, Melanoma

Background

Single cell RNA-sequencing (scRNAseq), along with the related method of single nucleus RNA-sequencing, now offer researchers unparalleled opportunities to interrogate cells as individuals. Methods have been developed to classify cell types; identify gene expression markers; infer lineages; learn gene regulatory relationships, and examine the effects of experimental manipulations on both levels of gene expression and cell type abundances [1–7]. Because scRNAseq data are noisy, reliable inference requires leveraging information across many cells, trading off sensitivity for statistical power. How to handle that tradeoff should depend, ideally, on one's goal. Unsupervised clustering of large numbers of transcriptionally very different cells ("cell types") into small numbers of groups of similar size allows for a great deal of latitude in aggregating information across cells; it is thus not surprising that many different clustering approaches perform well. Other tasks, such as ordering cells along a continuum of gene expression change, or picking out rare cell populations within much larger groups, are less forgiving, and a plethora of different approaches currently compete for investigators' attention [8–23]. Assessing the performance of such methods is frequently hindered by a lack of knowledge of ground truth.

A particularly challenging application of scRNAseq is the identification of patterns of gene co-expression. The identification of large-scale blocks of co-expressed genes—co-expression "modules"—can provide an alternative method for classifying cells when traditional clustering fails [24]. In contrast, smaller-sized blocks of gene co-expression have the potential to reflect true gene-regulatory networks that relate to specific functions [25]. This is because random transcriptional noise in gene circuits should induce weak but real correlations among regulatory genes and their targets. Indeed, it has long been proposed that gene regulatory links could be discovered solely from the weak gene expression correlations that one might encounter when studying otherwise homogenous populations of cells [26–32].

Unfortunately, identifying small yet significant gene expression correlations in single cell data requires a degree of statistical power that scRNAseq applications rarely strive for (and, to be fair, rarely need to). Yet, as greater numbers of scRNAseq datasets accumulate, with a growing trend toward increasing numbers of cells per dataset, we wondered whether substantial amounts of novel information about gene co-regulation might be accessible simply through a more in-depth examination of pairwise gene expression correlations.

One of the main challenges in pursuing such a program is the absence of an accepted statistical model for pairwise correlations in scRNAseq data. Only with a model can one define a null hypothesis by which to judge whether observations are significant. Unfortunately, with scRNAseq, there is not good consensus regarding the model to use for the data distributions of individual genes, much less their correlations. The common approach of fitting individual gene data to ad hoc analytical distributions (e.g. "zero-inflated negative binomial" [33]; reviewed by [34]), has met with frequent criticism that is difficult to dismiss [35–38]. One may seek to circumvent such concerns by attempting to learn empirical distributions on a case-by-case basis from data, but this typically requires making assumptions about the amount and distribution of actual biological variation in the data, which are frequently unknown.

Furthermore, pitfalls in implementing empirical methods can be hard to avoid, particularly with high-order statistical information, such as correlations. For example, the seemingly reasonable intuition that one might be able to construct the distribution of the correlation coefficient under the null hypothesis simply by randomly permuting elements is actually incorrect [39]. So, as we will show below, is the intuition that one can improve the detection of gene correlations by averaging over groups of similar cells (i.e., creating "metacells") prior to calculation of correlation coefficients.

One of the major obstacles to defining an appropriate data distribution for scRNAseq data is the fact that underlying sources of technical variation are not fully understood, nor is the range of biological variation in biologically "equivalent" cells fully known. Here we begin by re-considering these factors, and leveraging the work of others, in pursuit of an analytical model of null correlation distributions that makes the fewest ad hoc assumptions and minimizes adjustable parameters. We show that the approach that emerges has the power to identify subtle yet real correlations, both positive and negative, in scRNAseq data, even among genes in modest numbers of cells that are relatively sparsely sequenced. Ultimately this method should be applicable not only to the identification of gene regulatory interactions, but to more complex tasks based on gene–gene correlations—such as the identification of cellular trajectories [40] and "tipping points" [41]—as well as providing a means to achieve a more principled approach to basic, early steps in scRNAseq analysis—such as normalization, batch correction, feature selection and clustering.

Results

The significance of gene–gene correlations

The statistical significance of correlations is rarely discussed because, for many common kinds of data—those that are continuous and at least approximately normal in distribution—the magnitude of correlation and its significance are related in a simple way that depends only on the number of measurements, and not the data distributions. Owing to Fisher [42], for any Pearson correlation coefficient (PCC), the p value (probability of observing $|\mathrm{PCC}| \geq x$ by chance) may be estimated as $Erfc\left[\sqrt{(n-3)/2}\, arctanh(|x|)\right]$, where n is the number of samples, and $Erfc$ is the complement of the Error function (we refer to this expression henceforth as the "Fisher formula").

Because scRNAseq data are both discrete and generally not normally distributed, p values obtained using the Fisher formula cannot be accepted as accurate, but just how far off will they be? Fig. 1 explores that question through simulation. Assuming Poisson-distributed data, and gene expression vectors of 500 cells in length, the formula does quite poorly for vectors of mean < 1, i.e., where the expected proportion of zeros exceeds 37%, assigning p values that are too low for positive correlations, and too high for negative ones (Fig. 1A). For distributions with mean ≥ 1 (fewer than 37% zeros on average), the formula does reasonably well down to p values as low as 10^{-4} but deviates progressively thereafter. This degree of accuracy would be a problem for any genome-wide analysis of correlations: To analyze pair-wise correlations among m genes one must test $m(m-1)/2$ hypotheses. With values of m often ≥ 12,000, this amounts to $> 7 \times 10^8$ simultaneous tests, such that statistical significance of any single observation could potentially

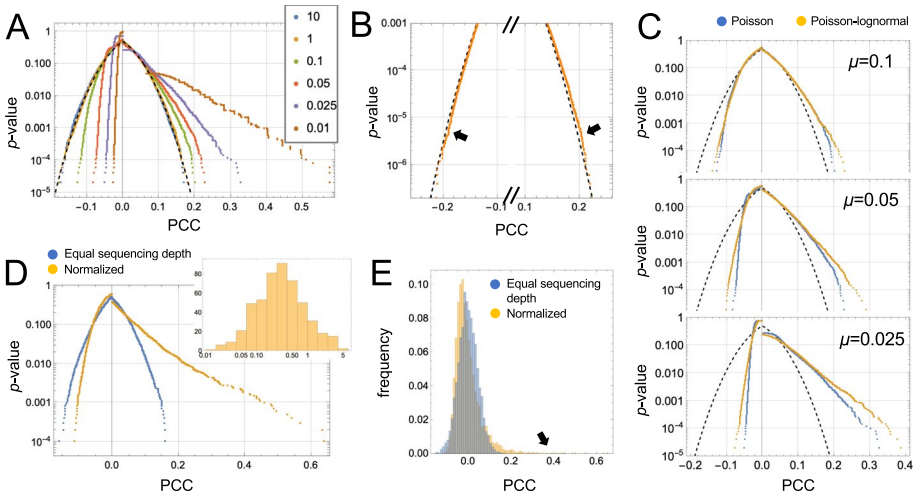


Fig. 1 Relationship between Pearson correlation coefficient, vector sparsity and *p* value. The panels compare *p* values determined empirically by correlating 50,000 pairs of random, independent vectors of length 500 with *p* values predicted by the Fisher formula. **A** Data were independent random variates from Poisson distributions with means as indicated. The dashed line shows the output of the Fisher formula. **B** Even for vectors drawn from a distribution with mean = 1, the Fisher formula significantly mis-estimated *p* values smaller than 10^{-4}. **C** Poor performance of the Fisher formula is worsened when data are drawn from a Poisson-log-normal distribution, rather than a Poisson distribution (in this case the underlying log-normal distribution had a coefficient of variation of 0.5). **D**–**E** Data were simulated under the scenario that gene expression is the same in every cell, but due to differences in sequencing depth, observed gene expression varies according to the depths shown in the inset to panel (**D**). In panel (**D**), true gene expression was adjusted so that observed gene expression after normalization would have a mean of 1, and the Pearson correlation coefficients (PCCs) obtained by correlating randomly chosen vectors are shown. Panel D plots empirically derived *p* values as a function of PCC, whereas panel E displays histograms of PCCs. Compared with data that do not require normalization, associated *p* values from normalized data are even more removed from the predictions of the Fisher formula

require *p* as low as 10^{-9}, a value for which we may estimate, by extrapolation, that the Fisher formula is highly inaccurate even for genes with mean expression = 1.

The simulations in Fig. 1A, B assume Poisson-distributed data but, as is often pointed out, scRNAseq data are usually over-dispersed relative to the Poisson distribution (more on this below). As shown in Fig. 1C, adding in such additional variance causes simulated data to deviate even further from the Fisher formula.

Further problems arise when considering that scRNAseq data always come from collections of cells with widely varying total numbers of UMI. Depending on the platform, such "sequencing depth" can vary over orders of magnitude, which is why normalization is usually considered a necessary early step in data analysis. Without normalization, it is obvious that many spurious gene–gene correlations would be detected, as any difference in sequencing depth between cells would, if not corrected for, induce positive correlation across all expressed genes.

As one might expect, normalizing individual reads by scaling them to each cell's sequencing depth eliminates this bias, restoring the expected value of PCC under the null hypothesis to zero. Yet normalization does not restore the *distribution* of PCCs to what it would have been had all cells been sequenced equally. The consequences can be dramatic, as shown in Fig. 1D, E, where we simulate a case in which "true" gene expression is the same in each of 500 cells, but observed gene expression is a

Poisson random variate from a mean that was scaled by a factor chosen from a distribution of cell-specific sequencing depths similar to what one might observe in a typical scRNAseq experiment, using the 10X Chromium platform (inset, Fig. 1D). Despite mean gene expression being ~ 1, the relationship between PCC and p value more closely resembles the case (in the absence of sequencing depth variation) where the mean is 0.01 (Fig. 1A). This is surprising, given that the average fraction of zeros in the normalized vectors was only 0.68 (which, for Poisson-distributed data, would be expected to occur at a mean expression level of 0.39). In short, for gene expression data resembling what is typically obtained in scRNAseq, the Fisher formula is highly unsuitable, for most pairs of genes, for estimating the significance of correlations.

An analytical model for the distribution of correlations

The data in Fig. 1 indicate that the relationship between PCCs and p values is highly sensitive both to data structure and procedures intended to "correct" for technical variation. Because of this, we were concerned that a suitable null model for the distribution of correlations might be difficult to estimate empirically, especially when the true biological variation in most datasets is unknown. We therefore turned to constructing a null model analytically, attempting to account for known sources of variation (beside meaningful biological variation). The three sources considered were (1) variation introduced by imperfect normalization; (2) technical variation due to random sampling of transcripts during library preparation and sequencing; and (3) variation due to stochasticity of gene expression. The last of these cannot properly be called technical variation—fluctuating gene expression is a biological phenomenon—but like technical noise, gene expression fluctuation is usually an unwanted source of variation, and its effects need somehow to be suppressed.

With regard to the first source, we follow [43] in correcting not the gene expression data points themselves, but rather their Pearson residuals. Traditionally, the Pearson residual P_{ij} for cell i and gene j, is defined as

$$P_{ij} = \frac{x_{ij} - \mu_j}{\sqrt{\mu_j}}$$

where x_{ij} is the gene expression value for cell i and gene j, and μ_j is the mean expression of gene j averaged over all cells. As the Pearson residual is mean-centered, its expectation value, $E[P_{ij}]$, is zero; thus, the average of Pearson residuals for a large number of x_{ij} drawn from a single distribution should approach zero. The average of squares of Pearson residuals can be seen, by inspection, to approach the variance divided by the mean of the distribution from which the x_{ij} derive. Variance divided by mean is also known as the Fano factor and is often used to assess whether data are consistent with a Poisson distribution since, for any Poisson distribution regardless of mean, $E[P_{ij}^2] = 1$.

Pearson residuals may also be used to construct PCCs, which are commonly defined in terms of variances and covariances, but may be equivalently expressed as:

$$\mathrm{PCC}_{a\times b} = \frac{\sum_{i=1}^{n} P_{ia}P_{ib}}{\sqrt{\sum_{i=1}^{n} P_{ia}^2 \sum_{i=1}^{n} P_{ib}^2}} = \frac{1}{(n-1)\sqrt{\phi_a \phi_b}} \sum_{i=1}^{n} P_{ia}P_{ib}$$

where $\mathrm{PCC}_{a\times b}$ is the Pearson correlation coefficient between genes a and b, n the number of cells, and ϕ_a and ϕ_b represent the Fano factors for genes a and b, respectively, i.e.

$$\phi_j = \frac{1}{n-1}\sum_{i=1}^{n} P_{ij}^2$$

Since $E[P_{ij}]=0$ for any gene, it follows that $E[\mathrm{PCC}_{a\times b}] = 0$, as long as all the expression values for gene a are drawn from a single distribution, and those for b are independently drawn from a single distribution.

However, in scRNAseq, unequal sequencing depth means that the expression values for any given gene are generally *not* drawn from a common distribution, but rather from one that is different for each cell. Interestingly, dividing x_{ij} in a Pearson residual by an appropriate scaling constant—i.e. normalizing the data—will restore $E[P_{ij}]$ to 0, but will not restore the higher moments of P_{ij}, e.g., $E[P_{ij}^2] \neq 1$. To capture the correct second moment, one must scale the value of μ_j inside each Pearson residual, rather than scaling x_{ij}. As [43] have pointed out, we can define a separate μ_{ij} for each cell and gene:

$$\mu_{\mathrm{ij}} = \frac{\sum_j x_{\mathrm{ij}} \sum_i x_{\mathrm{ij}}}{\sum_{ij} x_{\mathrm{ij}}}$$

and consequently, define a corrected Pearson residual as

$$P_{ij} = \frac{x_{\mathrm{ij}} - \mu_{\mathrm{ij}}}{\sqrt{\mu_{\mathrm{ij}}}}.$$

Although this transformation does not recover moments of the Pearson residual beyond the first two, it provides a principled alternative to traditional normalization. Moreover, by permitting calculation of a corrected Fano factor that has the appropriate expectation value under the assumption of Poisson distributed data, it can be used to test that assumption, in real data.

Source #2 refers to technical variation due to random, independent sampling of discrete numbers of transcripts. Although the sampling process in scRNAseq involves several discrete steps, including cell lysis, library preparation and DNA sequencing, several groups have argued that, at least at modest to low expression levels, simple Poisson "noise" can reasonably model the variation derived from these processes [35–38]. We accept this assumption here, but note that, in the following derivations, the Poisson distribution could just as easily be replaced by another known distribution, if it were adequately justified.

Finally, source #3 refers to the fact that "equivalent" cells are usually only equivalent in a time-averaged sense, i.e., transcript numbers will fluctuate around some mean value. Both theory and observation support the conclusion that these fluctuations can be large [30, 44–46]. The actual magnitude seems to differ for different categories of genes, but data from single-molecule transcript counting [44] suggest that, for most genes, the distribution of transcripts typically is approximately log-normal (consistent with the theoretical work of [47]), with a coefficient of variation in the range of 0.2–0.6.

Thus, even if library synthesis and sequencing performed identically across cells, one should not expect to observe Poisson-distributed reads. Under the reasonable assumption that gene expression fluctuations and sampling are independent, the variance of the combined process should be the sum of the variances of the composing processes. Since both processes have the same mean, we can re-state this thusly: the Fano factor of the combined process should be the sum of the Fano factors of the composing processes. One can then use this fact to further adjust the Pearson residual, and subsequently the Fano factor. Specifically, we define a "modified corrected Pearson residual" as:

$$P'_{\mathrm{ij}} = \frac{x_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}\left(1 + c_j^2 \mu_{ij}\right)}} \tag{1}$$

where c represents the coefficient of variation of gene expression for gene j. Accordingly, the expectation value for $\left(P'_{\mathrm{ij}}\right)^2$ becomes

$$\frac{\left\langle \frac{(x_{ij}-\mu_{ij})^2}{\mu_{\mathrm{ij}}} \right\rangle}{\left(1 + c_j^2 \mu_{\mathrm{ij}}\right)}$$

which resembles a Fano factor divided by $\left(1 + c_j^2 \mu_{\mathrm{ij}}\right)$. Since $c^2 \mu_{\mathrm{ij}}$ can alternatively be written as $\frac{\sigma_{\mathrm{ij}}}{\mu_{\mathrm{ij}}}$, where s is the standard deviation of gene expression fluctuations, the term $\left(1 + c_j^2 \mu_{\mathrm{ij}}\right)$ is simply the sum of the Fano factors for Poisson sampling (unity) and gene expression fluctuation ($c_j^2 \mu_{\mathrm{ij}}$). Dividing by that sum essentially removes the additional variance due to gene expression noise from the expectation value of $\left(P'_{\mathrm{ij}}\right)^2$, restoring that value to 1.

In this way, one may define a "modified corrected Fano factor" ϕ' equal to the expectation value of $\left(P'_{\mathrm{ij}}\right)^2$. Likewise, we may use ϕ' to define a modified corrected Pearson correlation coefficient, PCC′:

$$\mathrm{PCC}'_{a\times b} = \frac{1}{(n-1)\sqrt{\phi'_a \phi'_b}} \sum_{i=1}^{n} P'_{ia} P'_{ib}. \tag{2}$$

Notice that ϕ' provides a measure of the degree to which a gene's expression is more variable than expected by chance, and PCC′ provides a metric by which gene-pairs are judged more positively or negatively correlated than expected by chance, correcting in both cases for unequal sequencing across cells (without normalizing data) and the expected noisiness of gene expression.

To use these statistics in practice, one needs to know not only their expectation values but their full distributions under the null hypothesis. Constructing those analytically requires not only the coefficient of variation of gene expression noise, c, but the full distribution of that noise, which in the absence of information to the contrary, we will take to be log-normal [47], noting that any other distribution could as easily be substituted in the following discussion.

We thus treat x_{ij} as a Poisson random variable from a distribution whose mean is a log-normal random variable with a coefficient of variation of c (we refer to this compound distribution as Poisson-log-normal). Although an analytical form for the probability mass function of the Poisson-log-normal distribution is not known, we may derive analytical forms for an arbitrary number of its moments, as a function of μ (the mean) and c (see Methods). Thus, one can calculate for every cell i and gene j, given the observed values of μ_{ij}, the moments of the expected distributions of the P'_{ij} under the null hypothesis, and subsequently those for the $\left(P'_{ij}\right)^2$. From there one can calculate the moments of any number of products and sums of P'_{ij} and $\left(P'_{ij}\right)^2$, such that, eventually, the moments of ϕ' and PCC′ under the null hypothesis are ultimately obtained (see Appendix). Given a finite number of moments, one can estimate the tails of the distributions of these statistics (see Methods), allowing one to calculate the probability of extreme values of ϕ' and PCC′ arising by chance (p values).

This method, which we refer to as BigSur (Basic Informatics and Gene Statistics from Unnormalized Reads), provides an approach for discovering genes that are significantly variable across cells (based on ϕ'), and gene pairs that are significantly positively or negatively correlated (based on PCC′), automatically accounting for the widely varying distributions of these statistics as a function of gene expression level and vector length (number of cells). The one free parameter in the method, c, is relatively constrained, as its average value (over all genes) can be estimated from a plot of ϕ' versus mean expression (see below). In this manner, one can avoid the use of arbitrary thresholds or cutoffs in deciding which genes are significantly "highly" variable (e.g., for dimensionality reduction and cell classification) and which genes are significantly positively and negatively correlated (e.g., to discover gene expression modules and construct regulatory networks).

Performance on simulated data

In Fig. 2 we simulate gene expression data for 1000 genes and 999 cells, under the null model described above (i.e., complete independence), varying "true" mean expression widely and uniformly over the genes, such that the most highly expressed genes average 3467 transcripts/cell and the most lowly expressed 0.0351 per cell. "Observed" gene expression values are then obtained by randomly sampling from a Poisson log-normal distribution with $c = 0.5$, in which the gene-specific mean is first scaled in each cell according to a pre-defined distribution of sequencing depth factors (chosen to mimic typical ranges of sequencing depth when using the 10X Chromium platform). The result is a set of gene expression vectors of length 999, with means varying between 0.001 and 231.

As shown in Fig. 2A, for most genes with mean expression greater than ~ 0.1 reads/cell, uncorrected Fano factors exceed 1, and rise linearly with expression level. Normalizing the data—scaling expression in each cell to the relative sequencing depth of that cell—reduces high-expression skewing somewhat, but also elevates the Fano factors for genes with low expression, driving them closer to 2. These results, in which the majority of genes display Fano factors greater than 1, which rise further for highly expressed

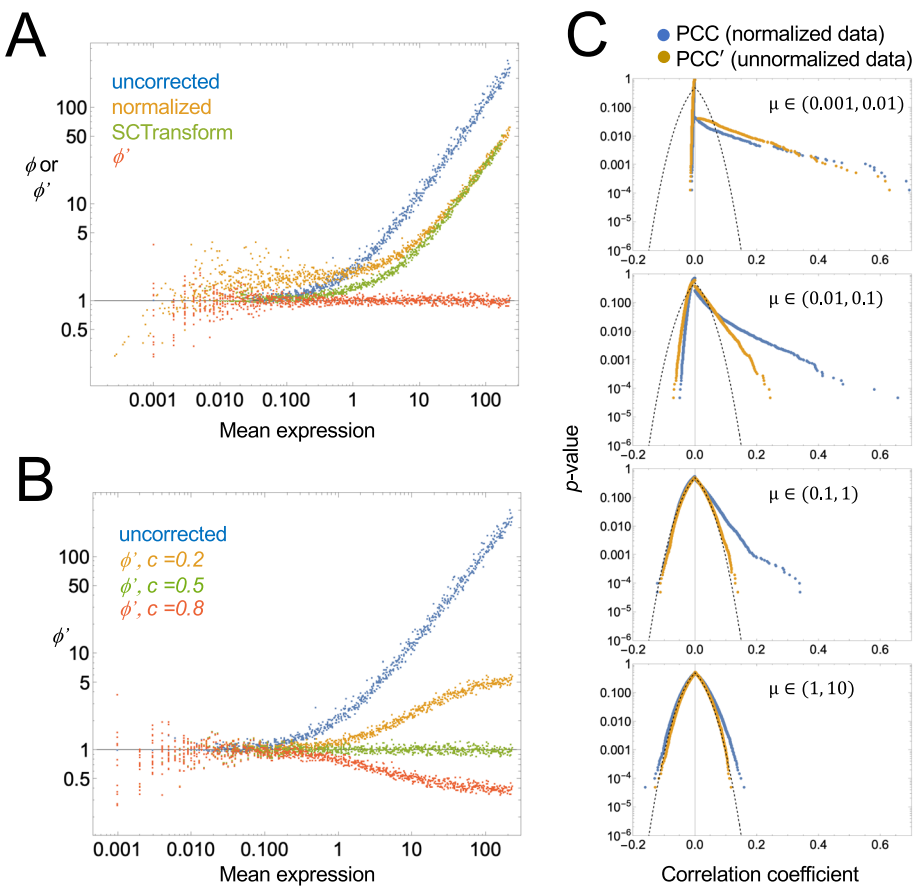


Fig. 2 Comparing uncorrected and modified corrected Fano factors and correlation coefficients. Random, independent, uncorrelated gene expression data was generated for 1000 genes in 999 cells, under the assumption that observations are random Poisson variates from a per-cell expression level that is itself a random variate of a log-normal distribution, scaled by a sequencing depth factor that is different for each cell (see methods). **A** Uncorrected (ϕ) or modified corrected (ϕ') Fano factors are plotted as a function of mean expression level for each gene. Uncorrected factors were calculated either without normalization, or with default normalization (scaling observations by sequencing depth factors, learned by summing the gene expression in each cell). Uncorrected Fano factors were also calculated using SCTransform [48] as an alternative to default normalization. Modified corrected Fano factors were obtained by applying BigSur to unnormalized data, using a coefficient of variation parameter of $c = 0.5$. **B** Modified corrected Fano factors (ϕ') were calculated as in A, but using different values of c. The data suggest that an optimal choice of c can usually be found by examining a plot of ϕ' versus mean expression. **C** Empirical p values associated with uncorrected (PCC) or modified corrected (PCC′) Pearson correlation coefficients were calculated for pairwise combinations of genes in bins of different mean gene expression level (μ); examples are shown for four representative bins (both genes derived from the same bin). With increasing gene expression levels, the p value versus PCC relationship begins to approach the Fisher formula (dashed curve), but it does so much sooner for PCC′ than PCC

genes, agree with the pattern most commonly seen in actual scRNAseq data. Values of the Fano factor for lowly expressed genes may be restored to near 1 by normalizing using SCTransform, an algorithm designed to correct for some of the variance-inflating aspects of normalization-by-scaling [48], but the presence of high Fano factors among the highly expressed genes persists.

In contrast, if we calculate the modified corrected Fano factor, ϕ', for each gene, using $c = 0.5$, we see that values are now centered around 1 at all expression levels (Fig. 2A).

Note that choosing different values of c produces consistent positive or negative skewing at mean gene expression values above 1 (Fig. 2B). Under the assumption that most genes in real data should not vary significantly across cells, one may therefore estimate the optimal choice of c for any data set by simply finding the value that minimizes such high-expression skewing of ϕ'.

Because it uses the moments of the Pearson residuals to calculate p values, BigSur assigns statistical significance to every gene's ϕ'. As expected, given that the data in Fig. 2A were random samples, no values of ϕ' were found to be statistically significant at a Bonferroni-corrected p value threshold of 0.05, or a Benjamini-Hochberg [49] false discovery rate of 0.05. Indeed, the lowest uncorrected p value associated with any of the 1000 genes in Fig. 2A was 0.001.

Similarly, when the same synthetic data are analyzed for gene–gene correlations, one may compare the PCC values produced directly from normalized expression data with the PCC′ values produced (from unnormalized data) by BigSur. As expected, both procedures return a distribution of values with zero mean, but PCC values are more broadly distributed than PCC′. Figure 2C shows the frequency at which various values of the correlation coefficient arise when vectors with different mean gene expression are correlated with each other. Although significant skewing from the Fisher formula is apparent, especially at low values of gene expression, it is much greater for PCC than PCC′. Indeed, for moderately expressed genes (e.g., 1–10 unique molecular identifiers [UMI] per cell), only PCC′ returns values whose distribution is relatively insensitive to expression level.

Performance on real data

To characterize the performance of BigSur on real data, we used the droplet-based sequencing data of Torre et al. [50], obtained from a clonal isolate of a human melanoma cell line grown in culture. In this data set, the number of cells is large (8640), data were validated on a second sequencing platform as well as by single molecule FISH, and the broad distribution of sequencing depths was typical of droplet-based sequencing.

We deliberately chose a clonal cell line because tissues always contain multiple cell types, i.e., groups of cells that express large sets of genes in cell-type specific ways. In such heterogeneous samples, genes that are associated with cell type identity will necessarily be strongly and densely correlated with each other; making the identification of correlations, in a sense, too easy—i.e., not a particularly good test of a method's performance—and not particularly informative (one may expect to identify as correlated more or less the same genes one would find by clustering cells by gene expression and testing for differential expression between clusters).

In contrast, the use of (ostensibly) homogeneous cells forces BigSur to operate on more subtle connections—for example, those involving fluctuations in function-specific gene regulatory networks—that cell clustering and differential expression would not easily detect.

Accordingly, scRNAseq data from these cells were subjected to minimal pre-processing prior to analysis (see Methods), such that expression values were analyzed for 14,933 genes. Total numbers of UMI per cell varied greatly, ranging from 67 to 90,494, with 90% of cells containing between 666 and 9004 UMI.

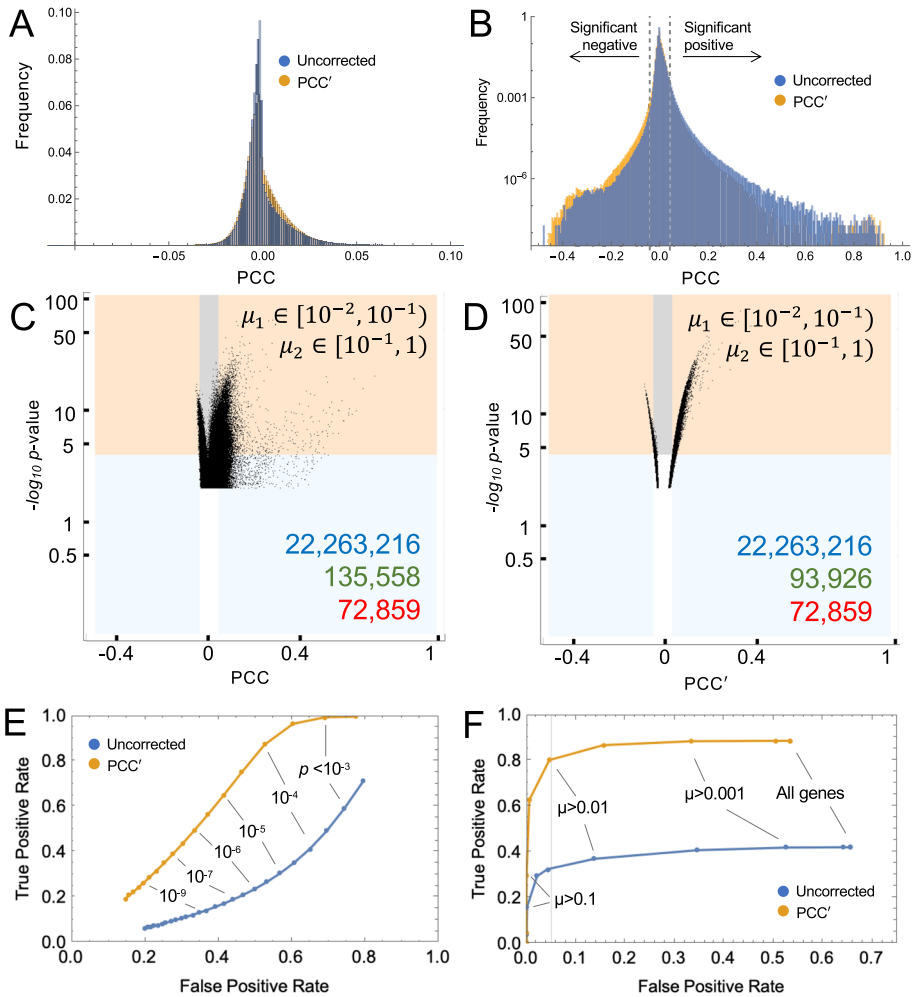


Fig. 3 Statistical significance of pairwise gene correlations in data from a clonal cell line. **A**, **B** Using scRNAseq data from a human melanoma cell line [50] (8640 cells × 14,933 genes), pairwise values of PCC were calculated from normalized data, and PCC′ from raw data. Histograms display the frequency of observed values (the logarithmic axis in B emphasizes low-frequency events). Notice in B how positive skewing, also seen in simulated data (Fig. 2), is less for PCC′ than PCC. Dashed lines in B show thresholds at which Fisher formula-derived *p* values would fall below $1.1 \times 10.^{-4}$. **C**, **D** Scatterplots showing *p* values assigned by BigSur to pairs of genes within two representative sets of bins of gene expression (for all pairwise combinations see Figs. S1, S2/Additional files 3, 4). The abscissa shows PCC (panel **C**) and PCC′ (panel **D**). The ordinate gives the negative log_{10} of *p* values determined by BigSur, i.e., larger values mean greater statistical significance. Orange and gray shading indicate gene pairs judged significant by BigSur (FDR < 0.02). Blue and orange show gene pairs that would have been judged statistically significant by applying the Fisher formula to the PCC or PCC′, using the same *p* value threshold as used by BigSur. The blue region contains gene pairs judged significant by the Fisher formula only, while the unshaded region shows gene pairs not significant by either method. Numbers in the lower right corner are the total numbers of possible correlations (blue), statistically significant correlations according to the Fisher formula (green), and statistically significant correlations according to BigSur (red). **E**, **F** ROC curves assessing whether the overall performance of the Fisher formula—applied either to PCC or PCC′—can be adequately improved either by using a more stringent *p* value cutoff (**E**) or limiting pairwise gene-correlations to those involving only genes with mean expression above a threshold level, μ (**F**)

First, we compared (uncorrected) PCCs for all gene–gene pairs, calculated using default-normalized expression data (i.e., data scaled to total UMI per cell), with PCC′ values returned by BigSur (Fig. 3A, B). In panel B, frequencies are scaled logarithmically, to better display the distribution of large absolute values. BigSur associated a

false discovery rate (FDR) of 2% to p values less than 1.15×10^{-4}, at which threshold it detected 639,789 correlations, 350,466 of which were positive. For uncorrected PCCs, the same p value cutoff would translate, using the Fisher formula, to $|\text{PCC}| > 0.041$, which is displayed as a dotted line in Fig. 3B. Using this threshold, 1,484,156 correlations would be considered significant. Comparison of histogram shapes shows that use of uncorrected PCCs particularly inflates positive correlations, especially large ones, and under-counts negative ones, which is consistent with the results obtained using simulated data (Fig. 2C).

To see how the discovery of correlations varied with gene expression level, we divided genes into 6 bins of different mean expression, and separately analyzed correlations between genes in all 21 possible combinations of bins. The full data are presented in Figs. S1 and S2 (Additional files 3, 4), with two representative panels in Fig. 3C, D. Each point represents a gene–gene pair. The value on the abscissa is either the uncorrected PCC (Figs. 3C and S1) or modified corrected PCC′ (Figs. 3D and S2), and the value on the ordinate is the $-log_{10}\, p$ value, as determined by BigSur (i.e., the larger the number, the lower the p value). Shaded territories mark those data points that were judged statistically significant (FDR < 0.02) either by BigSur (orange and gray), or when p values were calculated by the Fisher formula (blue and orange; for further details see figure legend). The data confirm that the Fisher formula returns an excess of correlations, compared to BigSur, albeit less severely for PCC′ than for PCC. Examination of the full dataset (Figs. S1, S2) shows that many more truly significant correlations are found among highly expressed genes; and the Fisher formula performs worst when at least one of the pairs in a correlation is a lowly-expressed gene. Indeed, as gene expression becomes high (e.g., mean value > 1 UMI per cell for both genes in a pair), the distribution of p values calculated by BigSur for PCC′ begins to approximate the Fisher formula reasonably well (Fig. S2), with deviation apparent only for very low p values ($-log_{10} p > 10$). This observation validates the accuracy of the method used by BigSur to recover p value distributions from the first five moments of the expected distributions of modified, corrected Pearson residuals (see "Methods" section).

Assuming, for the sake of illustration, that the correlations judged significant by BigSur represent ground truth, we may then calculate levels of true- and false-positivity when p values are calculated by feeding either PCC or PCC′ into the Fisher formula. This enabled us to ask whether applying a more stringent p value cutoff, or thresholding gene expression (i.e., excluding genes below a certain expression level), might enable this simpler, formulaic approach to achieve an acceptable level of sensitivity and specificity. As shown by the receiver-operating characteristic (ROC) curves in Fig. 3E, the performance of uncorrected PCCs is exceedingly poor regardless of p value threshold, with false positives exceeding true positives at all values. PCC′ does better, but to control FDR to < 10%, one still loses the ability to detect > 85% of true positives.

Arbitrarily thresholding gene expression performs somewhat better (Fig. 3F). For uncorrected PCCs, one must exclude all genes with average expression < 0.065 UMI/cell to control the FDR at 5%, which for this dataset means discarding 67% of all gene expression data, and in return recovering only 33% of true positives. PCC′ does much better here: we may recover 81% of true positives at an FDR of 5% by discarding

the ~ 52% of genes with the lowest expression. While the exact numbers are likely to vary for different data sets, these observations suggest that, if one is willing to sacrifice the power to identify a substantial minority of correlations, feeding PCC′ (but not PCC) into the Fisher formula can represent an acceptable and computationally fast alternative to *p* value identification by BigSur.

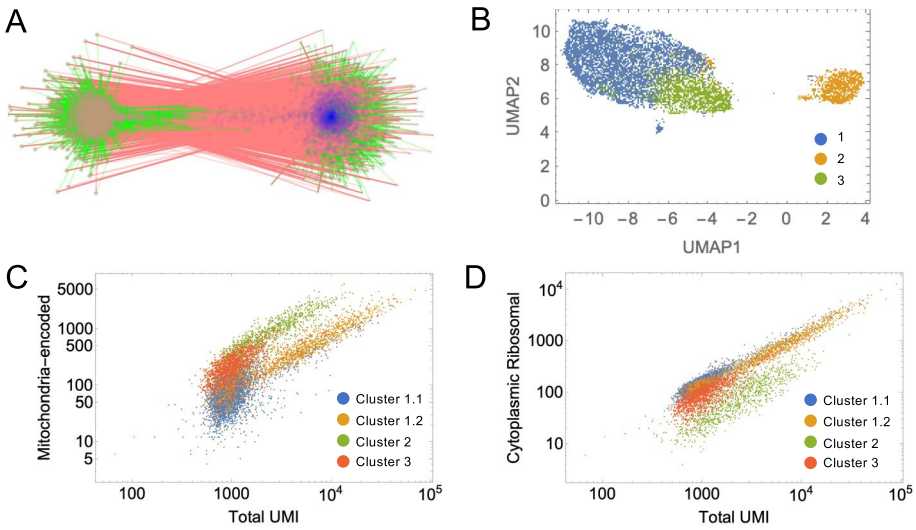
Clustering using correlations

Although BigSur found 639,789 statistically significant correlations in this dataset (about 0.57% of all possible pairwise correlations) the vast majority had quite small values of PCC′ (Fig. 3A, B), indicating that most statistically significant correlations were weak. To obtain a measure of correlation strength that could be compared across samples of different lengths (numbers of cells), we expressed each correlation in terms of an "equivalent" PCC, which is simply the PCC value that, for continuous, normally-distributed data of the same length, would have produced the same *p* value (by the Fisher formula). As shown in Fig. S3 (Additional file 5), only 4335 gene pairs displayed equivalent PCCs greater than 0.2 or less than − 0.2.

Yet, despite the weak strength of most correlations, there are good reasons to believe them to be biologically relevant. One of the simplest comes from examining the frequency at which correlations arise among paralogous genes and genes that encode proteins that physically interact. It is known that gene paralogs are often co-regulated [51] leading us to expect paralog pairs to be enriched among *bona fide* correlations. It is also reasonable to expect that transcripts encoding proteins that interact will be co-expressed at least some of the time. As it happens, among the correlations identified by BigSur we observe ~ 12 fold enrichment for paralogs and ~ 7.5 fold enrichment for genes encoding physically-interacting proteins (see "Methods" section).

To divide the full set of correlations into potentially interpretable groups, we used a random-walk algorithm [52] to identify subnetworks more highly connected internally than to other genes; we refer to these as gene communities. Most communities were of modest size, with 94 of the 96 largest containing between 4 and 392 genes each. However, the largest two contained 2160 and 1570 genes respectively, were very densely connected internally, and strongly anti-correlated with each other (Fig. 4A). These factors strongly suggest that these cells, despite being a clonal line, are heterogenous, falling into (at least) two distinct groups. Interestingly, the largest gene community contained virtually the entire set of mitochondrially-encoded mitochondrial genes, and the second largest contained virtually all protein-coding ribosomal genes (for both cytoplasmic and mitochondrial ribosomes). Using the top 50 most-highly connected genes (those with the largest positive equivalent PCCs) in each of these communities as features, we performed Leiden clustering on the modified corrected Pearson residuals of all 8640 cells, and easily subdivided them into three clusters of 5812, 1180 and 1638 cells, which we labeled as clusters 1,2 and 3, respectively (Fig. 4B; see "Methods" section).

We then analyzed each cluster independently by BigSur, re-calculating PCC′ values and statistical significance for each group of cells. Surprisingly, in each cluster BigSur again found two large, strongly anti-correlating communities, one of which contained the mitochondrial-encoded genes and the other the ribosomal protein genes. This led us to further subcluster cluster 1, again using the 50 most-highly connected genes in each



E

Cluster	Cell number	Significant intra-cluster correlations			Genes expressed	Genes involved in Significant correlations
		Positive	Negative	Total		
all cells	8640	350,466	289,323	639,789	14,933	11,961
1	5,812	21,203	8,589	29,792	14,933	4,405
1.1	4,226	1,382	536	1,918	14,786	1,281
1.2	1,582	18,353	4,702	23,055	14,915	2,519
2	1,180	20,390	4,992	25,382	14,698	2,672
3	1,638	1,869	366	2,235	14,232	918

Fig. 4 Clustering cells based on mitochondrial and ribosomal communities. **A** Correlations among the two largest gene communities detected by BigSur are shown. Vertices are genes, and edges—green and red—represent significant positive and negative correlations, respectively. Blue vertices represent members of the mitochondrially encoded gene community and brown vertices the ribosomal protein gene community (Labels have been omitted due to the large numbers of gene involved). **B** Using the top 50 most highly positively connected vertices in the two communities as features, cells were subjected to PCA and Leiden clustering; the three clusters recovered are displayed by UMAP. **C**, **D** After a second round of clustering of Cluster 1, the resulting four cell groups were analyzed for the distribution of expression of ribosomal protein-coding and mitochondrially-encoded genes, as a function of total UMI in each cell. The results show that the four clusters form distinct groups based on their relative abundances of ribosomal and mitochondrial genes. **E** Results of applying BigSur to each cluster. Note the large decrease in statistically significant correlations—especially negative correlations—in any of the clusters when compared with results obtained using all of the cells together (> 600,000 total correlations). This is consistent with heterogeneity in the original sample, causing large blocks of genes to correlate and anti-correlate

of these two communities as features and subdivided it into two groups of 4226 (cluster 1.1) and 1582 cells (cluster 1.2).

Variable expression of mitochondrially-encoded genes is a common finding in scRNAseq. Their presence at high levels (e.g. > 25% of total UMI) is usually interpreted as an indication of a "low quality" cell—potentially one in which the plasma membrane has ruptured and cytoplasm has been lost—or perhaps a cell in the process of apoptosis [53]. Closer examination of the cell clusters identified in this dataset suggests these phenomena are likely only part of the explanation. In Fig. 4C, we plot mitochondrial-encoded UMI versus total UMI for the entire set of cells, coloring them according to the four cell clusters mentioned above. Several distinct behaviors were noted. Cell

cluster 2 forms a coherent group with high mitochondrial expression that is linearly proportional to total UMI. Throughout this group, the percentage of total transcripts that is mitochondrial remains in a narrow band, with mean of 34% and coefficient of variation (CV) of 0.41. Cluster 3 displays low total UMI, and a mitochondrial fraction centered around a mean of 21% (CV = 0.43). Cluster 1.2 has high mitochondrial expression, but even higher total UMI, such that the mitochondrial fraction averages 9.7% (CV = 0.45), while cluster 1.1 has both low total UMI and low mitochondrial UMI (average mitochondrial fraction 8.2%, CV = 0.47).

In Fig. 4D, we also plot total UMI derived from ribosomal protein genes against total UMI. Of all the clusters, cluster 2 best fits the expectation for "damaged" cells: Mitochondrial and ribosomal UMI rise proportionately with total UMI (consistent with randomly varying sequencing depths), but the proportion of mitochondrial UMI is almost three times higher, the proportion of ribosomal UMI almost three times lower, and total UMI about 2.5 times lower than in cluster 1.2, the only other cluster that displays a wide range of sequencing depths. Yet it is curious that the mitochondrial proportions in cluster 2 are so narrowly distributed around a mean; one might expect variable degrees of cytoplasm leakage following cell damage to produce a broad distribution beginning near cluster 1.2 and gradually tapering off. The absence of such behavior suggests that such cells are not merely variably damaged during preparation, but represent a distinct, possibly pre-existing state, perhaps associated with some form of cell stress or death (although we see no significant enrichment of gene expression associated with apoptosis [54] in cluster 2).

Even if we remove cluster 2 from further analysis, clusters 1.1, 1.2 and 3 also differ between each other in relative proportions of mitochondrially-encoded, ribosomal and total transcripts; however, the way in which they do so is not suggestive of any simple mechanism of cell damage (and overall mitochondrial content is not in the range that

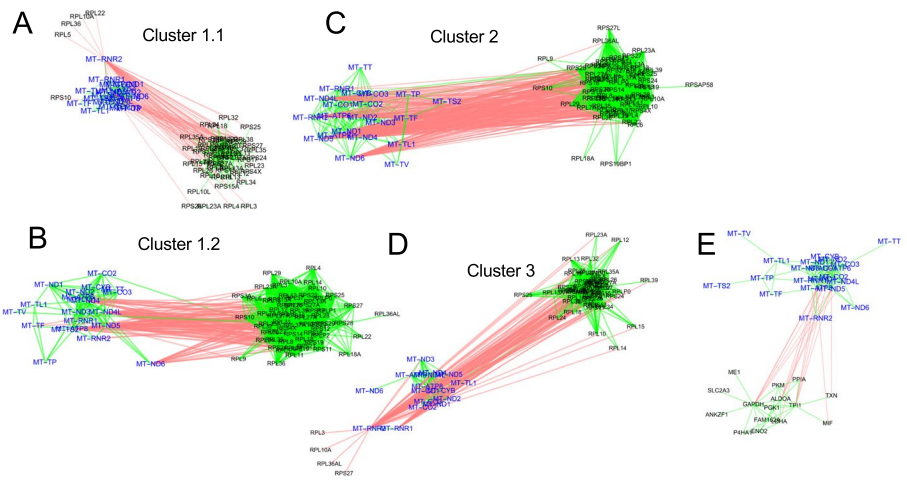


Fig. 5 Mitochondrial communities are a source of strong anti-correlations. Mitochondrially-encoded and ribosomal protein gene communities were identified in all clusters. Anti-correlations identified specifically between mitochondrially-encoded genes and ribosomal protein genes within the different clusters are shown in **A** (cluster 1.1), **B** (cluster 1.2), **C** (cluster 2) and **D** (cluster 3). Panel **E** shows additional anticorrelations in cluster 1.2 involving mitochondrially-encoded genes and glycolysis genes For ease of readability, mitochondrial genes have been highlighted in blue. Red links refer to negative correlations; green to positive

would necessarily lead to exclusion from downstream scRNAseq analyses). As mentioned above, using the genes in the mitochondrially-encoded- and ribosomal-enriched communities as features, analysis of gene correlations using BigSur separately on each of the four clusters revealed anti-correlating communities of mitochondrially-encoded and ribosomal genes in every case (Fig. 5A–D; Table 2). In fact, even after another round of subclustering of the subclusters derived from cluster 1 (cluster 1.1 and cluster 1.2) using these genes as features, BigSur still identified distinct, anti-correlating mitochondrially-encoded and ribosomal communities (not shown). These data strongly suggest that, notwithstanding effects of cell damage or cell death, there exists among healthy cells continuous, anti-correlated variation in both the mitochondrially-encoded and ribosomal protein genes, suggestive of some biologically meaningful regulatory relationship (discussed further below).

Analysis of gene communities

Figure 4E summarizes the results of using BigSur to identify significant correlations (FDR < 0.02) in each of the cell clusters described above (1.1, 1.2, 2 and 3). Because statistical power to detect correlations falls with number of cells analyzed, one might expect to see the fewest significant correlations in the smallest clusters, but this was not the case. Instead, the clusters with the largest number of significant correlations were those with the highest average sequencing depth (Fig. 4C), suggesting that data sparsity has an especially strong influence on correlation detection.

Schadt and colleagues have recently argued [55] that negative correlations may be considered a strong indicator of cell heterogeneity. These authors argue that minimization of the proportion of negative gene–gene correlations provides a principled metric for determining when to cease sub-clustering cells. Consistent with this view, we observed that, as cells were successively subclustered, the proportion of negative correlations fell from 45% to about 20% (Fig. 4E). The clusters with the lowest proportion of negative correlations were clusters 1.2, 2 and 3. As cluster 2 had very high levels of mitochondrially-encoded genes (33.7% of UMI), and cluster 3 had the lowest average UMI/cell (1273) of any cluster, we focused our analysis on cluster 1.2, which had both low levels of mitochondrial genes (9.7% of UMI) and high average UMI/cell (6252)..

Within cluster 1.2, we found that enrichment for correlated paralog pairs and genes whose products interact physically was much higher than when all 8640 cells had been considered together. Specifically, among the positive correlations, paralog pairs were enriched 55-fold, and links supported by protein–protein interactions 32-fold, over what would have been expected by chance. Indeed, 15% of all positive correlations in cluster 1.2 corresponded to links supported by known protein–protein interactions.

Table 1, Additional file 2: Table S1 and Figs. S4–S7 (Additional files 6, 7, 8, 9) summarize results for 13 of the largest gene communities detected in cluster 1.2, which collectively account for 1291 of the 2519 genes that showed any significant positive correlations. Table 1 groups the genes of each community into functional categories according to annotations found in the most recent release of the MSigDB database [56, 57]. Annotations identified using the Database for Annotation, Visualization and Integrated Discovery (DAVID [58]) are also shown in Additional file 2: Table S1. In 11 of the 13

Table 1 Gene communities identified in cell subcluster 1.2

Community	# of genes	Category	Gene names
A	253	Cell cycle, G2/M	ANLN, ANP32B, ANP32E, ARHGAP11A, ARHGEF39, ARL6IP1, ASPM, AURKA, AURKB, BIRC5, BUB1, BUB1B, BUB3, CALM3, CASC5, CBX1, CCNA2, CCNB1, CCNB2, CCNF, CDC20, CDC25B, CDC27, CDCA2, CDCA3, CDCA8, CDK5RAP2, CDKN1B, CDKN3, CENPA, CENPE, CENPF, CEP55, CEP70, CHEK2, CIT, CKAP2, CKAP2L, CKAP5, CKS1B, CKS2, CNTRL, CTCF, DBF4, DDX39A, DEPDC1, DEPDC1B, DKC1, DLGAP5, DNAJA1, DNAJB1, DTYMK, DYNLL1, ECT2, EGR1, FAM64A, FAM83D, FANCD2, FOXM1, G2E3, GAS2L3, GPSM2, GTSE1, H2AFV, H2AFX, H2AFZ, HJURP, HMG20B, HMGB2, HMGB3, HMGN2, HMMR, HN1, HNRNPD, HP1BP3, HSP90AA1, HSPA8, ILF2, INCENP, KIAA1524, KIF11, KIF14, KIF15, KIF18A, KIF20A, KIF20B, KIF23, KIF2C, KIF4A, KIF5B, KNSTRN, KPNA2, KPNB1, LBR, LIN54, LSM6, MAD2L1, MIS18BP1, MKI67, MPHOSPH9, MTF2, MZT1, NDC80, NEIL3, NEK2, NUCKS1, NUF2, NUP37, NUSAP1, OIP5, PBK, PCF11, PDS5B, PIF1, PLK1, PLK4, PPP1R12A, PPP2R5C, PRC1, PRR11, PSMC3, PSRC1, PTTG1, RACGAP1, RAD21, RANGAP1, RBM8A, RCCD1, RPS6KA5, SFPQ, SGOL1, SGOL2, SHCBP1, SMC4, SNRPA1, SNRPD1, SNRPD3, SPAG5, SPDL1, SRSF3, STMN1, TACC3, TICRR, TOP2A, TPX2, TRIP13, TROAP, TTK, TUBA1B, TUBA1C, TUBB2A, TUBB4B, UBE2C, UBE2S, UTP18, WHSC1, XPO1
		Cell cycle, other	ACTB, ACTL6A, BRD7, CACYBP, CENPN, CENPW, CHAMP1, DYNLT1, FXR1, H3F3B, HES1, KIAA0586, LMNB2, LYAR, MAGOHB, NCAPD2, NCAPG, PARPBP, PIM3, POLR2K, PPP4R2, RAN, RHEB, SETD2, SKA2, SNRPB, UBE2D3, YY1, ZWILCH
		Spliceosome	ACIN1, LSM4, LSM5, PPIH, RBMX, SRRM1, SRRT, SRSF7, WBP4
		Not accounted for (58/253) = 23%	*ABHD2, ACAT2, ACTG1, AFF4, AHSA1, ARPC5L, AZIN1, BNIP2, BRIX1, C1orf52, C5orf34, CCDC150, CCDC18, CCDC34, CHORDC1, DDX6, DIAPH3, DLEU2, EIF3J, EIF5, EXOSC3, FAM46A, FUBP1, FZD6, GNL2, HIBCH, HIRIP3, HMGB1, HMGN1, HSPH1, IDI1, IER2, IFI16, KPNA4, LEO1, LZIC, MITF, NAV2, PHAX, PHF19, PPID, PRDX3, PSPC1, PTGES3, RBBP6, RLF, RPL39L, SCLT1, SEC62, SUB1, SYTL2, TAF13, TFAM, TRMT10B, TWISTNB, UBALD2, UGDH, WDR1*

Table 1 (continued)

Community	# of genes	Category	Gene names
B	213	Cell cycle G1/S, DNA replication	ARL6IP6, ASF1B, ATAD2, ATAD5, BAZ1B, BLM, BRCA1, BRCA2, BRIP1, CASP8AP2, CCNE2, CDC16, CDC45, CDC6, CDCA5, CDCA7, CDK1, CENPQ, CHAF1A, CHAF1B, CHEK1, CLSPN, DDX11, DEK, DHFR, DNA2, DNMT1, DSCC1, DTL, DUT, E2F7, E2F8, EME1, ESCO2, EXO1, EZH2, FAM111A, FBXO5, FEN1, GEN1, GINS1, GINS2, GINS4, GMNN, HELLS, HIST1H4C, ICMT, KIAA0101, LIG1, MASTL, MCM10, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7, MCM8, MGME1, MSH2, MSH6, NASP, ORC1, ORC6, PCNA, PKMYT1, POLA1, POLD3, POLE3, POLQ, PRIM1, PRKDC, RAD18, RAD51, RAD51C, RBBP4, RBBP7, RBBP8, RBL1, RECQL, RFC1, RFC2, RFC3, RFC4, RFC5, RMI1, RNASEH2A, RPA2, RPA3, RRM1, SLBP, SMC1A, SMC2, SUPT16H, SVIP, TIMELESS, TIPIN, TK1, TMPO, TOPBP1, TYMS, UNG, USP1, USP37, WDHD1, WDR76, XRCC2, YEATS4, ZGRF1
		Cell cycle, other	CALM2, CCDC14, CDK11B, CENPH, CENPJ, CENPK, CENPM, CENPU, CEP152, CMC2, CSE1L, DNAJC9, DSN1, ERCC6L, EXOSC8, FBXO43, HAUS1, HAUS8, HPS4, ITGB3BP, KLHL23, KNTC1, LMNB1, MELK, MIS18A, MND1, MYBL1, NCAPD3, NCAPG2, NCAPH, NUP107, NUP85, PAICS, PSMC3IP, RANBP1, RTKN2, SPC25, SRSF10, STIL, SYNE2, TRIM37, TUBG1, VRK1
		DNA repair	ANKRD32, CDC5L, FANCB, FANCI, FANCL, FIGNL1, FUS, KIF22, MBD4, NUDT21, PARP2, PMS1, RAD51AP1, RAD54B, RIF1, SMCHD1, TTF2, UBE2T, USP10, XRCC5, ZWINT
		Histones	HIST1H1D, HIST1H1E
		Not accounted for (38/213) = 18%	*ACAA2, BAZ2B, BTG3, C19orf48, C21orf58, C3orf14, CBX5, CCDC15, CDCA4, CTDSPL2, CTNNAL1, DERA, FAM161A, FKBP5, GGCT, GGH, HMOX1, HNRNPAB, HSPB11, LGALS1, LSM8, MTHFD1, NAP1L4, POLR3K, POP7, PPM1G, PSIP1, PTPRG-AS1, RRM2, SDHA, SIVA1, SKA3, SLC43A3, SNRNP25, SRSF2, TEX30, TMEM106C, WDR34*

Table 1 (continued)

Community	# of genes	Category	Gene names
C	194	Unfolded protein response, ER stress	ASNS, ATF4, ATF6, ATP2A2, C19orf10, CALR, CANX, CAV1, CREB3L2, CRELD1, CRELD2, DDIT3, DDIT4, DNAJB11, DNAJB9, DNAJC10, DNAJC3, EIF4EBP1, ERP44, FAM129A, GFPT1, GLRX2, HERPUD1, HM13, HSP90B1, HSPA13, HSPA5, HSPA9, HYOU1, KDELR3, LMAN1, MANF, MTHFD2, NARS, NUPR1, OS9, P4HB, PDIA3, PDIA4, PDIA6, PPIB, PPP1R15A, PSAT1, RCN3, SDF2L1, SEC31A, SEL1L, SERP1, SESN2, SIL1, SPCS3, SRPR, SSR1, STT3B, TARS, TMCO1, TMED2, TRIB3, TXNIP, UGGT1, VEGFA, VIMP, WARS, WIPI1, XBP1, XPOT
		Other endoplasmic reticulum	ARF1, ARF4, ARFGAP3, ASPH, CALU, CLCN3, COPA, COPB1, COPB2, DDOST, FKBP2, GOLGA4, KDELR1, KDELR2, LAMB1, MAGT1, MIA3, MLEC, MTDH, NFE2L1, NUCB2, OSTC, PLOD3, PRNP, RCN1, RHOBTB3, RPN1, RPN2, RRBP1, SEC11C, SEC24D, SEC63, SLC35B1, SLC7A11, SMIM14, SPCS2, SSR2, SSR3, SUCO, TMED10, TMED9, TMF1, TRAM1
		tRNAs	AARS, EPRS, GARS, LARS, MARS, SARS, YARS
		Metal ion transport and homeostasis	ANXA2, ARHGEF2, ATP1A1, BEST1, CYB561, EDNRB, PDE4B, PEG10, SERPINE2, SHMT2, SLC39A14, SLC3A2, SLC5A3, TCEA1, TES
		Not accounted for (63/194) = 32%	*ACBD3, ALDH1L2, ANKRD11, ATP6V1F, BCAT1, BTG1, BUD31, C11orf24, C6orf48, CDH19, CDK2AP2, CITED1, CTC-425F1.4, DDR2, DUSP6, EIF1, FAM114A1, FBXO25, GADD45A, GAS5, GAS7, GDF15, GHITM, GOT1, GPNMB, HDLBP, IFRD1, IL1RAP, ITGA4, LARP1B, LIMA1, LMO4, LRIF1, LURAP1L-AS1, MAP4, MBNL2, MCF2L, MESDC2, MORF4L2, NRSN2-AS1, PHGDH, PHLDA1, PMP22, PPAPDC1B, PSAP, PYCR1, PYGB, RCAN1, S100A11, SELK, SEP15, SLC1A5, SLFN5, SORBS2, SPARC, SUPV3L1, TAX1BP1, TMEM263, TMX4, WDR26, ZEB2, ZFAS1, ZNF106*

Table 1 (continued)

Community	# of genes	Category	Gene names
D	158	Ribosomal proteins	FAU, RPL10, RPL10A, RPL11, RPL12, RPL13, RPL13A, RPL14, RPL15, RPL18, RPL18A, RPL19, RPL22, RPL23, RPL23A, RPL24, RPL26, RPL27, RPL27A, RPL28, RPL29, RPL3, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36, RPL36AL, RPL37, RPL37A, RPL38, RPL39, RPL4, RPL5, RPL6, RPL7A, RPL8, RPL9, RPLP0, RPLP1, RPLP2, RPS10, RPS11, RPS12, RPS13, RPS14, RPS15, RPS15A, RPS16, RPS17, RPS18, RPS19, RPS2, RPS20, RPS21, RPS23, RPS24, RPS25, RPS27, RPS27A, RPS28, RPS29, RPS3, RPS3A, RPS4X, RPS5, RPS6, RPS7, RPS8, RPS9, RPSA, UBA52
		Protein translation	EEF1A1, EEF1B2, EEF1D, EEF2, EIF3E, EIF3F, EIF3H, EIF3K, NACA, NACA2, PABPC1, PRKCSH, SEC11A, SSR4
		Ribosome biogenesis	GLTSCR2, NHP2
		Iron metabolism	FTH1, FTL
		Not accounted for (66/158) = 42%	*ABHD14B, APP, ARL2, ATP5G2, C14orf2, C4orf48, C7orf55, CCDC86, CCDC88C, CD109, CEP350, COMMD6, COX5B, COX7A2L, CST3, CUEDC2, DCBLD2, ETV5, FBXO32, FIBP, HIF1A, HLTF, IFITM3, IMPDH2, ITSN2, LRRC75A, LRRC75A-AS1, METTL12, MIA, MME, MT-ND6, NAA38, NAP1L1, NDUFB7, NDUFS4, NIN, OAZ1, OST4, PABPC3, PGLS, PPFIA1, PRDX5, PRSS23, RB1CC1, ROMO1, RP11-193E15.4, RP11-356J5.12, RP11-669N7.2, SAT2, SERF2, SF3B6, SH3KBP1, SNHG19, SNHG5, SNHG6, SUCLG2, TOMM7, TPT1, TSPO, UBL5, UQCRB, UQCRH, UQCRHL, VPS28, WNK1*
E	120	Cellular respiration	ACAT1, ATP5B, ATP5G1, ATP5G3, ATP5J, ATP5J2, ATP5L, ATP5O, BNIP3, BNIP3L, C17orf89, C1QBP, CHCHD2, COX17, COX5A, COX6B1, COX6C, COX7C, COX8A, CYCS, HMBS, LGALS3, MMADHC, MRPL11, MRPL12, MRPL21, MRPL41, MRPS16, NDUFA2, NDUFA4, NDUFA8, NDUFB3, NDUFB4, NDUFB9, NDUFC1, NDUFC2, NDUFS5, NDUFS6, NPM1, PARK7, PFDN2, PFN1, PHB, PHB2, PSMB4, SDHB, SLC25A3, SLC25A36, SLC25A39, SLIRP, TIMM13, TRAP1, UQCR10, UQCRQ, VDAC1
		Glycolysis	ALDOA, ANKZF1, CD44, ENO2, FAM162A, GAPDH, LDHA, ME1, MIF, P4HA1, PGK1, PKM, PPIA, SLC2A3, TPI1, TXN
		Hypoxia, oxidative stress	ADM, ARHGDIA, ATXN2L, C20orf27, EIF5A, GBE1, GLO1, HINT1, HSPB1, ITGAE, KDM3A, NME1, NRN1, POLR2L, PSMA4, PSMB2, PSMB6, S100A10, STRA13, TAF1D, TNS1, TRIB2, VIM
		Splicing	SNRPD2, SNRPF, THOC7, YBX1
		Endopeptidase inhibition	CST1, CST4, CSTB
		Not accounted for (18/120) = 15%	*APRT, BTF3, C12orf57, CCL28, CFL1, DARS, EIF4A1, EIF4A2, H1F0, HPCAL1, LMAN2, RP11-58E21.4, SDF4, SLAMF9, TMED4, TRAPPC1, TXNDC17, USP11*

Table 1 (continued)

Community	# of genes	Category	Gene names
F	110	Pigmentation, melanosome	APOE, ASAH1, ATOX1, B2M, BIRC7, CALM1, CAPG, CD320, CD63, CDK2, CLEC11A, CTSA, CTSB, CTSD, CTSH, DAB2, EMP1, ERP29, GNPTAB, GPR56, GRN, GSTO1, GUSB, GYPC, HMCN1, IDH1, MFGE8, MLANA, MLPH, MRPL44, MZT2B, NDUFB1, NGFRAP1, NPAS2, NPC2, NUMA1, PMEL, PRDX1, PRDX4, PSMA7, RAB27A, RAB38, RAI14, RLBP1, SDCBP, SLC24A5, TMEM98, TYR, UBR5, WSB1
		Myeloid phenotypes (monocyte, macrophage, basophil, dendritic cell)	ACOT7, AKR1A1, ARPC1B, ATP6AP1, ATP6V0E1, BHLHE41, BTN2A2, C21orf91, CD59, CD9, CHCHD6, CPM, DBI, HEXB, HIPK2, ITM2C, LGALS3BP, LITAF, LRPAP1, MDH1, METTL9, NDUFB6, PEBP1, QPCT, RGS10, SPAG9, SUMO2, TIMP1, TMEM147, TMSB4X, TOP1, TRIM63
		Not accounted for (28/110) = 25%	*AC104655.3, AFG3L2, BCAR3, CAPN3, CFAP61, COL4A3BP, CYTL1, G3BP1, G3BP2, GJB1, HTATSF1, KIAA1456, LHFPL3-AS1, LONRF1, MAGED1, MAGI1, MPG, PCCB, PDE3B, PLCH1-AS1, PMP2, POLR2F, RNF19A, RP4-718J7.4, SCML4, SEPT6, STK32A, TIMM50*
G	70	HNRNPs	HNRNPA2B1, HNRNPA3, HNRNPH3, HNRNPK, HNRNPM, HNRNPR, HNRNPU
		Other RNA processing	DDX1, DDX21, DNTTIP2, EIF4A3, EIF4G1, ESF1, LUC7L3, MYH10, NCL, NOLC1, NOP56, NOP58, NSUN2, NUDT1, PA2G4, PNO1, PNPT1, PRPF40A, PUS7, SET, SNRPC, SNRPE, SNRPG, SSB, SYNCRIP, TPRKB, TSR1, WDR3
		Other RNA binding proteins	BZW1, CNBP, EIF3A, EIF3B, GSPT1, LARP4, SERBP1
		Protein chaperones	DNAJC2, DNAJC8, CCT4, CCT5, CCT6A, HSP90AB1, HSPD1, HSPE1, NUDCD1, STIP1, TCP1
		Not accounted for (17/70) = 24%	*ARPC2, ATP6V1C1, CEBPZ, GPATCH4, KDM5A, KIAA0020, MACF1, MRPL19, NOM1, PSMD14, PSME4, PTMA, TLN1, TPM3, UTP11L, YWHAQ, ZC3H15*
H	59	Mitochondrially-encoded genes	MT-ATP6, MT-ATP8, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-RNR1, MT-RNR2, MT-TF, MT-TL1, MT-TP, MT-TS2, MT-TT, MT-TV
		Regulation of expression of mitochondrially-encoded genes	LRPPRC
		Protein ubiquitination	CBLB, DDX3X, DZIP3, HECTD1, MYCBP2, NBR1, TBL1XR1
		Hypoxia/stress response	CALD1, DST, FOS, IGF1R, IGFBP5, JUNB, MXI1, ZMYND8
		Not accounted for (20/59) = 34%	*AC021451.1, ADCY1, AKAP9, BPTF, CAPN2, COX6A1, CSDE1, DDX17, FCHSD2, GOLGB1, KRIT1, MEF2A, PCLO, PLCB4, PPP1R9A, SPEN, SRRM2, TDRD3, TIA1, ZBTB38, ZC3H13, ZKSCAN1, ZNF704*
I	45	S100 proteins	S100A1, S100A13, S100A4, S100A6
		Endosome function	ARFIP1, INPP5F, PIK3C3
		Lipoprotein metabolism	APOC1, APOD
		Not accounted for (36/45) = 80%	*ATP6V0B, ATPIF1, BHLHE40, CDKN2A, CKLF, COPS6, CTHRC1, FRMD4A, FXYD3, HOXB2, MGST3, MRPS21, MT2A, NDUFA5, NFKBIZ, PAXIP1-AS1, PLEKHA4, PTPRE, RABAC1, RAMP1, RPS27L, SEMA3B, SH3BGRL3, SLC20A1, STRIP2, TDRKH, TM4SF1, TMEM258, TMSB10, TNFRSF12A, TPD52L1, TSC22D1, TUBA1A, USP53, VGF, ZNHIT1*

Table 1 (continued)

Community	# of genes	Category	Gene names
J	29	Hypoxia	AKAP12, KLHL24, RND3, SAMD4A, SAT1
		Golgi and lysosome	EXOC1, GOLGA8B, LRRK2, PRELP, RAB30, SMPD1, SPG11
		Not accounted for (17/29) = 59%	*ARID5B, BCL2L11, CPQ, CRYL1, CSAG1, EYA3, HIST1H1C, INTU, KDM1B, LPP, RAB17, RP11-258C19.7, SH3BGRL, STARD13, TRIQK, UPF2, ZNF451*
K	20	Cholesterol, sterol biosynthesis	ACSS2, C14orf1, CYP51A1, DHCR24, DHCR7, EBP, FDFT1, FDPS, HMGCR, HMGCS1, INSIG1, LDLR, MSMO1, SC5D, SCD, SQLE, STARD4, TMEM97, LPIN1
		Not accounted for (1/20) = 5%	*PRRX1*
L	12	Interferon response	CEACAM1, HERC5, IFI44, IFIH1, IFIT2, IFIT3, ISG15, PMAIP1
		Not accounted for (4/12) = 33%	*KIN, MRPL55, P2RX7, PPM1K*
M	7	Regulation of growth factor signaling	FAM98B, LEMD3, PTPN1
		RNA splicing	SAP18
		Not accounted for (3/7) = 43%	*GALNT2, MRPL57, SLC2A11*

Communities containing at least 7 genes are shown. Genes were assigned to annotations after manually exploring overlaps with all MSigDB datasets (category labels used here often reflect the merger of redundant or semi-redundant gene sets). Genes marked "Not accounted for" failed to overlap significantly with known gene sets (i.e., overlaps involved at most two genes, or accounted for less than 1% of genes in the dataset)

communities (accounting for 1217 of the 1291 genes), over 50% of the genes could be associated with just one or a few functional annotations (Table 1).

For example, communities A and B consist primarily of genes related to the cell cycle, with more than 62% of the genes in community A known to be preferentially associated with the G2 and M phases of the cell cycle. About 71% of community B is associated with the cell cycle, the majority of these being associated with G1 and S phase. These communities were also correlated with each other, but the community-finding algorithm easily subdivided them.

Most other communities are at best weakly correlated with either of the cell-cycle communities. In community C, 34% of genes encode proteins involved in the unfolded protein response and/or endoplasmic reticulum (ER) stress, and other ER proteins account for another 22% of genes. ER stress is commonly observed in cancer, and the unfolded protein response is specifically and strongly activated in melanoma cells [59–61].

Community D contains almost all genes encoding protein components of cytoplasmic ribosomes, plus a large set of genes that regulate ribosome biogenesis or function. Overall, ribosome-related genes make up 57% of this community. Community E combines genes involved in cellular respiration, glycolysis, and oxidative and hypoxic stress, which collectively account for 78% of this community.

Fifty genes in community F, nearly half the total, are associated with melanin synthesis and melanosome biogenesis, and include traditional melanocyte markers such as *MLANA*, *PMEL*, *RAB27A* and *TYR*. Another subset in community F, consisting of 32 genes, shares annotations related to markers of myeloid cell types (monocytes, macrophages, basophils, etc.) but largely consists of ubiquitously expressed genes involved

Table 2 Mitochondrially-encoded and ribosomal protein gene communities identified following unsupervised clustering of each of the four cell clusters, 1.1, 1.2, 2 and 3

Cell cluster	Community	Number of genes	Gene names
1.1	Mitochondrially-encoded	110	AKAP9, ANKRD37, ARGLU1, ATP1A1, BHLHE40, C21orf58, CD109, CDK5RAP3, CIR1, COPB1, CTSC, DDIT4, DDX17, DDX5, DMTF1, DNAJB1, DST, EIF2A, EIF3D, EIF4A2, EIF4G1, EMP1, EPRS, EWSR1, FAM168A, FNBP4, FUS, GAS5, GOLGA8B, GPNMB, GRN, HERC5, HSP90AB1, HSPA9, HSPD1, IFIT1, IFIT2, IFIT3, IGF2R, ING2, IRF1, ISG15, IVNS1ABP, LAMB1, LGALS3BP, LYST, MACF1, MAN2C1, MCM6, **MT-ATP6, MT-ATP8, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-ND6, MT-RNR1, MT-TD, MT-TF, MT-TL1, MT-TP, MT-TS2, MT-TV,** MXI1, MYC, MYO10, NCOA3, NDUFB9, P4HA1, P4HB, PABPC1L, PDIA3, PDIA6, PMAIP1, PMEL, PNISR, PON2, PPFIA1, PRKDC, PSAP, PTPRM, RASGRP3, RGS1, RRBP1, SAT1, SEC31A, SERINC1, SF3B1, SLC1A3, SLC2A3, SLC35F6, SLC3A2, SNHG12, SORBS2, SPEN, SPTBN1, SRRM2, TAF1D, TLN1, TSR1, TYR, UBR5, VIM, ZFAS1
1.2	Mitochondrially-encoded	59	AC021451.1, ADCY1, AKAP9, BPTF, CALD1, CAPN2, CBLB, COX6A1, CSDE1, DDX17, DDX3X, DST, DZIP3, FCHSD2, FOS, GOLGB1, HECTD1, IGF1R, IGFBP5, JUNB, KRIT1, LRPPRC, MEF2A, **MT-ATP6, MT-ATP8, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-RNR1, MT-RNR2, MT-TF, MT-TL1, MT-TP, MT-TS2, MT-TT, MT-TV,** MXI1, MYCBP2, NBR1, PCLO, PLCB4, PPP1R9A, SPEN, SRRM2, TBL1XR1, TDRD3, TIA1, ZBTB38, ZC3H13, ZKSCAN1, ZMYND8, ZNF704
2	Mitochondrially-encoded	50	APOD, APP, ATP6AP1, BSG, CD109, COX6A1, CPVL, CTSA, CTSD, CTSK, GPNMB, GRN, GUSB, IGSF8, LGALS3BP, LRPAP1, MAGED2, MCUR1, MGST3, **MT-ATP6, MT-ATP8, MT-CO1, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-RNR1, MT-RNR2, MT-TF, MT-TL1, MT-TP, MT-TS2, MT-TT, MT-TV,** NCSTN, NDUFB9, P4HB, PCNXL2, PLAT, PLTP, PMEL, PSAP, RPN2, SPARC, SRPX, TYR
3	Mitochondrially-encoded	15	COX6A1, GPNMB, **MT-ATP6, MT-ATP8, MT-CO2, MT-CO3, MT-CYB, MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND4L, MT-ND5, MT-TL1,** PMEL
1.1	Ribosomal protein genes	48	FTH1, **RPL10, RPL11, RPL12, RPL13, RPL13A, RPL18, RPL19, RPL24, RPL26, RPL27, RPL27A, RPL28, RPL30, RPL31, RPL34, RPL35, RPL37, RPL37A, RPL38, RPL8, RPLP0, RPLP1, RPLP2, RPS11, RPS12, RPS13, RPS14, RPS15, RPS15A, RPS16, RPS17, RPS18, RPS19, RPS2, RPS20, RPS21, RPS23, RPS27, RPS27A, RPS29, RPS3, RPS5, RPS6, RPS8,** SNHG5, TPT1, UBA52
1.2	Ribosomal protein genes	158	ABHD14B, APP, ARL2, ATP5G2, C14orf2, C4orf48, C7orf55, CCDC86, CCDC88C, CD109, CEP350, COMMD6, COX5B, COX7A2L, CST3, CUEDC2, DCBLD2, EEF1A1, EEF1B2, EEF1D, EEF2, EIF3E, EIF3F, EIF3H, EIF3K, ETV5, **FAU**, FBXO32, FIBP, FTH1, FTL, GLTSCR2, GNB2L1, HIF1A, HLTF, IFITM3, IMPDH2, ITSN2, LRRC75A, LRRC75A-AS1, METTL12, MIA, MME, MT-ND6, NAA38, NACA, NACA2, NAP1L1, NDUFB7, NDUFS4, NHP2, NIN, OAZ1, OST4, PABPC1, PABPC3, PGLS, PPFIA1, PRDX5, PRKCSH, PRSS23, RB1CC1, ROMO1, RP11-193E15.4, RP11-356J5.12, RP11-669N7.2, **RPL10, RPL10A, RPL11, RPL12, RPL13, RPL13A, RPL14, RPL15, RPL18, RPL18A, RPL19, RPL22, RPL23, RPL23A, RPL24, RPL26, RPL27, RPL27A, RPL28, RPL29, RPL3, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36, RPL36AL, RPL37, RPL37A, RPL38, RPL39, RPL4, RPL5, RPL6, RPL7A, RPL8, RPL9, RPLP0, RPLP1, RPLP2, RPS10, RPS11, RPS12, RPS13, RPS14, RPS15, RPS15A, RPS16, RPS17, RPS18, RPS19, RPS2, RPS20, RPS21, RPS23, RPS24, RPS25, RPS27, RPS27A, RPS28, RPS29, RPS3, RPS3A, RPS4X, RPS5, RPS6, RPS7, RPS8, RPS9, RPSA,** SAT2, SEC11A, SERF2, SF3B6, SH3KBP1, SNHG19, SNHG5, SNHG6, SSR4, SUCLG2, TOMM7, TPT1, TSPO, UBA52, UBL5, UQCRB, UQCRH, UQCRHL, VPS28, WNK1

Table 2 (continued)

Cell cluster	Community	Number of genes	Gene names
2	Ribosomal protein genes	440	ACTB, ACTG1, AKR1B1, ALDOA, ANAPC13, ANP32A, ANP32B, ANXA5, AP2S1, APOA1BP, APRT, ARL6IP5, ARPC1A, ARPC1B, ARPC2, ARPC3, ATOX1, ATP5B, ATP5E, ATP5EP2, ATP5F1, ATP5G1, ATP5G2, ATP5G3, ATP5I, ATP5J, ATP5J2, ATP5L, ATP5O, ATP6V0B, ATP6V1F, ATP6V1G1, ATPIF1, BCAS3, BCCIP, BOLA3, BTF3, BUD31, C11orf31, C12orf57, C14orf2, C17orf89, C19orf10, C19orf53, C1QBP, C20orf27, C8orf33, CACYBP, CALM1, CALM2, CAPG, CAPZB, CBX1, CBX3, CCNI, CCT7, CD59, CD63, CDC42, CDKN2A, CETN2, CFL1, CHCHD2, CHCHD6, CHCHD7, CHMP2A, CHMP4B, CKS2, CNBP, CNIH1, COA4, COX14, COX17, COX4I1, COX5A, COX5B, COX6B1, COX6C, COX7A2, COX7B, COX7C, COX8A, CST3, CSTB, CTHRC1, CTSH, CWC15, CYC1, CYCS, CYTL1, DBI, DDOST, DGUOK, DNAJB1, DNAJC8, DRG1, DSTN, DTD1, DUT, DYNLL1, DYNLRB1, ECHS1, EDF1, EEF1A1, EEF1B2, EEF1D, EIF1, EIF2S2, EIF3H, EIF3I, EIF4EBP1, EIF5A, EMC4, ENO1, ENY2, ERH, ESD, **FAU**, FBL, FIS1, FKBP1A, FKBP2, FTH1, FTL, FXYD5, GADD45GIP1, GAPDH, GHITM, GLO1, GMNN, GNAS, GNB2L1, GNG12, GPX1, GPX4, GSTO1, GSTP1, GUK1, GYPC, H2AFZ, H3F3B, HDDC2, HINT1, HIST1H4C, HMGB1, HMGN2, HNRNPC, HSBP1, HSPB1, HSPE1, INSIG1, ITGB1BP1, LAGE3, LAMTOR1, LAMTOR4, LAMTOR5, LAPTM4B, LDHA, LDHB, LGALS1, LGALS3, LINC00998, LRRC75A, LRRC75A-AS1, LSM4, LSM7, METTL9, MIA, MIF, MKKS, MLANA, MMADHC, MRPL12, MRPL13, MRPL21, MRPL33, MRPL48, MRPL51, MRPL57, MRPS33, MRPS34, MRPS6, MT2A, MTPN, MYEOV2, MYL6, MZT2B, NAA38, NACA, NAP1L1, NDUFA1, NDUFA12, NDUFA13, NDUFA4, NDUFA5, NDUFA8, NDUFAB1, NDUFB1, NDUFB10, NDUFB11, NDUFB2, NDUFB3, NDUFB4, NDUFC1, NDUFC2, NDUFS5, NDUFS6, NEDD8, NGFRAP1, NHP2, NHP2L1, NME1, NME4, NOL7, NPC2, NPM1, NSA2, NUDT1, NUTF2, OAZ1, OLA1, OST4, OSTC, PA2G4, PABPC1, PABPC3, PARK7, PDAP1, PDCD5, PEBP1, PFDN2, PFDN4, PFDN5, PFN1, PHB, PHB2, PKM, PMP22, POLR2F, POLR2J, POLR2L, POMP, PPDPF, PPIA, PPIB, PPT1, PRDX1, PRDX5, PRDX6, PRELID1, PSMA5, PSMA7, PSMB4, PSMB6, PSMB7, PSME1, PSMG1, PTMA, PTTG1IP, PYURF, RAB2A, RAB38, RAB7A, RAN, RBX1, REXO2, RGS10, RHOA, ROMO1, RP11-669N7.2, RP11-831H9.11, **RPL10, RPL10A, RPL11, RPL12, RPL13, RPL13A, RPL14, RPL15, RPL18, RPL19, RPL22, RPL23, RPL23A, RPL24, RPL26, RPL27, RPL27A, RPL28, RPL29, RPL3, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36, RPL36AL, RPL37, RPL37A, RPL38, RPL39, RPL4, RPL5, RPL6, RPL7A, RPL8, RPL9, RPLP0, RPLP1, RPLP2, RPS10, RPS11, RPS12, RPS13, RPS14, RPS15, RPS15A, RPS16, RPS17, RPS18, RPS19, RPS19BP1, RPS2, RPS20, RPS21, RPS23, RPS24, RPS25, RPS27, RPS27A, RPS27L, RPS28, RPS29, RPS3, RPS4X, RPS5, RPS6, RPS7, RPS8, RPS9, RPSA,** RSL1D1, RSL24D1, RTFDC1, S100A1, S100A10, S100A11, S100A13, S100A6, S100B, SCAND1, SEC11A, SEC61G, SEP15, SERF2, SERP1, SET, SF3B6, SH3BGRL, SH3BGRL3, SHFM1, SKP1, SLC25A3, SLC25A5, SLIRP, SLMO2, SMS, SNHG5, SNHG6, SNHG8, SNRPC, SNRPD1, SNRPD2, SNRPD3, SNRPE, SNRPF, SNRPG, SOD1, SPCS1, SRP14, SRP72, SRSF3, SSBP1, SSR3, SSR4, ST13, STMN1, SUB1, SUMO2, SUMO3, TBCA, TCEA1, TCEAL8, TCEB1, TCEB2, THOC7, TIMP1, TMA7, TMBIM6, TMED2, TMEM14A, TMEM14B, TMEM14C, TMSB10, TMSB4X, TOMM20, TOMM6, TOMM7, TPD52, TPI1, TPT1, TRAPPC1, TSPAN3, TUBA1B, TXN, TXN2, TXNDC17, UBA52, UBB, UBE2L3, UBE2N, UBL5, UBXN1, UQCC2, UQCR10, UQCRB, UQCRH, UQCRQ, USMG5, UTP11L, UXT, VDAC1, VIM, YBX1, YWHAE, YWHAZ, ZNF706, ZNHIT1
3	Ribosomal protein genes	63	ACTG1, ATOX1, CD63, CHCHD2, COX7C, **FAU**, FTH1, FTL, H2AFZ, HSPE1, LGALS3, MLANA, PRDX1, PSMA7, **RPL10, RPL11, RPL12, RPL13, RPL13A, RPL15, RPL18, RPL19, RPL23, RPL24, RPL26, RPL27, RPL27A, RPL28, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL37, RPL37A, RPL38, RPL5, RPL8, RPLP0, RPLP2, RPS11, RPS12, RPS13, RPS14, RPS15, RPS15A, RPS18, RPS2, RPS20, RPS23, RPS24, RPS25, RPS27A, RPS29, RPS3, RPS4X, RPS6, RPS8,** TMSB10, TPI1, TPT1, UBA52, UQCRB

Mitochondrially-encoded and ribosomal protein genes are highlighted in bold

in processes such as lipid biosynthesis. At least one of these genes, *TRIM63*, is known to be strongly associated with melanoma [62], and is a validated target of MITF [63], the primary transcription factor controlling expression of pigmentation genes.

Community G contains multiple genes involved in RNA processing, including seven genes encoding members of the ubiquitously expressed heterogeneous nuclear ribonucleoprotein (HNRNP) family, as well as genes encoding protein chaperones of the HSP10, HSP40 HSP60, HSP90, and CCT families. Community H contains most of the mitochondrially-encoded genes, as well as genes involved in protein ubiquitination and the hypoxia stress response.

Community I combines four genes encoding S100-family calcium binding proteins, as well as various genes associated with endosome function and lipoprotein metabolism. Unlike other communities, in this community most genes do not fall into large groups with traditional annotations. However, many of them are genes known to be strongly associated with melanoma. These include *APOC1*, *APOD*, *CDKN2A*, *CTHRC1*, *FXYD3*, *INPP5F*, *MT2A*, *S100A4, SEMA3B, TMSB10* and *VGF* [64–67], suggesting this community is detecting genes co-regulated by drivers of a melanoma-specific cell state.

Community J combines genes associated with the response to hypoxia with genes involved in Golgi body and lysosome function. Nearly all the genes in community K are associated with cholesterol/sterol biosynthesis and homeostasis; the only exception is *PRRX1*, which serves as the transcriptional co-regulator of serum response factor. In oligodendrocytes at least, it has been shown that PRRX1 is necessary for the expression of cholesterol biosynthesis genes [68].

Community L consists mainly of genes annotated as related to interferon signaling; these genes are also the primary drivers of the cellular anti-viral response. Finally, community M contains several genes associated with growth factor signaling.

Figures S4–S7 (Additional files 6, 7, 8, 9) display the gene–gene linkages within these communities graphically. Here, genes are shown as light blue disks, except for transcription factors which appear as yellow squares. The areas of the disks and squares are proportional to the relative mean expression of the genes (absolute scaling differs between panels, as it was adjusted to enhance the readability of each figure). Green lines denote significant positive correlations (no negative correlations were observed within any of the communities shown). In several cases, a smaller version of each graph, in which links supported by protein–protein interaction data have been overlayed with brown lines, is displayed as an inset; for some communities, only the graphs containing these highlights are shown. Examination of Figs. S4–S7 shows that genes associated with a single functional annotation, or that encode directly-interacting proteins, are often especially densely interconnected, causing them to cluster together (e.g., genes encoding directly physically interacting proteins in A and B, unfolded protein response genes in C, ribosomal subunit genes in D, glycolysis genes in E, S100 genes in I, etc.).

Whereas the clustering of genes into communities had been carried out based on positive gene–gene correlations, plotting pairs of communities together enabled the evaluation of negative correlations between them, the vast majority of which involved the mitochondrially-encoded genes which, as mentioned previously, strongly anti-correlated with ribosomal protein genes (Fig. 5). Mitochondrial genes also anti-correlated with

some of the genes in communities A, F, E and G; for example, anticorrelations between mitochondrial genes and genes encoding glycolytic enzymes are shown in Fig. 5E.

Figure 5 and Table 2 also document that the anti-correlation between mitochondrial and ribosomal genes is as apparent in cell clusters 1.1, 2 and 3 as it is in cluster 1.2. This is interesting insofar as the features used to subdivide cells into these clusters included most of mitochondrially-encoded and ribosomal genes. What this suggests is that there are not two separable populations of cells overexpressing mitochondrial versus ribosomal genes. Instead, the data suggest that the observed anti-correlations are not themselves sufficiently correlated with each other to drive cell clustering. In other words, BigSur seems to able to detect local gene–gene relationships that could not have been identified by comparing differential expression of genes in any possible grouping of cells.

Estimating the accuracy of BigSur

The association of a small number of annotation categories with nearly all of the gene communities that BigSur found, in an unsupervised manner, in scRNAseq data strongly suggest that many of the correlations detected by BigSur are true positives. Nevertheless, in nearly all communities some genes did not fall into obvious categories (Table 1). Do these just reflect the incompleteness of existing annotations, or did BigSur produce false positive results beyond the ~ 2% expected from the FDR cutoff that was used?

Rarely in biology does one have ground truth information with which to settle this question, but one can still perform informative tests. For example, one can consider a subset of the genes that form a functional category and ask whether BigSur correctly identifies other members of the category as being correlated with them. An example is shown in Fig. 6A, where we started with a 27-gene panel, the "Reactome Cholesterol Biosynthesis" gene set from MSigDB (*ACAT2, ARV1, CYP51A1, DHCR24, DHCR7, EBP, FDFT1, FDPS, GGPS1, HMGCR, HMGCS1, HSD17B7, IDI1, IDI2, LBR, LSS, MSMO1, MVD, MVK, NSDHL, PPAPDC2, PMVK, SC5D, SQLE, SREBF1, SREBF2, TM7SF2)*. That panel includes many, but not all, genes known to be involved in cholesterol metabolism. The graph displays all statistically significant positive correlations involving any of these genes that were detected in cell cluster 1.2 (members of the gene panel are indicated with large font, blue lettering).

All detected members of the gene set formed a single network, with many internal positive linkages. Closely connected to this network were eight additional genes (outlined in red) that are not part of the Reactome Cholesterol Biosynthesis set, but belong either to the "Hallmark" cholesterol homeostasis set from MSigDB (*ACSS2, ACTG1, LDLR, SCD, STARD4, TMEM97*); are designated as core cholesterol metabolism genes in recent literature (*ACLY* [69]; *C14or1/ERG28* [70]); or encode sterol binding proteins that serve as feedback controllers of cholesterol biosynthesis (*INSIG1* [71]). Highlighted in orange are two additional genes that regulate lipid biosynthesis more generally, *LPIN1* and *PRRX1*.

Four additional genes are highlighted in gray, as it is possible they are also related to cholesterol metabolism: AKR1A1, CALR, PRDX1, and PMEL. *AKR1A1* encodes an enzyme of the aldo/keto reductase superfamily, which reduces a wide variety of carbonyl-containing compounds to their corresponding alcohols; its orthologues *AKR1C1* and *AKR1D1* are well known to play a role in the metabolism of steroid hormones but

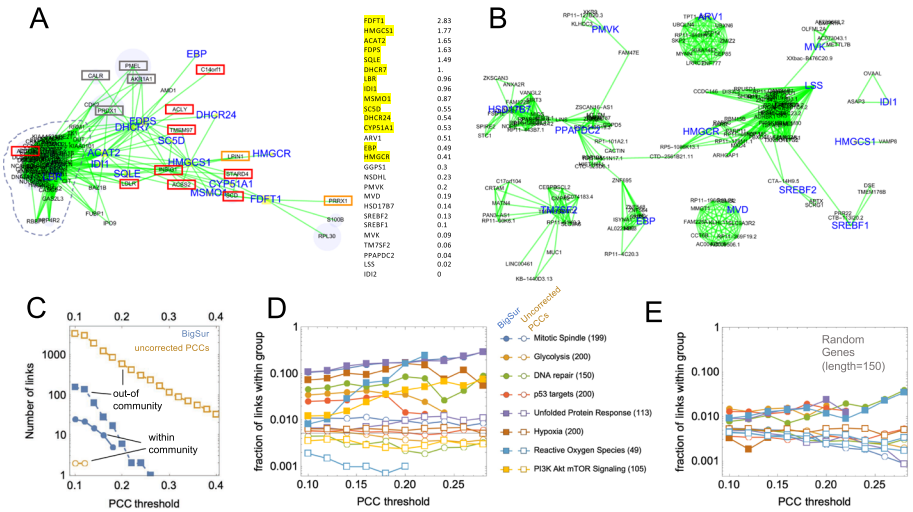


Fig. 6 Comparing the ability of BigSur and uncorrected PCCs to identify biologically significant correlations. Using the data from cell cluster 1.2, BigSur identified positive correlations and their *p* values for all genes, converting the *p* values to equivalent PCCs. In addition, the same starting data were also normalized and uncorrected PCCs obtained. **A**–**D** compare the correlations identified by the two methods. **A** Genes identified by BigSur as correlating with the Reactome Cholesterol Metabolism gene set. All significant gene–gene correlations detected by BigSur involving genes in this 27-gene set (the names of which are shown at right) and all other genes in the genome are shown graphically. Genes belonging to the set are highlighted in blue; additional known cholesterol metabolism-related genes are highlighted with rectangles. Green lines show significant positive correlations. Dashing surrounds a group of cell cycle genes that correlate with the dual-function gene *LBR*. Expression levels (mean UMI per cell after normalization) for each of the genes in the set are shown at right; genes highlighted in yellow are those that displayed any significant correlations. **B** Correlation communities for the same gene set, panel identified using uncorrected PCCs, visualized as in panel (**A**). Note that the target genes are scattered among 5 disconnected communities and are connected with genes with no obvious relationship to cholesterol metabolism. **C** Positive correlations involving any of the genes belonging to the Reactome Cholesterol Metabolism gene set were divided into "within-community" links and "out-of-community" links. With BigSur (filled symbols), the ratio of within-community to out-of-community links is much higher than with uncorrected PCCs (open symbols), suggesting the latter produce much less enrichment for functionally relevant connections. **D** Analyses similar to that in panel C were performed for eight additional MSigDB "Hallmark" gene sets. Plotted are the proportions of correlations that are within-group over a range of PCC thresholds. Numbers of genes in each panel are shown in parentheses. **E** Analyses similar to that in panel (**D**) were performed with 150-gene sets chosen at random after first removing genes with mean expression below 0.052 (so that median expression for random genes was similar to that of the functional panel genes

AKR1A1 is thought to act on non-steroid compounds (perhaps this assumption should be revisited). *CALR*, encodes calreticulin, a protein that promotes folding, oligomeric assembly and quality control in the endoplasmic reticulum. As cholesterol synthesis takes place in the ER membrane, it is perhaps not surprising that expression of *CALR* and cholesterol synthesis genes would be co-regulated. *PRDX1* encodes peroxiredoxin, an antioxidant enzyme that reduces hydrogen peroxide and alkyl hydroperoxides and plays a key role in maintaining redox balance. In macrophages, PRDX1 has been shown to be critical for cholesterol efflux during autophagy [72]. *PMEL* encodes a major component of the melanosome. Interestingly, it has been reported that cholesterol strongly stimulates melanogenesis in melanocytes and melanoma cells [73].

In addition to these genes, a set of tightly interconnected genes that have no known association with cholesterol metabolism is peripherally connected to this community (outlined by a dotted line in Fig. 6A). A large proportion of these genes encode proteins

associated with the cell cycle, especially with functions associated with mitosis, such as mitotic checkpoint events, and cyto- and karyokinesis. Strongly connected to these mitotic genes is *LBR*, which encodes the Lamin B receptor. LBR is a multifunctional protein: it plays a key role in cholesterol biosynthesis, catalyzing the reduction of the C14-unsaturated bond of lanosterol, but also anchors the nuclear lamina and heterochromatin to the inner nuclear membrane, and in so doing is strongly regulated by phosphorylation during the cell cycle [74]. A functional link between expression of cholesterol biosynthesis genes and mitosis may arise from that fact that cholesterol synthesis normally rises dramatically during G1 phase and if prevented from doing so will result in G1 arrest [75]. It thus may make sense to upregulate the production of cholesterol biosynthetic genes as cells go into mitosis, so they are available to act in the G1 phase that immediately follows.

To the right of the graph in Fig. 6A the genes of the Reactome Cholesterol Biosynthesis set are listed alongside their relative expression levels (UMI/cell after normalization) in this data set. Highlighted in yellow are those for which BigSur identified significant positive correlations, nearly all of which were toward the higher end of expression levels. These results illustrate an important limitation of any correlation methodology, which is that statistical power depends on the number of non-zero entries in the data, which will, in turn, reflect expression level, sequencing depth, and the number of cells analyzed. Examination of the raw data indicates that the point at which BigSur begins to fail to identify significant correlations occurred, for these genes, when the total number of non-zero entries (among 1582 cells) fell to about 320; this corresponded to a mean expression level of about 0.3 UMI/cell. For more strongly correlated genes one should of course expect fewer entries to be necessary to achieve significance, but based on the results here we suggest that, when seeking to identify the kinds of weak correlations associated with noise-coupled gene-regulatory networks, a minimum of several hundred non-zero entries per gene may be a reasonable threshold. In practice, one should be able to adjust the number of genes that fulfill this criterion either by varying the number of cells analyzed or varying the depth of sequencing (or both).

Figure 6B plots the results of the identical exercise as in Fig. 6A, carried out using uncorrected PCCs obtained directly from normalized gene expression data, rather than using PCC′ and BigSur. A threshold of PCC = 0.29 was selected so that the number of Reactome Cholesterol Biosynthesis genes that exhibited above-threshold correlations was the same as in Fig. 6A. In this case, however, the cholesterol biosynthesis genes did not form a single community, but rather separated into 5 disconnected groups. Most groups contained links to many genes, none of which appeared, on inspection, to bear any relationship to cholesterol metabolism. These data make the important point that *most of the links detected using uncorrected PCCs are likely to be false positives.*

One way to extend this analysis to other functional gene sets is suggested by the observation that, among the correlations in Fig. 6A involving Reactome Cholesterol Biosynthesis genes, 25 occur between member genes ("within-community") while 157 involve other genes ("out-of-community"). In contrast, in Fig. 6B there are 143 out-of-community links and zero in-community ones. Figure 6C shows how the numbers of within-community and out-of-community links change as the PCC significance threshold is changed, both for BigSur and uncorrected PCCs. These results suggest that the fraction

of all correlations that is in-community can be used as a measure of the specificity with which functionally relevant correlations are identified by any given method. In Fig. 6D, we used this approach to contrast the performance of BigSur (filled symbols) with uncorrected PCCs (open symbols) for 8 different gene sets, each of which contained between 49 and 200 members. Using BigSur, the proportion of links that were in-community varied between ~ 1% and 30% depending on the gene set, usually increasing as the equivalent PCC threshold was made more stringent. Using uncorrected PCCs, the proportion of in-community links varied between 0.2% and 1% and did not change appreciably with PCC threshold. For comparison, Fig. 6E performs a similar analysis using 8 "random" gene sets—sets of 150 genes randomly selected from genes with similar overall expression to those in the gene sets used in Fig. 6D. The results suggest that the performance of BigSur on functionally related genes is far above chance level, while uncorrected PCCs perform at or around that level.

Induced changes in gene expression are associated with increased gene–gene correlation

If BigSur detects physiologically meaningful gene correlations, one might expect genes induced in response to specific perturbations to be correlated with each other. In other words, among genes that differ *between* treated and untreated conditions, one might also expect to observe groups of the same genes displaying correlations *within* each individual condition—presumably because they represent genes regulated by common upstream signals. To test this, we turned to a scRNAseq study in which differentiated human airway epithelial cells were treated with IL-13, a model of cytokine-induced asthma [76]. We focused on a single subset of cells, those labeled "defense secretory", as similar numbers of them were captured in both treated and untreated conditions (see Methods).

Data from control and IL13-treated cells were then analyzed separately using BigSur (Fig. S8/Additional file 10). First, we focused on the 419 genes reported to be upregulated by IL13 [76]. Within the treated group we found 313 of these genes were also correlated with each other in a single, large community (Fig. S8A). Interestingly, 45 of these genes were correlated in control cells as well (Fig. S8B), suggesting that a module of co-regulated gene expression is already active prior to treatment.

Next, we asked whether BigSur could identify co-regulated genes that had not been picked up as differentially expressed (perhaps because their expression levels did not change sufficiently after treatment). We focused on genes that correlated with one particular gene, MUC16, which encodes an IL13-induceable mucin thought to play a prominent role in asthma-associated mucus obstruction of the airway [76]. We found 168 genes to be significantly positively correlated with *MUC16* in IL13-treated cells (Fig. S9/Additional file 11), only 23 of which were among those differentially expressed in response to IL13 [76]. Among the remaining 145 genes were found *MUC4*, another mucin gene; *SYTL2*, a paralog of *SYT2*, which encodes a limiting factor in mucin secretion [77]; and *THSD7B*, which encodes a direct interactor of *SYT2* and *SYTL2*.

Also among the genes that correlated with *MUC16* in IL13-treated cells were 5 of 11 genes previously identified by GWAS as associated with risk of childhood asthma: *PDE4D*, *PTPRD*, *RBFOX1*, *ROBO2*, and *RYR2* [78]; *MUC16* also correlated with *ROBO1*, which encodes an interaction partner of *ROBO2*. The top-ranked GWAS gene, *RYR2*,

encodes a calcium channel thought to play an important role in asthma pathogenesis [78]. Neither RYR2, nor any of the other 11 GWAS associated genes, were detected as differentially expressed after IL13 treatment, although two genes encoding proteins that interact with RYR2 (*CALM1* and *CMK2D*) were [76]. These data suggest that the analysis of gene correlations has the potential to substantially augment biological insights that come from studies of differential gene expression.

Aggregation into "meta-cells" impedes the identification of gene correlations

The ability of BigSur to detect biologically relevant gene correlations with controllable false discovery reflects its ability to handle the highly skewed distributions of scRNAseq data, distributions that are usually dominated by many zeros. In the scRNAseq literature, several other approaches have been proposed to compensate for data sparsity, such as filling in zeros by imputation (estimating values based on closely related cells) or aggregating groups of cells (merging measurements in similar cells, or across technical replicates of the same cells). Aggregation essentially replaces groups of cells with a smaller number of "metacells" or "pseudocells", essentially implementing a small-scale version of "pseudobulking" [79–83]. This approach has proved useful in improving the accuracy of identifying differentially expressed genes, and in analyses in which genes are clustered into weighted gene co-expression networks [84, 85].

To the extent that aggregation reduces data sparsity, one might expect it also to improve accuracy in discovering gene–gene correlations. In fact, at least one group [86] has proposed this explicitly, developing a method that starts with normalized, log-transformed scRNAseq data, and averages values among groups of cells that are judged sufficiently similar (based on Leiden clustering). Using this method, they claimed that one can enhance the ability to detect meaningful correlations in scRNAseq data.

We investigated this claim, however, and came to the opposite conclusion. We discovered that grouping cells into metacells markedly inflates false positive correlations, a phenomenon not explored by [86]. This effect is easily demonstrated by applying their procedure to synthetic data in which gene expression values were sampled randomly and independently from Poisson-lognormal distributions (coefficient of variation of underlying lognormal distribution = 0.5), with cells sequenced to different depths, as in Figs. 1 and 2. Using such data as a starting point, we calculated correlation coefficients six different ways (Fig. 7A–C), using the Fischer formula to define the threshold for statistical significance (with PCC′ we also used a Benjamini–Hochberg adjusted FDR threshold of < 2%, as determined by BigSur). In such a data set, an accurate method should detect *no* significant correlations, positive or negative.

As expected, using PCCs calculated directly from raw, unnormalized data, nearly 100 million artifactually significant positive correlations were identified, and these were not eliminated by data normalization (scaling UMI values to sequencing depth). Log-transformation of the data (using a "pseudocount" of 1) reduced the number of false positives to 382,511, probably because it greatly reduced the range of variability between high and low values. In contrast, when log-transformed values were grouped and averaged according to the procedure of [86], clustering the 3570 cells and grouping them into 100 metacells, we observed a massive inflation of large positive and negative values of PCC (Fig. 7B, C), among which over 2 million were judged significant by the Fischer formula

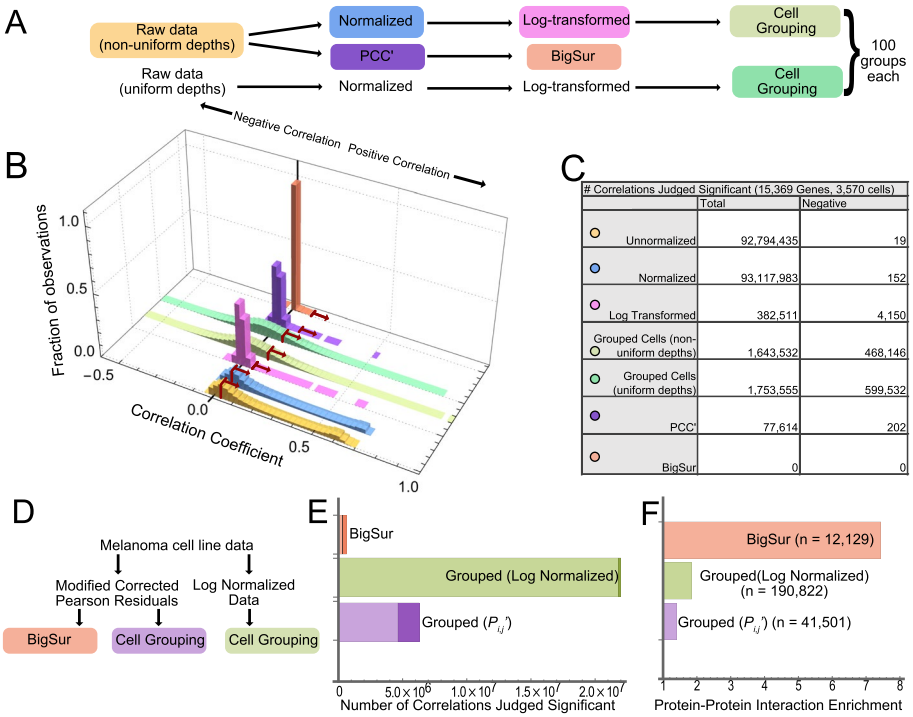


Fig. 7 Grouping into "metacells" creates false positive correlations. **A**–**C** Analysis of synthetic, uncorrelated data. **A** Pipelines for data processing. Each box denotes a step at which a Pearson correlation coefficient (PCC or PCC′) was calculated, with the color of the box corresponding to the colors used in the following plots. **B** Histograms of correlation coefficients obtained from the data sets in panel (**A**). Arrows show the thresholds above which observed correlations were judged statistically significant. **C** Numbers of correlations in panel (**B**) that were judged to be significant ($p < 0.02$). **D**–**F** Analysis of melanoma cell line data. **D** Pipeline for data processing. The colors of each box correspond to the colors in panels (**E**, **F**). **E** Number of correlations judged significant in data sets in D. Darker shading denotes the negative correlations; lighter are positive correlations. **F** Enrichment for known protein–protein interactions among the correlations shown in (**E**). The value of *n* in each case gives the absolute number of protein–protein interactions. Of the 12,129 interactions detected by BigSur, 11,373 were also detected using grouping of log-normalized data and 10,773 were detected using grouping of modified corrected Pearson residuals (10,324 were shared among all three)

(the formula automatically takes into account the reduced number of cells). This was not just a consequence of failure to compensate for variable sequencing depth, because even when we created data with equal sequencing depth across all cells, we still observed that grouping into metacells produced just as many false correlations (Fig. 7B, C).

In contrast, PCC′, calculated from unnormalized data, did much better, producing fewer than 78,000 false correlations, as judged using the Fischer formula, and no correlations, positive or negative, when the FDR was appropriately calculated using BigSur.

These results indicate that, even if grouping cells increased the likelihood of detecting true correlations, the effect would likely be overwhelmed by the creation of so many false ones. To assess the net effect, we compared the method of [86] with BigSur using the full set of melanoma cells analyzed in Fig. 3. As diagramed in Fig. 7D, we performed grouping both on Log-transformed normalized data, as per [86], and on modified corrected Pearson residuals (P_{ij}′), hoping that the latter might better correct for variable sequencing depth. Consistent with what was seen with synthetic data, grouping produced many more correlations than BigSur, although the use of modified corrected Pearson residuals reduced this to some extent (Fig. 7E).

To assess the recovery of true positives, we quantified enrichment for pairs of genes with known protein–protein interactions (Fig. 7F), as such genes are more likely than random genes to be co-expressed. As described earlier, such interactions were enriched more than seven-fold among the pairs of genes found significantly positively correlated by BigSur. In contrast, even though grouping into metacells identified significant correlations among more total gene pairs with known protein–protein interactions, the enrichment over what would have been expected by chance was less than two-fold, suggesting an accuracy barely better than chance levels.

The reason grouping generates so many false positive correlations appears to be generic to the process of aggregating cells by similarity, and not specific to details of the method of grouping of [86], as we can see similar effects using synthetic data and other, simpler algorithms for grouping cells. These observations nicely illustrate how, when dealing with correlations, even seemingly innocuous empirical fixes can create unexpected problems, and underscore the value of BigSur in provided a principled, theory-grounded approach to assessing both the magnitude and significance of gene–gene correlation.

Discussion

The above results suggest that *bona fide* communities of co-regulated genes can be identified with high specificity by carefully mining weak correlations within groups of a thousand or fewer relatively homogeneous cells. BigSur achieves the accuracy to do this first by correcting measures of correlation for unequal sequencing depth and the added variance contributed by gene expression noise, and subsequently by estimating an individual p value for each gene pair—thereby overcoming the strong effect of gene expression distribution on the likelihood of correlation arising by chance. In contrast, the use of uncorrected PCCs to identify co-regulated genes performed poorly on both synthetic and real scRNAseq data, as did at least one method of compensating for data sparsity by cell aggregation.

In developing BigSur, we sought to avoid normalization steps and expression thresholds and to minimize user-provided parameters to the greatest extent possible. The major user input to BigSur, besides a UMI matrix, is a coefficient of variation for gene expression noise, c, which can be quickly estimated by fitting a plot of the modified corrected Fano factor against gene expression level (Fig. 2B). In reality, the magnitude of the noise of gene expression may be different for different genes, and for some the Poisson log-normal distribution may not be the best approximation of the noise. Although these factors likely degrade the performance of BigSur, we note that the value of c only significantly impacts the highly expressed genes, for which (due to low sparsity) the detection of significant correlations is a less challenging task.

A more subtle source of potential error comes from fact that, in calculating modified corrected Pearson residuals, the value of μ_{ij} used by BigSur is determined empirically from a finite set of cells, i.e., it is an estimator of μ_{ij}. Furthermore, whereas it is an unbiased estimator when μ is Poisson-distributed, this is not generally the case for more skewed distributions, such as Poisson-lognormal [43]. It is unknown whether these sources of error have much impact on the determination of correlation coefficient p values by BigSur, and additional work will be necessary to investigate this question.

Despite these concerns, the ability of BigSur to identify gene communities that are closely related in function (Table 1), as well as add new, functionally related genes into known gene sets (Figs. 6, S9), suggests that it already operates at a level of performance that can be useful to cell and tissue biologists. Particularly interesting are the questions it raises about linkages within communities—for example, why do genes encoding cystatin endopeptidase inhibitors (*CST1*, *CST4*, *CSTB*) correlate with genes involved in glycolysis and cellular respiration (community E)? Why do genes encoding protein chaperones correlate with genes involved in RNA processing (community G)? Why do genes involved in Golgi and lysosome function correlate with genes related to hypoxia (community J)?

The data in Table 1 also suggest that new cell biology may be discovered by examining genes labeled "not accounted for", i.e., genes that are associated with a community but do not, as a group, overlap substantially with any mSigDB dataset. For example, 38 of the 213 genes that correlate with cell cycle community B are not currently annotated as cell-cycle related. Among the strongly-coupled unfolded-protein response genes in community C, one also finds genes encoding secreted (*GDF15, PSAP, SPARC*) and cell surface molecules (*DDR2*, *GPNMB*, *ITGA4, IL1RAP, PMP22, SLC1A5*); the potential relationship of these genes to cellular stress responses may deserve heightened attention. Indeed, each of the communities in Table 1 suggests new and unexpected forms of co-regulation of gene expression. It will be interesting to see how many of these are reproduced in other cell types—a task that can be efficiently approached by applying BigSur to the large number of existing scRNAseq data sets.

It is instructive to note that few of the gene–gene relationships detected by BigSur could have been revealed by the typical analytical steps of cell clustering and identification of differentially expressed genes, as most of the gene communities identified here are not associated with sufficient total variation in gene expression to be useful drivers of cell clustering. On the other hand, clustering did play an important role here in reducing the heterogeneity of the sample to which BigSur was applied. Although most of the gene communities that were identified using the 1582-cell subcluster 1.2 were also observed when BigSur was applied to the entire sample of 8640 cells (not shown), clusters were more difficult to visualize and analyze in the latter case, thanks to a large background of cell-type (or cell-state)-specific gene expression (which generated numerous additional correlations). This experience suggests that iterative application of BigSur analysis and clustering (potentially using correlated gene communities as features) can provide a useful pipeline for identifying meaningful gene communities of manageable size.

It is interesting that one of the strongest axes of variation we detected in this study involved mitochondrially-encoded genes and genes coding for ribosomal subunits, with both communities strongly anti-correlating with each other (both before and after "damaged" cell removal and subclustering). Because the mitochondrial community consists only of mitochondrially-encoded, and not cytoplasmically-encoded, mitochondrial genes, the simplest interpretation is that this community reflects cell-to-cell variation in the number of mitochondria (or, more accurately, mitochondrial genomes). What then might explain anti-correlation between mitochondrial number and transcripts for ribosomal proteins? Given that protein synthesis requires both specialized machinery (ribosomes) and a source of energy (mitochondria), one might expect to see positive, rather than negative, correlation between the agents that orchestrate these processes. Yet this

intuition is correct only in a long-time-averaged sense and does not necessarily apply if demand for protein synthesis fluctuates. In mammalian cells, ribosomal protein mRNAs are long-lived, with half-lives in the 5–10 h range [87], whereas mitochondria can replicate on a time scale of 1–2 h [88]; accordingly, one process may systematically lag the other, producing the kind of anti-correlation we observe here. While this interpretation is speculative, it demonstrates how the analysis of gene–gene correlations can motivate new hypotheses about transcriptome-scale regulation of cell biology.

Under the expectation that most networks of gene correlation reflect shared transcriptional regulation, we might have expected to identify upstream transcription factors more frequently in most of the networks we discovered. In some cases, this clearly did occur: For example, the transcription factors *ATF4*, *AT6*, and *XBP1*, which were detected in community C, are known controllers of the unfolded protein response, the components of which show up in the same community. Transcription factors related to cell cycle progression and DNA damage-repair strongly associated with cell-cycle communities A and B. *ZKSCAN1*, a transcription factor targeted to mitochondria [89] associates with the mitochondrially-encoded gene community H. However, in many cases, expected transcription factors (e.g., *SREBF1* or *SREBF2* in community K) were not observed. This may reflect limitations in statistical power, as transcription factor genes tend to be expressed at a somewhat lower level than other genes—although in this data set expression of the average observed transcription factor was only about half that of the average observed gene. Another likely explanation is that gene regulation is often achieved through the post-translational modification of transcription factors, rather than regulation of their mRNAs. Alternatively, it may reflect the importance of factors that act post-transcriptionally, such as miRNAs (which were not assessed in this data set), in controlling gene expression. In future, it will be important to extend BigSur to take account not only of miRNAs but also of "multi-omic" features, such as chromatin accessibility.

Although we have focused here primarily on the use of BigSur as a tool for discovery of gene–gene correlations, it is worth pointing out that the intermediate steps in the BigSur pipeline produce useful tools for other analytical procedures, some of which were exploited here. For example, methods for feature selection for cell clustering commonly involve thresholds (e.g., expression levels) and cutoffs (e.g., numbers of features) that are arbitrary, and may not be ideal choices for every data set. Use of the modified corrected Fano factor ϕ', and its associated p values, can provide a less arbitrary approach to feature selection, which can outperform other methods on challenging tasks, such as finding rare cell types, or subclustering cells that differ only modestly [90]. In addition, as we did here when dividing cells into subclusters based on the expression of ribosomal and mitochondrial genes (Fig. 4), one may also use communities of correlated genes themselves as features for clustering—in this way leveraging not just variation but co-variation to drive clustering. Finally, whereas it is common practice to cluster cells based on their normalized expression values, the matrix of modified corrected Pearson residuals that BigSur calculates almost certainly provides a more accurate starting point for clustering, as it avoids artifacts introduced by normalization, and corrects for the inflated variance associated with highly expressed genes.

Conclusions

It has long been suspected that, by mining gene expression correlations within single cell transcriptomes, it should be possible not only to identify groups of genes that distinguish among cell types, but also discover small-scale gene regulatory networks that operate within cell types. We have provided evidence here that this goal can be achieved, first by modifying and correcting the metric of correlation and then using an analytical approach to assign statistical significance to every gene pair. When applied to scRNAseq data, this dramatically reduced the number of false positives that would have been identified by other methods, and enabled the identification of biologically relevant gene communities, both known and novel.

Methods

As the first step in calculating ϕ' and PCC$'$, we begin with "raw" (neither normalized nor log-transformed) UMI data and calculate, for each gene in each cell, a cell- and gene-specific Pearson Residual, defined according to Eq. 1. To do so, we first calculate a cell- and gene-specific expected value (μ_{ij}) which we obtain for each gene by averaging its values over all cells, then scaling that in each cell by to the relative proportion of total UMI that each cell contains. This is essentially the same procedure used in the simplest form of normalization, except that, rather than normalize the data matrix, we are normalizing the term for the mean in the Pearson residual.

Next, each modified Pearson residual is divided by $\sqrt{\left(1+\mathrm{c}^2\mu_{ij}\right)}$, where c is a constant between 0.2 and 0.6. Ideally, c should be chosen individually for each gene, depending on prior knowledge of the level of gene expression stochasticity, however, in the absence of prior knowledge we typically find c empirically (see Fig. 2B). It should be noted that, for many scRNAseq data sets, most values of μ_{ij} will be <1, meaning the effect of the choice of c on most of the data is often relatively small. To obtain ϕ' for each gene, the Pearson residuals for each cell are squared, summed, and divided by $n-1$, where n is the number of cells.

To calculate p values for ϕ' any given gene, we consider the null hypothesis to be the expected number of transcripts in each cell is μ_{ij}, i.e., the total number of transcripts across all the cells partitioned in proportion to the number of total UMI in each cell. As noted above, because BigSur determines the value of μ_{ij} empirically—by summing up genes UMI across all cells and multiplying by the fraction of total UMI in each cell—μ_{ij} is technically an estimator of the cell-specific expectation value for each gene and cell.

To calculate p we need to know how the sums of squared Pearson residuals should be distributed for any given set of μ_{ij} and c. As discussed above, we take μ_{ij} to have a Poisson-log-normal distribution, allowing us to calculate the moments of the distribution of squared Pearson residuals from the moments of the Poisson and log-normal distributions, and from there the moments of the distribution of sums of squared Pearson residuals. In the end we obtain, for each gene j, a finite set of moments (typically 4 or 5) for the distribution of ϕ' that would be expected under the null hypothesis for that particular gene, given the values $\mu_{1j}, \mu_{2j}, \mu_{3j} \ldots \mu_{nj}$ and c. We then use Cornish-Fisher approximation of the Edgeworth expansion [91] as a reasonably computationally efficient way to approximate the p value associated with any given observation of ϕ', given that distribution.

The procedure for obtaining PCC′ proceeds in the same fashion, starting with the same modified corrected Pearson residuals, but now taking the dot product of the vectors of Pearson residuals for each pair of genes, and dividing by $n-1$ times the geometric mean of the ϕ' values for those genes (Eq. 2). Moments of the expected distributions of PCC′ are calculated analytically in exactly the manner described above, with the slight complication that the ϕ' terms in the denominator are not strictly independent of the Pearson residuals in the numerator, but to a good first approximation may be treated as such, as they aggregate information across all the Pearson residuals. The Cornish-Fisher approximation is again used, as described above, to assign p values to both tails of the resulting distribution. In solving the 4th degree polynomial equations produced by this method (a computationally slow step), we improve speed by sacrificing accuracy specifically for those p values that clearly fall below thresholds of interest.

Because the number of possible gene–gene correlations scales roughly as the square of the number of genes, the need to correct for multiple hypothesis testing is particularly acute. Given that the observations are not independent from one another, Bonferroni correction is clearly too conservative, and thus we use the Benjamini–Hochberg algorithm [49] for controlling the false discovery rate.

Simulating scRNAseq data

In Fig. 2, we simulate the expression of 1000 genes across 999 equivalent cells, where by equivalent we mean that, for each gene, a single "target" transcript level was chosen from a log-normal distribution, the mean of which was selected so that the logarithm of its value varied uniformly across the gene set, over the range from 0.035 to 3467 transcripts per cell. The coefficient of variation of each log-normal distribution was taken to be 0.5 for all genes. Next, we generated a set of 999 scaling factors, drawn from a log-normal distribution with mean of 1 and coefficient of variation of 0.75, selected so that the logarithm of the value varied vary smoothly across the range. Finally, we generated a UMI value to each gene in each cell that was a random variate from a Poisson distribution with a mean equal to the target for that gene times the scaling factor for that cell. The result was a set of gene expression vectors of length 999, with mean values varying between 0.001 and 231.

In Fig. 7A–C, synthetic data were generated as in Fig. 2, except that the number of genes was increased to 15,369 and the number of cells increased to 3570. Grouping cells into 100 metacells was carried out as described by [86]. In Fig. 7D–F, the full melanoma cell dataset was used (17,451 genes × 8640 cells; see below), and grouping was again performed to create 100 metacells.

Analysis of melanoma cell line data

Data from droplet-based sequencing of subcloned WM989 melanoma cells (GEO accession GSE99330), which had been pre-processed to remove UMI judged not to be associated with true cells, were imported and further pre-processed in the following way: First we removed all known pseudogenes (comprehensive lists of human pseudogenes were obtained from HGNC and BioMART). Pseudogenes derived by gene duplication or retrotransposition are often highly homologous to their parent genes, creating ambiguity in

the mapping of the short-read sequences used in scRNAseq. When pseudogenes were not removed from analysis, we frequently detected strong correlations between pseudogenes and parent genes that very likely represented the effects of ambiguous mapping, rather than true correlation. After pseudogene removal, the number of detected genes was 27,526. As both theory (Fig. 1) and experience indicated that statistically significant correlations were usually unobservable for genes with UMI in fewer than 0.001% of cells, we further eliminated genes expressed in fewer than 8 of the 8640 cells; this reduced the size of the gene set to 17,451, and the number of possible unique correlations to 152,259,975 (while not strictly necessary, this step reduces computational time by ~ 2.5 fold, since the number of correlations to test varies approximately quadratically with the number of genes).

Analysis of cultured airway epithelial cell data

Data from a study comparing untreated and acutely IL13-treated cultured human airway epithelial cells were downloaded from GSE145013 [76]. To confine analysis of correlations to a relatively homogeneous cell type, we subsetted only those cells clustered into groups composed primarily of the "secretory" cell type; specifically, related clusters c5 and c6 contained most of the secretory cells of the treated and untreated samples, respectively, and analysis was confined to cells of these clusters. This resulted in a dataset of 407 IL13-treated 303 untreated cells. After removal of pseudogenes, and gene with fewer than 80 total UMI, each dataset contained the same set of 15,268 genes.

Calculating paralog pair and protein-interaction enrichment scores

A curated list of 3132 paralogous pairs of human genes was downloaded from [92]. A list of physical human protein–protein interactions was downloaded from BioGrid (https://downloads.thebiogrid.org/File/BioGRID/Release-Archive/BIOGRID-4.4.218/BIOGRID-MV-Physical-4.4.218.tab3.zip) and supplemented with additional data from HIPPIE v2.3 (http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/), to produce a list of 790,008 gene pairs. To calculate enrichment, we first calculated the fraction of statistically significantly positively correlated gene pairs identified by BigSur that overlapped with either the paralog-pair or protein–protein interaction pair database. Next, we removed from the databases all gene pairs involving genes not detected in the scRNAseq data and divided the remaining number by the total number of possible gene–gene correlations (i.e. $m(m-1)/2$, where m is the number of genes in the scRNAseq data) to yield the expected frequency of paralogous or interacting pairs under the hypothesis they are randomly distributed among all possible pairwise correlations. The ratio between the observed frequency and expected frequency was considered to be the fold enrichment.

Extracting (and pooling) gene communities

BigSur generates a matrix in which rows and columns are genes, and entries are signed equivalent PCCs—which are derived by using the inverse of the Fisher formula on the p values returned by BigSur, together with the signs of the values of PCC′. Although it is a derived quantity, the equivalent PCC is a useful form in which to store correlation data, not only because it is signed, but also because it adjusts for differences in data length

(number of cells), so that similar "strengths" of correlation would be expected to translate into similar equivalent PCCs, even across samples with very different numbers of cells (cell number has a large effect on the relationship between correlation strength and p value).

Only equivalent PCCs that were judged statistically significant according to a user-supplied threshold (e.g., a Benjamini–Hochberg FDR) were included, all others being set to zero. This matrix was converted to an unweighted adjacency matrix (all nonzero values replaced with 1) and the *walktrap* algorithm (with a default setting of 4 steps) was used to identify initial communities [52]. Because this produces communities connected by both positive and negative links, each community was then subjected to a second round of community-finding, after first setting negative links to 0, thus allowing sub-communities that negatively correlate with each other to be separated.

To identify instances in which walktrap had subdivided communities too finely, we manually examined the number of positive correlations between genes in each community and each other community, recursively merging communities in which the number of inter-community correlations was particularly large (compared with the number of possible links between communities). In addition, in rare cases in which communities returned were very large (e.g., in the thousands), we subdivided them by applying walktrap an additional time.

Cell clustering based on correlated features

Feature selection refers to the process of identifying genes that capture important dimensions of variation on which cells may be clustered. A variety of approaches have been proposed for identifying such genes, and many work equivalently under most circumstances (with tens of thousands of genes, clustering is often a highly over-determined problem). Under challenging circumstances (e.g. when the number of true difference separating clusters is small, or the number of cells in a state is small), we have shown [90] that ϕ' is a measure of variability at least as good as others, and because BigSur returns both ϕ' and p values, one may avoid selecting too large a set of features (which can defeat clustering algorithms).

It has also been pointed out, however, that not only the statistical features of individual genes, but also their interdependencies (i.e., correlations) should ideally be used to inform clustering [24]. We recognized that the communities identified by BigSur represent ideal sources of features, particularly if we emphasize those community members that are the most highly connected to each other. We also recognized that the modified corrected Pearson residuals generated by BigSur provide a more sensible set of vectors to use as the input to clustering than either raw or normalized UMIs (for the same reasons that ϕ' and PCC$'$ are improvements over their unmodified, uncorrected forms). Briefly, we used the modified corrected Pearson residuals to construct shared nearest neighbor graphs based on the top 50 principal components, with $k=20$ and a Jaccard similarity threshold of 1/15; this was followed by Leiden clustering [93] and UMAP visualization, essentially as in the Seurat analysis package [94].

With this approach, we found that well-defined clusters can often be reliably obtained using sets of as few as 50–75 highly connected genes. This was used to repetitively

subcluster the scRNAseq data from WM989 melanoma cells, at each step running BigSur and using the 50 most connected genes of the clusters containing the majority of ribosomal genes, together with the 50 most connected genes of those containing the majority of mitochondrially encoded genes, as features.

Graphical display of correlations

Matrices representing statistically significant correlations were plotted using the *GraphPlot* function of Mathematica software, in which vertices were arrayed either by Spring Embedding or Spring Electrical Embedding. Edges were colored green when positive and red when negative. Vertex locations were first determined according to the graph produced after deleting negative edges, after which vertices connected only by negative edges were added in. Symbols used to represent vertices were scaled so that their areas were proportional to the mean expression level of the gene represented. Correlation strengths (the absolute values of the equivalent PCCs) are not represented on these images.

Abbreviations

BigSur	Basic Informatics and Gene Statistics from Unnormalized Reads
CV	Coefficient of variation
ER	Endoplasmic reticulum
FDR	False discovery rate
PCA	Principal component analysis
PCC	Pearson correlation coefficient
PCC′	Modified, corrected Pearson correlation coefficient
ROC	Receiver operating characteristic
scRNAseq	Single cell RNA sequencing
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05926-z.

Additional file 1: Table S1. Analysis of gene communities using the Database for Annotation, Visualization and Integrated Discovery. Gene communities in Table 1 were analyzed using DAVID functional analysis [58] and the following databases: UP_KW_BIOLOGICAL_PROCESS, UP_KW_CELLULAR_COMPONENT, UP_KW_MOLECULAR_FUNCTION, UP_KW_PTM, UP_SEQ_FEATURE, GOTERM_BP_DIRECT, GOTERM_CC_DIRECT, GOTERM_MF_DIRECT, BBID, BIOCARTA, and KEGG_PATHWAY. Results shown are those that met default threshold requirements to display in the Functional Annotation Chart. Results for each community are given on a separate spreadsheet.

Additional file 2: Appendix. Mathematical formulae and derivations

Additional file 3: Figure S1. Significance of uncorrected Pearson correlation coefficients, as calculated by BigSur versus the Fisher formula, binned by gene expression. scRNAseq data were as described in Fig. 3. Data points representing pairs of genes were divided into 21 bins based on the mean expression levels of each gene, and the results for each bin were plotted as described in Fig. 3C. The abscissa shows PCCwhile the ordinate gives the negative log10 of *p* values determined by BigSur, i.e., larger values mean greater statistical significance. Orange and gray shading indicate gene pairs judged significant by BigSur. Blue and orange show gene pairs that would have been judged statistically significant by applying the Fisher formula to the PCC, using the same p-value threshold as used by BigSur. The blue region contains gene pairs judged significant by the Fisher formula only, while the unshaded region shows gene pairs not significant by either method. Numbers in the lower right corner of each panel are the total numbers of possible correlations, statistically significant correlations according to the Fisher formula, and statistically significant correlations according to BigSur.

Additional file 4: Figure S2. Significance of modified corrected Pearson correlation coefficients,, as calculated by BigSur versus the Fisher formula, binned by gene expression. scRNAseq data were as described in Fig. 3. Data points representing pairs of genes were divided into 21 bins based on the mean expression levels of each gene, and the results for each bin were plotted as in Fig. S1. The inset compares the PCC′-*p* value relationship determined by BigSurwith that predicted by the Fisher formula, showing that, for highly expressed genes, the two methods agree well.

Additional file 5: Figure S3. Distribution of equivalent PCCs. The *p* values obtained by BigSur for the melanoma cell line were transformed using the inverse of the Fisher formula to a set of "equivalent" PCCs. Since the Fisher formula operates on the absolute values of correlations, each calculated equivalent PCC was assigned the sign of the PCC′ value for the same gene pair. Equivalent PCCs provide a measure of correlation strength that can be compared across data sets with differing numbers of cells. They may be understood as a measure of how strongly correlated two normally distributed vectorswould need to be to produce the observed *p* value. Here, only those gene pairs judged significant by BigSur are shown. The fact that so many weakly correlated gene pairsare nevertheless statistically significant is a function of the long vector length in this experiment.

Additional file 6: Figure S4. Gene communities A and B from cell cluster 1.2. Green edges depict significant correlations. Transcription factor vertices are displayed as yellow boxes with gene names in blue. In the boxed insets the same graphs are overlayed in brown to highlight links supported by known protein-protein interactions.

Additional file 7: Figure S5. Gene communities C, and D from cell cluster 1.2. Genes and links are highlighted as in Fig. S4.

Additional file 8: Figure S6. Gene communities E, F, G and H from cell cluster 1.2. Genes and links are highlighted as in Fig. S4.

Additional file 9: Figure S7. Gene communities I, K, L and M from cell cluster 1.2. Genes and links are highlighted as in Fig. S4. In communities L and M, links supported by known protein-protein interactions are highlighted in brown.

Additional file 10: Figure S8. Graphical representation of positive correlations among genes significantly upregulated in lung epithelial cells treated with IL13 [76]. Of 419 upregulated genes, 313 form a single connected community in the treated cells, whereas 45 of those correlate in the untreated group. Green lines represent statistically significant positive correlations. Transcription factors are marked with blue text and a yellow box.

Additional file 11: Figure S9. Gene-gene correlation among genes that positively correlate with MUC16 in IL13-treated secretory lung epithelial cells. Genes were clustered using complete-linkage hierarchical clustering. Darker color indicates stronger correlation; white indicates no significant correlation

Acknowledgements
We thank Weining Shen (UCI) for advice on statistical analysis. We acknowledge Mika Caldwell and Yilun Zhu for helpful feedback and testing of code.

Author contributions
KS conceived of the work, developed and implemented code, was responsible for posting code and data to github, and contributed to writing of the manuscript. ED conceived of the work and contributed to writing of the manuscript. JG helped conceive of the study and contributed to the development of code. SA provided financial support and edited the manuscript. QN provided financial support, advice on code development and edited the manuscript. AL conceived of the work, developed and implemented code, and contributed to writing of the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by NIH Grants CA217378, AR075047, DE019638 and the NSF-Simons Center for Multiscale Cell Fate Research (NSF 1763272). K.S. and E.D. acknowledge support from NIH training Grant GM136624.

Availability of data and materials
The datasets generated and code used for analysis during the current study are available in the BigSurM Github repository, https://github.com/landerlabcode/BigSurM/tree/0c128246bed1969b371e12bdbd094e49436e2e34/Datasets.

Code availability
R and Mathematica implementations are available under the Lander lab profile on GitHub (https://github.com/landerlabcode/).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests

Received: 15 November 2023 Accepted: 9 September 2024
Published online: 18 September 2024

References

1. Tritschler S, Buttner M, Fischer DS, Lange M, Bergen V, Lickert H, Theis FJ. Concepts and limitations for learning developmental trajectories from single cell genomics. Development. 2019;146:dev170506.

2. Tam PPL, Ho JWK. Cellular diversity and lineage trajectory: insights from mouse single cell transcriptomes. Development. 2020;147:dev179788.
3. Nguyen H, Tran D, Tran B, Pehlivan B, Nguyen T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Brief Bioinform. 2021;22:bbaa190.
4. Xie B, Jiang Q, Mora A, Li X. Automatic cell type identification methods for single-cell RNA sequencing. Comput Struct Biotechnol J. 2021;19:5874–87.
5. Junttila S, Smolander J, Elo LL. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. Brief Bioinform. 2022;23:bbac286.
6. Das S, Rai A, Rai SN. Differential expression analysis of single-cell RNA-Seq data: current statistical approaches and outstanding challenges. Entropy (Basel). 2022;24:995.
7. Simmons S. Cell type composition analysis: comparison of statistical methods. bioRxiv 2022**:**2022.2002.2004.479123.
8. Wang H, Ma X. Learning discriminative and structural samples for rare cell types with deep generative model. Brief Bioinform. 2022;23:bbac317.
9. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525:251–5.
10. Jiang L, Chen H, Pinello L, Yuan GC. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016;17:144.
11. Herman JS, Sagar N, Grun D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. Nat Methods. 2018;15:379–86.
12. Jindal A, Gupta P, Jayadeva, Sengupta D. Discovery of rare cells from voluminous single cell expression data. Nat Commun. 2018;9:4719.
13. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. Characterization of cell fate probabilities in single-cell data with Palantir. Nat Biotechnol. 2019;37:451–60.
14. Wegmann R, Neri M, Schuierer S, Bilican B, Hartkopf H, Nigsch F, Mapa F, Waldt A, Cuttat R, Salick MR, et al. CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. Genome Biol. 2019;20:142.
15. Dong R, Yuan GC. GiniClust3: a fast and memory-efficient tool for rare cell type identification. BMC Bioinform. 2020;21:158.
16. Gerniers A, Bricard O, Dupont P. MicroCellClust: mining rare and highly specific subpopulations from single-cell expression data. Bioinformatics. 2021;37:3220–7.
17. Bej S, Galow AM, David R, Wolfien M, Wolkenhauer O. Automated annotation of rare-cell types from single-cell RNA-sequencing data through synthetic oversampling. BMC Bioinform. 2021;22:557.
18. Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, Lickert H, Ansari M, Schniering J, Schiller HB, et al. Cell Rank for directed single-cell fate mapping. Nat Methods. 2022;19:159–70.
19. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, Myung P, Plikus MV, Nie Q. Inference and analysis of cell-cell communication using Cell Chat. Nat Commun. 2021;12:1088.
20. Zhang L, Nie Q. scMC learns biological variation through the alignment of multiple single-cell genomics datasets. Genome Biol. 2021;22:10.
21. Sha Y, Wang S, Bocci F, Zhou P, Nie Q. Inference of intercellular communications and multilayer gene-regulations of epithelial-mesenchymal transition from single-cell transcriptomic data. Front Genet. 2020;11:604585.
22. Bocci F, Zhou P, Nie Q. spliceJAC: transition genes and state-specific gene regulation from single-cell transcriptome data. Mol Syst Biol. 2022;18:e11176.
23. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol. 2022;40:245–53.
24. Bageritz J, Willnow P, Valentini E, Leible S, Boutros M, Teleman AA. Gene expression atlas of a developing tissue by single cell expression correlation analysis. Nat Methods. 2019;16:750–6.
25. Kolodziejczyk AA, Kim JK, Tsang JC, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Buhler M, Liu P, et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell. 2015;17:471–85.
26. Pedraza JM, van Oudenaarden A. Noise propagation in gene networks. Science. 2005;307:1965–9.
27. Becskei A, Kaufmann BB, van Oudenaarden A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. Nat Genet. 2005;37:937–44.
28. Stewart-Ornstein J, Weissman JS, El-Samad H. Cellular noise regulons underlie fluctuations in *Saccharomyces cerevisiae*. Mol Cell. 2012;45:483–93.
29. Padovan-Merhar O, Raj A. Using variability in gene expression as a tool for studying gene regulation. Wiley Interdiscip Rev Syst Biol Med. 2013;5:751–9.
30. Munsky B, Neuert G, van Oudenaarden A. Using gene expression noise to understand gene regulation. Science. 2012;336:183–7.
31. Warmflash A, Dinner AR. Signatures of combinatorial regulation in intrinsic biological noise. Proc Natl Acad Sci U S A. 2008;105:17262–7.
32. Gupta A, Martin-Rufino JD, Jones TR, Subramanian V, Qiu X, Grody EI, Bloemendal A, Weng C, Niu SY, Min KH, et al. Inferring gene regulation from stochastic transcriptional variation across single cells at steady state. Proc Natl Acad Sci U S A. 2022;119:e2207392119.
33. He Z, Pan Y, Shao F, Wang H. Identifying differentially expressed genes of zero inflated single cell RNA sequencing data using mixed model score tests. Front Genet. 2021;12:616686.
34. Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. 2022;23:27.
35. Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nat Genet. 2021;53:770–7.
36. Choi K, Chen Y, Skelly DA, Churchill GA. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. Genome Biol. 2020;21:183.

37. Wang J, Huang M, Torre E, Dueck H, Shaffer S, Murray J, Raj A, Li M, Zhang NR. Gene expression distribution deconvolution in single-cell RNA sequencing. Proc Natl Acad Sci U S A. 2018;115:E6437–46.
38. Kim JK, Kolodziejczyk AA, Ilicic T, Teichmann SA, Marioni JC. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat Commun. 2015;6:8687.
39. DiCiccio CJ, Romano JP. Robust permutation tests for correlation and regression coefficients. J Am Stat Assoc. 2017;112:1211–20.
40. Jin S, MacLean AL, Peng T, Nie Q. scEpath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. Bioinformatics. 2018;34:2077–86.
41. Yang XH, Goldstein A, Sun Y, Wang Z, Wei M, Moskowitz IP, Cunningham JM. Detecting critical transition signals from single-cell transcriptomes to infer lineage-determining transcription factors. Nucleic Acids Res. 2022;50:e91.
42. Fisher RA. Statistical methods for research workers. 11th ed. Edinburgh: Oliver and Boyd; 1950.
43. Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. 2021;22:258.
44. Bahar Halpern K, Tanami S, Landen S, Chapal M, Szlak L, Hutzler A, Nizhberg A, Itzkovitz S. Bursty gene expression in the intact mammalian liver. Mol Cell. 2015;58:147–56.
45. Thattai M, van Oudenaarden A. Intrinsic noise in gene regulatory networks. Proc Natl Acad Sci U S A. 2001;98:8614–9.
46. Schwabe A, Rybakova KN, Bruggeman FJ. Transcription stochasticity of complex gene regulation models. Biophys J. 2012;103:1152–61.
47. Beal J. Biochemical complexity drives log-normal variation in genetic expression. Eng Biol. 2017;1:55–60.
48. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296.
49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57:289–300.
50. Torre E, Dueck H, Shaffer S, Gospocic J, Gupte R, Bonasio R, Kim J, Murray J, Raj A. Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. Cell Syst. 2018;6(171–179):e175.
51. Ibn-Salem J, Muro EM, Andrade-Navarro MA. Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res. 2017;45:81–91.
52. Pons P, Latapy M. Computing communities in large networks using random walks. In: Computer and information sciences—ISCIS 2005, vol. 3733. Berlin: Springer; 2005, pp. 284–293.
53. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17:29.
54. Langemeijer SM, Mariani N, Knops R, Gilissen C, Woestenenk R, de Witte T, Huls G, van der Reijden BA, Jansen JH. Apoptosis-related gene expression profiling in hematopoietic cell fractions of MDS patients. PLoS ONE. 2016;11:e0165582.
55. Tyler SR, Lozano-Ojalvo D, Guccione E, et al. Anti-correlated feature selection prevents false discovery of subpopulations in scRNAseq. Nat Commun. 2024;15:699. https://doi.org/10.1038/s41467-023-43406-9.
56. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34:267–73.
57. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
58. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.
59. Sykes EK, Mactier S, Christopherson RI. Melanoma and the unfolded protein response. Cancers (Basel). 2016;8:30.
60. Rather RA, Bhagat M, Singh SK. Oncogenic BRAF, endoplasmic reticulum stress, and autophagy: crosstalk and therapeutic targets in cutaneous melanoma. Mutat Res Rev Mutat Res. 2020;785:108321.
61. Manga P, Choudhury N. The unfolded protein and integrated stress response in melanoma and vitiligo. Pigment Cell Melanoma Res. 2021;34:204–11.
62. Netanely D, Leibou S, Parikh R, Stern N, Vaknine H, Brenner R, Amar S, Factor RH, Perluk T, Frand J, et al. Classification of node-positive melanomas into prognostic subgroups using keratin, immune, and melanogenesis expression patterns. Oncogene. 2021;40:1792–805.
63. Rambow F, Job B, Petit V, Gesbert F, Delmas V, Seberg H, Meurice G, Van Otterloo E, Dessen P, Robert C, et al. New functional signatures for understanding melanoma biology from tumor cell lineage-specific analysis. Cell Rep. 2015;13:840–53.
64. Wouters J, Kalender-Atak Z, Minnoye L, Spanier KI, De Waegeneer M, Bravo Gonzalez-Blas C, Mauduit D, Davie K, Hulselmans G, Najem A, et al. Robust gene expression programs underlie recurrent cell states and phenotype switching in melanoma. Nat Cell Biol. 2020;22:986–98.
65. Eriksson J, Le Joncour V, Nummela P, Jahkola T, Virolainen S, Laakkonen P, Saksela O, Holtta E. Gene expression analyses of primary melanomas reveal CTHRC1 as an important player in melanoma progression. Oncotarget. 2016;7:15065–92.
66. Tsoi J, Robert L, Paraiso K, Galvan C, Sheu KM, Lay J, Wong DJL, Atefi M, Shirazi R, Wang X, et al. Multi-stage differentiation defines melanoma subtypes with differential vulnerability to drug-induced iron-dependent oxidative stress. Cancer Cell. 2018;33(890–904): e895.
67. Verfaillie A, Imrichova H, Atak ZK, Dewaele M, Rambow F, Hulselmans G, Christiaens V, Svetlichnyy D, Luciani F, Van den Mooter L, et al. Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. Nat Commun. 2015;6:6683.
68. Wang J, Saraswat D, Sinha AK, Polanco J, Dietz K, O'Bara MA, Pol SU, Shayya HJ, Sim FJ. Paired related homeobox protein 1 regulates quiescence in human oligodendrocyte progenitors. Cell Rep. 2018;25(3435–3450):e3436.

69. Blattmann P, Henriques D, Zimmermann M, Frommelt F, Sauer U, Saez-Rodriguez J, Aebersold R. Systems pharmacology dissection of cholesterol regulation reveals determinants of large pharmacodynamic variability between cell lines. Cell Syst. 2017;5(604–619): e607.
70. Capell-Hattam IM, Fenton NM, Coates HW, Sharpe LJ, Brown AJ. The non catalytic protein ERG28 has a functional role in cholesterol synthesis and is coregulated transcriptionally. J Lipid Res. 2022;63:100295.
71. Gong Y, Lee JN, Brown MS, Goldstein JL, Ye J. Juxtamembranous aspartic acid in Insig-1 and Insig-2 is required for cholesterol homeostasis. Proc Natl Acad Sci U S A. 2006;103:6154–9.
72. Jeong SJ, Kim S, Park JG, Jung IH, Lee MN, Jeon S, Kweon HY, Yu DY, Lee SH, Jang Y, et al. Prdx1 (peroxiredoxin 1) deficiency reduces cholesterol efflux via impaired macrophage lipophagic flux. Autophagy. 2018;14:120–33.
73. Schallreuter KU, Hasse S, Rokos H, Chavan B, Shalbaf M, Spencer JD, Wood JM. Cholesterol regulates melanogenesis in human epidermal melanocytes and melanoma cells. Exp Dermatol. 2009;18:680–8.
74. Nikolakaki E, Simos G, Georgatos SD, Giannakouros T. A nuclear envelope-associated kinase phosphorylates arginine-serine motifs and modulates interactions between the lamin B receptor and other nuclear proteins. J Biol Chem. 1996;271:8365–72.
75. Singh P, Saxena R, Srinivas G, Pande G, Chattopadhyay A. Cholesterol biosynthesis and homeostasis in regulation of the cell cycle. PLoS ONE. 2013;8:e58833.
76. Jackson ND, Everman JL, Chioccioli M, Feriani L, Goldfarbmuren KC, Sajuthi SP, Rios CL, Powell R, Armstrong M, Gomez J, et al. Single-cell and population transcriptomics reveal pan-epithelial remodeling in type 2-high asthma. Cell Rep. 2020;32:107872.
77. Tuvim MJ, Mospan AR, Burns KA, Chua M, Mohler PJ, Melicoff E, Adachi R, Ammar-Aouchiche Z, Davis CW, Dickey BF. Synaptotagmin 2 couples mucin granule exocytosis to Ca^{2+} signaling from endoplasmic reticulum. J Biol Chem. 2009;284:9781–7.
78. Ding L, Abebe T, Beyene J, Wilke RA, Goldberg A, Woo JG, Martin LJ, Rothenberg ME, Rao M, Hershey GK, et al. Rank-based genome-wide analysis reveals the association of ryanodine receptor-2 gene variants with childhood asthma among human populations. Hum Genomics. 2013;7:16.
79. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, Meir Z, Hoichman M, Lifshitz A, Tanay A. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. 2019;20:206.
80. Ben-Kiki O, Bercovich A, Lifshitz A, Tanay A. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. Genome Biol. 2022;23:100.
81. Persad S, Choo ZN, Dien C, Sohail N, Masilionis I, Chaligne R, Nawy T, Brown CC, Sharma R, Peer I, et al. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nat Biotechnol. 2023;41:1746–57.
82. Lun ATL, Marioni JC. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. Biostatistics. 2017;18:451–64.
83. Bilous M, Tran L, Cianciaruso C, Gabriel A, Michel H, Carmona SJ, Pittet MJ, Gfeller D. Metacells untangle large and complex single-cell transcriptome networks. BMC Bioinform. 2022;23:336.
84. Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. Cell Rep Methods. 2023;3:100498.
85. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, Hudelle R, Qaiser T, Matson KJE, Barraud Q, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021;12:5692.
86. Xu H, Hu Y, Zhang X, Aouizerat BE, Yan C, Xu K. A novel graph-based k-partitioning approach improves the detection of gene-gene correlations by single-cell RNA sequencing. BMC Genomics. 2022;23:35.
87. Eisen TJ, Eichhorn SW, Subtelny AO, Lin KS, McGeary SE, Gupta S, Bartel DP. The dynamics of cytoplasmic mRNA metabolism. Mol Cell. 2020;77(786–799):e710.
88. Davis AF, Clayton DA. In situ localization of mitochondrial DNA replication in intact mammalian cells. J Cell Biol. 1996;135:883–93.
89. Yao Z, Luo J, Hu K, Lin J, Huang H, Wang Q, Zhang P, Xiong Z, He C, Huang Z, et al. ZKSCAN1 gene and its related circular RNA (circZKSCAN1) both inhibit hepatocellular carcinoma cell growth, migration, and invasion but through different signaling pathways. Mol Oncol. 2017;11:422–37.
90. Dollinger E, Silkwood K, Atwood S, Nie Q. A principled, robust approach to feature selection in single cell transcriptomics. bioRxiv 2023;to be submitted.
91. Cornish EA, Fisher RA. Moments and cumulants in the specification of distributions. Rev Int Stat Inst. 1938;5:307–20.
92. Hu Y, Ewen-Campen B, Comjean A, Rodiger J, Mohr SE, Perrimon N. Paralog Explorer: a resource for mining information about paralogs in common research organisms. Comput Struct Biotechnol J. 2022;20:6570–7.
93. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. Eur Phys J B. 2013;86:471.
94. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.

Publisher's Note