Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles

Sameer S. Bajikar[a,1], Christiane Fuchs[b,c,1], Andreas Roller[b,c], Fabian J. Theis[b,c,2], and Kevin A. Janes[a,2]

[a]Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22908; [b]Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health, 85764 Neuherberg, Germany; and [c]Institute for Mathematical Sciences, Technical University Munich, 85747 Garching, Germany

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved December 27, 2013 (received for review June 18, 2013)

Regulated changes in gene expression underlie many biological processes, but globally profiling cell-to-cell variations in transcriptional regulation is problematic when measuring single cells. Transcriptome-wide identification of regulatory heterogeneities can be robustly achieved by randomly collecting small numbers of cells followed by statistical analysis. However, this stochastic-profiling approach blurs out the expression states of the individual cells in each pooled sample. Here, we show that the underlying distribution of single-cell regulatory states can be deconvolved from stochastic-profiling data through maximum-likelihood inference. Guided by the mechanisms of transcriptional regulation, we formulated plausible mixture models for cell-to-cell regulatory heterogeneity and maximized the resulting likelihood functions to infer model parameters. Inferences were validated both computationally and experimentally for different mixture models, which included regulatory states for multicellular function that were occupied by as few as 1 in 40 cells of the population. Importantly, when the method was extended to programs of heterogeneously coexpressed transcripts, we found that population-level inferences were much more accurate with pooled samples than with one-cell samples when the extent of sampling was limited. Our deconvolution method provides a means to quantify the heterogeneous regulation of molecular states efficiently and gain a deeper understanding of the heterogeneous execution of cell decisions.

noise | morphogenesis | breast cancer | systems biology

Cell-to-cell differences in transcriptional or posttranslational regulation can give rise to heterogeneous phenotypes within a population (1–7). There are several elegant techniques for monitoring regulatory states in single cells after a network of marker and effector genes has been identified (8–13). However, the options are much more limited when seeking to discover novel states without a predefined network. At the transcript level, global methods have been developed to profile single cells by oligonucleotide microarrays (14, 15) or RNA sequencing (16–19). However, generally such approaches overlook the considerable technical variation in RNA extraction (20) and reverse transcription (21) when applied to the limited starting material of single cells. Single-cell profiles also retain the biological noisiness (22) associated with each cell's isolation and handling. These confounding sources of variation cannot be separated from reproducible heterogeneities in regulation unless many (>50) cells are individually profiled (9). Therefore, challenges remain for single-cell methods to discover regulatory heterogeneities in a reliable, unbiased, and efficient way.

An attractive alternative to single-cell methods is to analyze sets of population-averaged data and define regulatory signatures for discrete subpopulations. Existing approaches for transcriptomic data are able to deconvolve mixed cellular states computationally, but they require hundreds of coexpressed markers (23) or calibration with purified cell populations (24, 25). Usually, the size or identity of regulatory states is not defined beforehand and their discovery is what motivates the study (9, 11, 26). Certain states may also lack well-defined surface markers that would allow purification. It thus remains unclear whether computational inference with multiple cell averages can track quantitative characteristics of regulatory states not previously thought to exist.

As a hybrid between single-cell and mixture-based approaches, we previously developed a technique that applies probability theory to transcriptome-wide measurements (27). The method begins with random collections of up to 10 cells isolated in situ where cell-to-cell regulatory heterogeneities could possibly reside. Each of these "stochastic samples" is then profiled for overall mRNA expression by using a heavily customized cDNA amplification procedure together with oligonucleotide microarrays (20, 27). The process of random sampling is repeated 15–20 times to build a distribution of 10-cell averages. Transcripts with stark cell-to-cell variations can be distinguished statistically because of binomial fluctuations in single-cell expression that convolve their 10-cell averages. Last, candidate heterogeneities are clustered on a gene-by-gene basis according to the patterns of their sampling fluctuations to indicate putative regulatory states in single cells (27).

Stochastic-profiling experiments are quantitative and highly reproducible as a result of the 10-fold increase in starting material compared with a single cell (20). However, a recognized drawback of the approach is that explicit information about single cells is "lost" in the 10-cell averages. Here, we report that one can recover this information computationally and reconstruct the single-cell distribution of regulatory states with remarkable accuracy. Our method combines maximum-likelihood estimation with mixture models that are grounded in known mechanisms of transcriptional regulation. This approach of maximum-likelihood inference quantifies the single-cell characteristics of each regulatory state, including the probability that a cell will reside in one state or the

Significance

Cell-to-cell variations in gene regulation occur in a number of biological contexts, such as development and cancer. Discovering regulatory heterogeneities in an unbiased manner is difficult owing to the population averaging that is required for most global molecular methods. Here, we show that we can infer single-cell regulatory states by mathematically deconvolving global measurements taken as averages from small groups of cells. This averaging-and-deconvolution approach allows us to quantify single-cell regulatory heterogeneities while avoiding the measurement noise of global single-cell techniques. Our method is particularly relevant to solid tissues, where single-cell dissociation and molecular profiling is especially problematic.

Author contributions: S.S.B., C.F., A.R., F.J.T., and K.A.J. designed research; S.S.B., C.F., and A.R. performed research; S.S.B. and C.F. contributed new reagents/analytic tools; S.S.B., C.F., F.J.T., and K.A.J. analyzed data; and S.S.B., C.F., and K.A.J. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]S.S.B. and C.F. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: kjanes@virginia.edu or fabian.theis@helmholtz-muenchen.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1311647111/-/DCSupplemental.

other. Our predictions are validated with independent gene-specific observations in single cells, and we demonstrate for one very rare state (~2–3% of the population) that it is important for normal morphogenesis of breast epithelial cells in 3D culture. Last, we show that, when sampling is limited to fewer than 20 observations, the parameterization of regulatory states is substantially more accurate when given 10-cell data compared with one-cell data. Maximum-likelihood inference now enables stochastic profiling to bridge the gap between -omics datasets and single-cell information.

Results

Probability Models for Heterogeneous Transcriptional Regulation. To make reliable single-cell inferences, it was critical to start with simple probabilistic models of gene expression that were biologically accurate. Our method considers genes that exhibit two distinct regulatory states in a population of cells (19, 20, 27). Within each state, the cell-to-cell variation of expression was originally described by a lognormal distribution according to measurements of high-copy transcripts in single mammalian cells (28, 29). We tested whether there was a mechanistic foundation for using two lognormal subpopulations by examining a standard model of regulated gene expression (30, 31) (*SI Appendix*, Fig. S1). In this model, transcript levels per cell are determined by the kinetics of polymerase binding–unbinding, transcriptional elongation, and mRNA degradation. The relative magnitudes of the kinetic rate parameters together govern the steady-state distribution of transcripts in the population (32), allowing different regulatory states to be simulated.

For parameter sets where the probability of observing zero transcripts per cell was near zero, we found that the lognormal distribution was a suitable approximation of basal expression (Fig. 1*A*, blue). Parameter sets yielding median expression levels as low as 20 copies per cell showed only minor skewness in quantile–quantile comparisons with a lognormal distribution (Fig. 1*A*, blue inset). Starting with this basal distribution, we simulated a second cellular regulatory state by increasing the rate of polymerase binding, decreasing the rate of mRNA degradation, or both (Fig. 1*A*, orange). The apparent rate of polymerase binding increases upon recruitment by transcription factors that are expressed or activated heterogeneously within a population of cells (4, 5). Conversely, mRNA stabilization occurs post-transcriptionally through dedicated signal-transduction pathways activated by environmental stresses and proinflammatory stimuli (33). We found that either mechanism of gene up-regulation led to right-shifted distributions that were lognormal (Fig. 1*A*, orange inset). These simulations indicated that lognormal random variables were appropriate for the regulated expression of mid- to high-abundance transcripts.

One drawback of the lognormal distribution is that it has no support at zero copies (34), making it poor for capturing low-abundance genes that are completely silenced in some cells. To identify an alternative in this circumstance, we reconfigured the parameters of the model and defined a steady-state population where most cells would contain close to zero transcripts (Fig. 1*B*, blue). As noted before (32), this regulatory state was best captured by an exponential distribution (Fig. 1*B*, blue inset). Importantly, we found that when the kinetic parameters of a basal exponential state were modified to create a second right-shifted state (Fig. 1*B*, orange), the resulting distributions were lognormal (Fig. 1*B*, orange inset). Together, we conclude that the basic mechanisms of gene expression lead to steady-state distributions described by probability models that are relatively simple.

Deconvolution of Random 10-Cell Averages by Maximum-Likelihood Inference. Our results from the gene-expression model suggested that single-cell regulatory heterogeneities could be depicted as a mixture of two lognormal states or as a mixture of an exponential state and a lognormal state (Fig. 1). Either mixture gives rise to a probability distribution characterized by four key parameters. The lognormal–lognormal (LN–LN) mixture requires the log-mean expression of the two regulatory states (μ_1 and μ_2), the log-SD for biological noise (σ), and the expression frequency (F) describing the probability that cells will occupy the higher regulatory state (step 1, Fig. 2*A*). (For simulations, the two lognormal states are assumed to share a common σ, but in practice we test whether

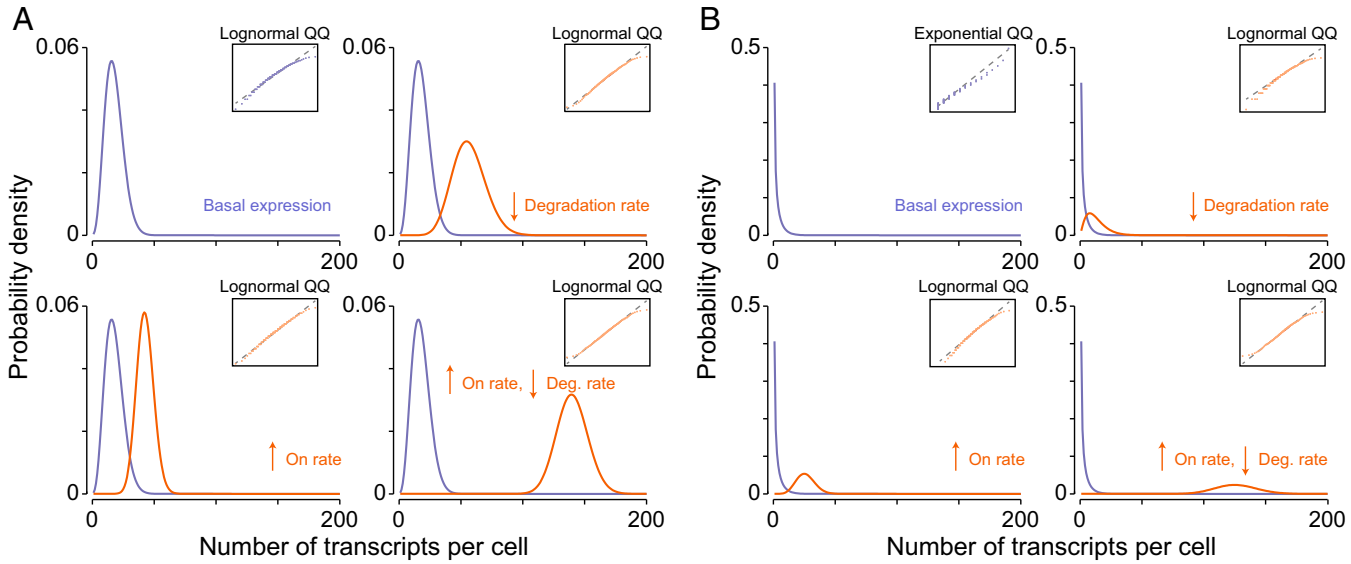

Fig. 1. Simple probability models capture regulated changes in gene expression. (*A* and *B*) Probability densities for the number of transcripts per cell were calculated using a kinetic model (30, 31) whose parameters led to basal regulatory states (blue) with either nonzero copies per cell in *A* or with near-zero copies per cell in *B*. The basal-regulatory states were compared with a lognormal distribution in *A* or an exponential distribution in *B* through a quantile–quantile (QQ) plot (blue insets). A second, induced regulatory state (orange) was created by increasing the polymerase binding rate (*Lower Left*), decreasing the transcript degradation rate (*Upper Right*), or both (*Lower Right*) in the model (*SI Appendix*, Fig. S1). All induced regulatory states were compared with a lognormal distribution through a QQ plot (orange insets).

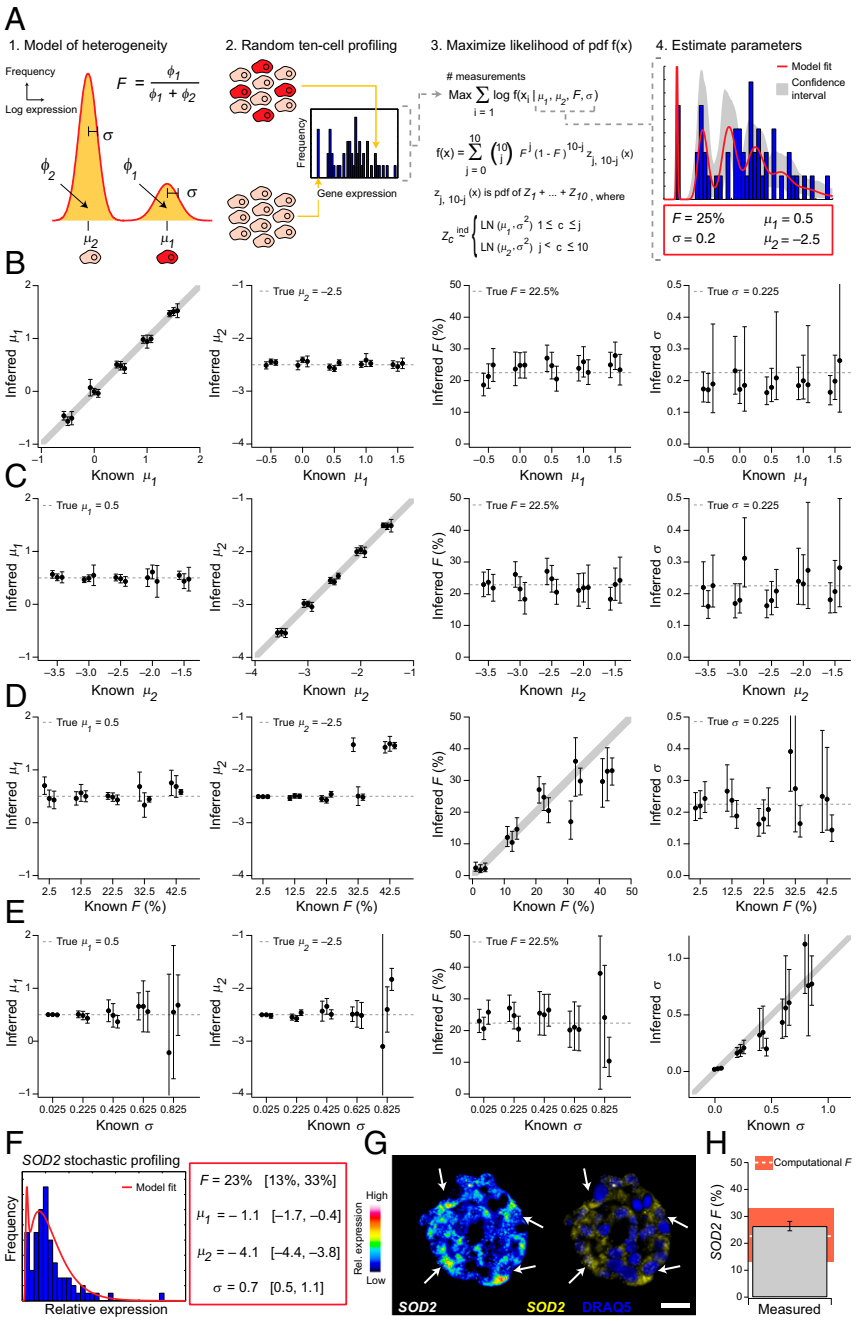

Fig. 2. Inferring cellular subpopulations by maximum-likelihood inference of stochastic 10-cell samples from an LN–LN mixture of regulatory states. (*A*) The maximum-likelihood approach involves four steps. (1) A model of heterogeneous gene regulation is posed, where single cells are assumed to express genes at a low or high level with a common coefficient of variation for both subpopulations. The weight of each subpopulation is defined by the integrated single-cell expression distribution of the subpopulation (ϕ_1 and ϕ_2). The four parameters of the model are the log-mean expression for each subpopulation (μ_1 and μ_2), the proportion of cells in the high subpopulation (F), and the common log-SD of expression (σ). (2) Random 10-cell samples are collected to build a distribution of measurements for inference by the model. (3) Based on the model in step 1, a likelihood function is derived (*Methods*). (4) The likelihood function is then maximized by searching through the four parameters of the model to identify those that are most likely given the experimental observations. Additionally, we obtain measures of confidence for each estimated parameter (gray). (*B–E*) Accurate prediction of single-cell parameters from simulated 10-cell samples. Ten-cell expression data were simulated using different values of (*B*) μ_1, (*C*) μ_2, (*D*) F, and (*E*) σ (solid gray line) and then estimated by maximum likelihood. For each group of simulations, the remaining three model parameters were kept fixed at (*C–E*) $\mu_1 = 0.5$, (*B*, *D*, and *E*) $\mu_2 = -2.5$, (*B–C* and *E*) $F = 22.5\%$, and (*B–D*) $\sigma = 0.225$ (dashed gray line). Solid gray line shows the one-to-one mapping of inferred-to-known parameter value. Off-diagonal plots are categorical plots of the fixed parameter estimates for a given value of the perturbed parameter. Graphs show the parameter estimates together with 95% maximum-likelihood CIs from three independent sets of 50 10-cell samples. (*F–H*) Prediction and validation of expression frequency for the heterogeneous transcript, *SOD2*, during breast-epithelial acinar morphogenesis. (*F*) Distribution of 81 10-cell qPCR measurements of *SOD2* in outer ECM-attached epithelial cells and estimated subpopulation distribution (red line). Maximum-likelihood parameters (red box) are shown with 95% CI in brackets. (*G*) Representative RNA FISH image of endogenous *SOD2* expression. A pseudocolored image (*Left*) is shown alongside a two-color image with DRAQ5 counterstain to visualize nuclei (*Right*). Arrows indicate ECM-attached cells with high *SOD2* expression. (Scale bar, 20 μm.) (*H*) Percentage of cells showing high expression of *SOD2* by RNA FISH (gray bar) compared with the maximum-likelihood estimate of F (white dashed line). RNA FISH data are shown as the mean percentage with 95% CI of ECM-attached cells showing high expression of *SOD2*. Maximum-likelihood predictions are shown as the parameter point estimate (white) with 95% CI (red).

inferences are improved when each lognormal state is allowed its own noise parameter; see below.) Thus, an LN–LN gene that is expressed at an approximately eightfold higher level in 20% of the population with a coefficient of variation (CV) of ∼50% is captured by $\mu_1 - \mu_2 = 2$, $F = 20\%$, and $\sigma = 0.48$.

The exponential–lognormal (EXP–LN) mixture also requires σ and F, along with a single log-mean for the high lognormal state (μ) and a rate parameter for the low exponential state (λ) (step 1, *SI Appendix*, Fig. S2). The rate parameter relates to how quickly the lower distribution decays above zero copies per cell. For example, a rate parameter of $\lambda = 1$ creates a distribution that has ∼37% overlap with that of a high lognormal state of $\mu = 0.5$ and $\sigma = 0.225$ whereas $\lambda = 3$ causes only a ∼6.3% overlap. We modeled two distinct regulatory states by restricting the simulations to rate parameters that caused negligible overlap with the high lognormal state ($\lambda > 3$). Together, the different mixture models enabled us to simulate stochastic-profiling data by summing the expression of 10 cells randomly sampled from the appropriate two-state distribution (step 2, Fig. 2*A* and *SI Appendix*, Fig. S2).

To infer the most-likely parameters from a collection of random 10-cell samples, we derived maximum-likelihood estimators for the LN–LN and EXP–LN mixtures (*Methods* and *SI Appendix*, *Methods*). Maximum-likelihood estimation requires a defined probability density function (pdf). The stochastic-sampling pdf is the convolution of 10 binomial choices drawn from the two underlying distributions in the mixture (step 3, Fig. 2*A* and *SI Appendix*, Fig. S2). The pdf has a ≤11-modal shape where each mode corresponds to choosing 0–10 cells from the high regulatory state. The most-likely parameter combination was calculated by maximizing the likelihood function (*Methods*), yielding parameters

with interval estimates that best explained the data (step 4, Fig. 2*A* and *SI Appendix*, Fig. S2). By performing this maximum-likelihood estimation, we could "invert" stochastic profiling data to infer single-cell characteristics from 10-cell samples.

Theoretical and Experimental Validation of Maximum-Likelihood Inference. We evaluated the performance of our approach by using computational simulations of 10-cell samples with known distribution parameters. First, it was important to identify the minimum number of random samples needed to ensure accurate parameter estimation. Given hundreds to thousands of samples, we found that robust and accurate estimates were obtained for all model parameters irrespective of the mixture type (*SI Appendix*, Fig. S3 *A* and *B*). Conversely, with very few samples (~20 or fewer), the convolved distributions were incompletely populated and our resulting estimates were highly uncertain and sometimes inaccurate for the LN–LN and EXP–LN mixtures. The transition between the two regimes occurred at 50–100 samples, which we defined as the approximate number of data points required for effective maximum-likelihood inference of single transcripts.

We next used simulations to identify the parameter ranges where maximum-likelihood inference makes accurate estimates of each regulatory state. Starting with the LN–LN mixture, we perturbed μ_1, μ_2, σ, or F individually while keeping the other three parameters fixed and simulated 50 random 10-cell samples. For a wide range of subpopulation log-means (μ_1 and μ_2), maximum-likelihood inference accurately inferred model parameters with negligible bias (Fig. 2 *B* and *C*). We also observed good performance when altering the expression frequency (F). Accuracy declined near $F = 50\%$, when the two subpopulations offset one another and disguise as a distribution with large σ (Fig. 2*D*). Nevertheless, the estimation procedure still accurately and confidently captured ~70% of the total parameter space ($F = 0$–35% over the range of 0–50%). For the log-SD (σ), performance declined only when this parameter was extremely high (Fig. 2*E*). Parameter estimates were accurate until σ reached ~0.8, corresponding to a ~95% CV that is higher than nearly all genes examined thus far (35, 36). None of the mixture parameters could be reliably inferred from higher-order moments of the 10-cell distributions, although low F or high σ correlated with a slight increase in skewness (*SI Appendix*, Fig. S4). These results indicated that maximum-likelihood inference could extract parameters that were otherwise inaccessible by descriptive statistics.

We repeated the simulations for the EXP–LN mixture and arrived at very similar conclusions. As long as λ and μ were large enough to prevent overlap of the two regulatory states, we found that parameter estimates were accurate, although the variance of inferred σ was somewhat higher than in the LN–LN mixture (*SI Appendix*, Fig. S5). Together, these simulations suggested that maximum-likelihood inference is able to deconvolve a wide range of regulatory heterogeneities from 10-cell samples.

To examine the accuracy of maximum-likelihood inference with real 10-cell samples, we focused on expression of the detoxifying enzyme superoxide dismutase 2 (*SOD2*) during breast-epithelial acinar morphogenesis. We used a culture model in which immortalized human breast epithelial cells are grown as single-cell clones in reconstituted basement-membrane ECM to form 3D organotypic spheroids (37, 38). Earlier stochastic-profiling studies of developing spheroids had suggested that there were two *SOD2* regulatory states among the ECM-attached cells (27, 39). To apply maximum-likelihood inference, we deeply sampled *SOD2* expression by quantitative PCR (qPCR) in 81 random samples of 10 ECM-attached cells (Fig. 2*F*, *Left*). Using these data, we maximized the likelihood of the LN–LN and EXP–LN models, as well as that of a relaxed LN–LN model, which allowed each regulatory state to have its own log-SD (σ_1 and σ_2). The three estimates were compared by using the Bayesian information criterion (BIC) score to calculate the quality of the fit relative to the number of inferred parameters (*SI Appendix*, Table S1). The best overall estimate was the mixture model that parameterized two distinct regulatory states with the lowest BIC score.

For the 10-cell measurements of *SOD2*, we found that the LN–LN mixture was slightly preferred over the EXP–LN mixture (Fig. 2*F*, *Right* and *SI Appendix*, Table S1). The inability to discriminate clearly between these two models was likely caused by the basal regulatory state, which could be described as an exponential distribution ($\lambda = 46$) or a lognormal distribution with a very small log-mean ($\mu_2 = -4.1$) given the sampling data. Regardless, the two models predicted similar *SOD2* expression frequencies among ECM-attached cells: 23% (13–33%) for the LN–LN mixture vs. 19% (12–27%) for the EXP–LN mixture. To determine the accuracy of this shared prediction, we directly measured F in 3D spheroids by RNA FISH (Fig. 2*G*). Scoring individual cells with high *SOD2* fluorescence intensity, we calculated an expression frequency of ~26%. This measurement closely agreed with the inferred parameter of the LN–LN mixture (the better-scoring model; Fig. 2*H* and *SI Appendix*, Table S1) and lay within the estimated confidence interval (CI) of the EXP–LN mixture. By resampling the 10-cell *SOD2* data, we found that at least 50 observations were required to arrive at an accurate result (*SI Appendix*, Fig. S3*C*), confirming our earlier estimates using simulated data (*SI Appendix*, Fig. S3 *A* and *B*). The *SOD2* parameterization suggested that maximum-likelihood inference could correctly extract single-cell information from 10-cell sampling data.

Maximum-Likelihood Inference of Coordinated Stochastic Transcriptional Profiles. Programs of gene expression are often controlled by common upstream factors that enforce the regulatory state. We reasoned that coordinated single-cell gene programs would be the product of an overarching regulatory heterogeneity characterized by a shared F. If true, then it should be possible to estimate the expression frequency more confidently and with fewer samples by extending maximum-likelihood inference to gene clusters with coordinated 10-cell fluctuations.

We extended the approach as follows (Fig. 3*A*). First, each gene within the cluster was assigned its own μ_1 and μ_2 (or μ and λ for the EXP–LN mixture) to account for gene-to-gene differences in expression level and detection sensitivity. Next, we assumed that the genes within a cluster share a common F and σ (or F, σ_1, and σ_2 in the relaxed LN–LN mixture), implying a shared mechanism of regulation (39, 40). Therefore, each mixture model of a cluster of g genes involved $2g + 2$ or $2g + 3$ parameters. Even for small gene programs ($g \leq 10$), this parameter search space was too large for nonconvex optimization methods to maximize the global likelihood function quickly (*Methods*). To increase the speed and efficiency of estimation, the cluster was broken down into smaller four-gene subgroups to infer log-means for each gene in the subgroup together with local estimates of F, σ, and λ (steps 1 and 2, Fig. 3*A*). After log-means were locally estimated, the remaining parameters were globally inferred by remaximizing the likelihood function for the entire gene cluster while retaining the local gene-specific estimates of μ_1 and μ_2 (LN–LN mixture) or μ (EXP–LN mixture) (steps 3 and 4, Fig. 3*A*). As before, selection of the LN–LN, relaxed LN–LN, and EXP–LN mixture model was made according to the lowest BIC score (*SI Appendix*, Table S1). This revised formulation of maximum-likelihood inference enabled accurate and confident estimates of the expression frequency while requiring only approximately one-third of the sample size (*SI Appendix*, Fig. S6).

We tested our extension of maximum-likelihood inference by extracting from an earlier study two coexpression clusters that were completely uncharacterized (27) (*SI Appendix*, Fig. S7). These clusters contained one to two dozen genes with strongly coordinated expression fluctuations across 16 samples of 10 ECM-

attached cells, but the patterns of fluctuation were markedly different (Fig. 3 *B* and *C*). Accordingly, when we inferred the parameters for the two clusters, the model predicted two very different expression frequencies. The first "infrequent" gene cluster was predicted to be up-regulated in ~25% of the ECM-attached population (Fig. 3*B*). The LN–LN mixture model was preferred over the EXP–LN or relaxed LN–LN mixtures (*SI Appendix*, Table S1), although all three models converged upon similar values for *F*. By contrast, the expression frequency of the "rare" second cluster was predicted to be ~10% by the LN–LN mixture (Fig. 3*C*), which was the best-scoring model of the three (*SI Appendix*, Table S1). Our parameterization of the two clusters emphasizes the mosaicked regulatory states that evolve even in a very simple model of tissue architecture (27, 38, 41).

To test whether the predicted values of *F* were accurate within the coexpressed clusters, we designed and validated riboprobes for four or five genes in each cluster and quantified their frequency of high expression by RNA FISH (*SI Appendix*, Fig. S8 *A* and *B*). We found that transcripts in the infrequent expression cluster were strongly expressed in three to five ECM-attached cells per acinus cross-section (Fig. 3 *D* and *F* and *SI Appendix*, Fig. S9*A*), yielding an average expression frequency of ~25%. Conversely, genes in the rare expression cluster (Fig. 3 *E* and *G* and *SI Appendix*, Fig. S9*B*) were strongly expressed in one or two ECM-attached cells per acinus cross-section, consistent with an expression frequency of ~10%. The expression frequencies of both clusters closely agreed with the inferred *F* parameters, suggesting that our extended inference approach was effective and accurate.

We evaluated the estimates of expression frequency more broadly by selecting four additional clusters from the same dataset for parameterization (*SI Appendix*, Fig. S7) (27). The clusters showed distinct fluctuation patterns and consequently led to *F* estimates that ranged from less than 5% to greater than 25% (Fig. 4 *A*–*D*, *Upper*). We validated riboprobes for a representative gene in each cluster and scored the expression frequency (Fig. 4 *A*–*D*, *Lower* and *SI Appendix*, Fig. S8*C*). Together with the earlier clusters, we observed a strong correlation between the expression frequency inferred computationally and the manual counts obtained by RNA FISH ($R = 0.89$, Fig. 4*E*). The accuracy of the manual counts was further confirmed by correlation with an expression-frequency index derived from digital image analysis of segmented acini (*Methods* and *SI Appendix*, Fig. S10). Taken together, these data indicate that maximum-likelihood inference accurately infers single-cell expression frequencies from cluster-wide patterns of 10-cell fluctuations.

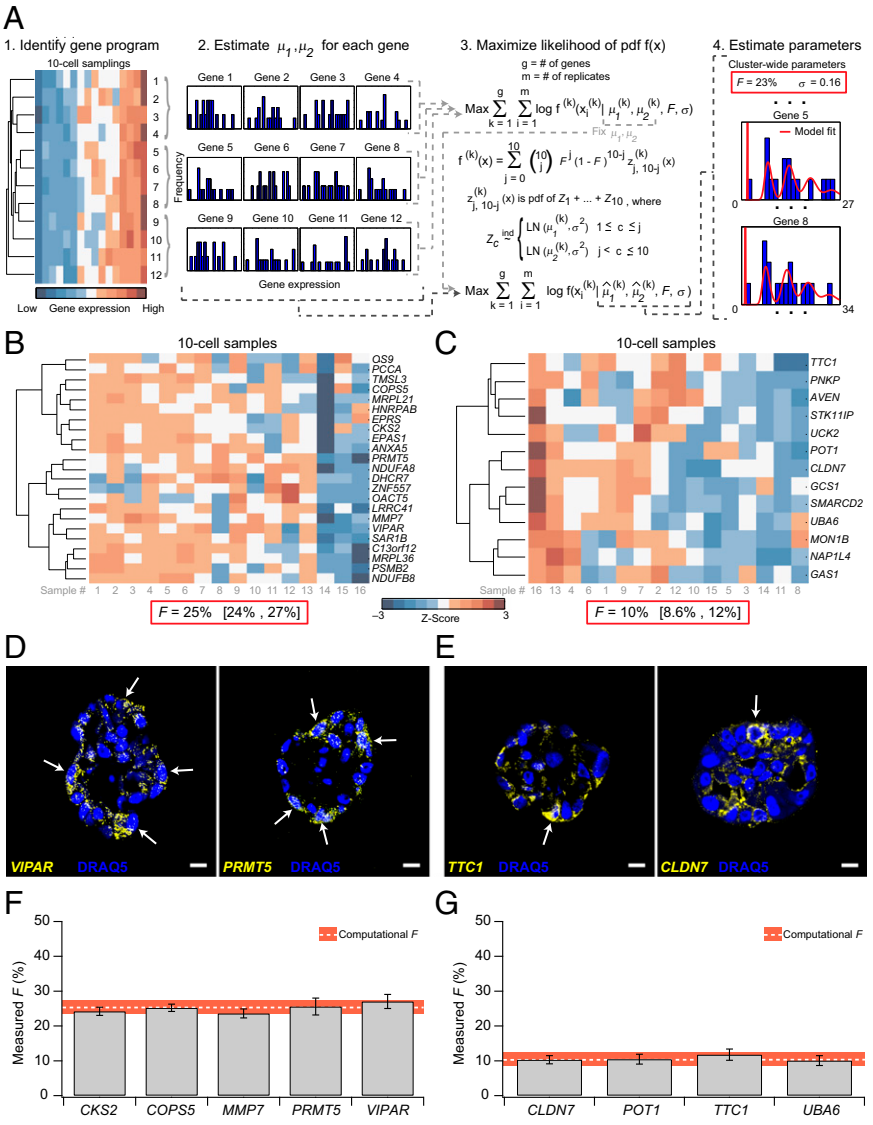

Fig. 3. Maximum-likelihood inference accurately estimates subpopulation frequencies from 10-cell gene-expression clusters. (*A*) The maximum-likelihood approach was modified for gene clusters with coordinated 10-cell sampling fluctuations as follows. (1) Global gene measurements are grouped and assumed to share a common *F* and σ. (2) An expression cluster of interest is divided into four-gene subsets for the first round of parameter estimation of μ_1 and μ_2 for each gene in the subset. (3) A maximum-likelihood estimator is derived based on an expanded version of the model in Fig. 2*A*, where each gene *k* in a group of genes, {*1, . . ., g*}, has its own $\mu_1^{(k)}$ and $\mu_2^{(k)}$. The likelihood function is maximized to infer $\mu_1^{(k)}$ and $\mu_2^{(k)}$ locally. (4) The likelihood function is then remaximized for the entire dataset keeping the log-mean estimates ($\hat{\mu}_1^{(k)}$ and $\hat{\mu}_2^{(k)}$) fixed to provide clusterwide estimates of *F* and σ. Note that each gene has a different range of gene expression to reflect differences in overall expression levels, which are captured in the model predictions as well. (*B*–*G*) Prediction and validation of expression frequency for heterogeneously expressed gene programs during breast-epithelial acinar morphogenesis. (*B* and *C*) Heat map of clustered 10-cell transcriptional profiles (27). Gray labels indicate the 10-cell sample numbers from Fig. 3*B*. Maximum-likelihood estimate of expression frequency (red box) is shown with 95% CI in brackets for each cluster. Note that the two gene clusters are predicted to have substantially different frequencies of high expression based on their 10-cell sampling fluctuations. (*D* and *E*) Representative RNA FISH images of transcripts from (*D*) the infrequent cluster and (*E*) the rare cluster. Images are shown with DRAQ5 counterstain to visualize nuclei. Arrows show ECM-attached cells with high expression. (Scale bar, 10 μm.) (*F* and *G*) Percentage of cells showing high expression by RNA FISH (gray bar) of a subset of genes in each cluster compared with the maximum-likelihood estimate of *F* (white dashed line). RNA FISH data are shown as the mean percentage with 95% CI of ECM-attached cells showing high expression. Maximum-likelihood predictions are shown as the parameter point estimate (white) with 95% CI (red).

Identification of a Peculiar, Very Rare Transcriptional Regulatory State. Maximum-likelihood inference provides critical information about the state distribution and expression frequency of any gene cluster identified by stochastic profiling to be heterogeneously regulated. As a proof-of-concept application, we screened gene clusters from the 3D profiling data (27) to identify unusual regulatory states that warranted follow-up study. One cluster was notable among those surveyed because the predicted expression frequency of the high regulatory state was very rare (F = 2.3%). The very rare cluster was also distinguished by its strong concordance with the relaxed LN–LN mixture compared with the alternative mixture models (*SI Appendix*, Table S1). Moreover, the log-mean of the low regulatory state was extremely low (μ_2 ~–3.3), suggesting that the cluster was at or below detection in the population. Within this coexpression cluster, we recognized several genes that were strongly associated with breast cancer, including the breast cancer susceptibility gene *BRIP1* [alternatively called *FANCJ* or *BACH1* (42)], the breast cancer-associated gene *IRF2* (43), and the zinc-finger gene *HIVEP2*, which is frequently down-regulated or mutated in breast cancer (44, 45) (Fig. 5*A*). We speculated that genes within the cluster were tightly regulated so that they could be activated in a restricted cellular context where their expression was critical.

Among the genes in the very rare cluster, we were most intrigued by the phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit δ isoform (*PIK3CD* alternatively called *p110δ*). Three-dimensional breast epithelial cultures abundantly express two other PI3K isoforms, *PIK3CA* and *PIK3CB* (Fig. 5*B* and *SI Appendix*, Fig. S11), and it is generally thought that any PI3K isoform can support proliferation and survival (46). Nevertheless, we found that the low-copy expression of *PIK3CD* was transcriptionally up-regulated with delayed kinetics compared with the other PI3K isoforms (Fig. 5*B*), suggesting a unique regulatory mechanism. When *PIK3CD* abundance was visualized in single cells by RNA FISH, we observed a striking pattern. Most cells lacked *PIK3CD* or expressed it at very low levels; however, we consistently identified a sporadic subpopulation of cells (roughly one or two cells every other acinus cross-section) with high *PIK3CD* expression (Fig. 5*C* and *SI Appendix*, Fig. S8*D*). The overall frequency of cells in the *PIK3CD*^{hi} state was somewhat higher than the estimates of F for the cluster, but the inferred frequency agreed with the very rare expression of two other members of the cluster, *FEM1A* and IRF2 (*SI Appendix*, Fig. S12). Together, these observations pointed to an acute (and likely transient) regulatory event triggering the selective induction of cluster genes in single ECM-attached cells.

We next asked whether *PIK3CD* was specifically important for normal acinar morphogenesis. To eliminate the very rare *PIK3CD*^{hi} subpopulation, we perturbed p110δ by two independent methods: RNA interference and the p110δ-specific small-molecule inhibitor, IC87114 (ref. 47, Fig. 5*D*, and *SI Appendix*, Fig. S13). When shPIK3CD cells were placed in 3D culture, we found that acini were larger and distorted, suggesting a defect in proliferation arrest. Using phosphorylated Rb (pRb) as a proliferative marker, we observed that shPIK3CD acini were still cycling after 15 d of 3D culture when shGFP control acini had quiesced (Fig. 5 *E* and *F*). Furthermore, when control cells were cultured with IC87114, we observed sustained proliferation that phenocopied *PIK3CD* knockdown (Fig. 5 *E* and *F*). These data together indicate that p110δ activity stemming from the very rare *PIK3CD*^{hi} regulatory state is critical for normal proliferation arrest of breast epithelia in 3D culture. More broadly, our results with the very rare cluster illustrate how maximum-likelihood inference can be used to hone in on gene programs with an expression frequency or regulatory pattern of interest.

Comparison with Alternative Deconvolution Methods. We compared the performance of our method to other computational approaches for deconvolving mixed populations (48–50). The alternative methods invoked different mathematical formalisms—Bayesian statistics (48), nonnegative matrix factorization (49), and principal component analysis (50)—and none had previously been applied to transcriptional profiles of small samples. Using the sampling fluctuations within the infrequent, rare, and very rare clusters, we attempted estimates of expression frequency and found that all were substantially less accurate than our approach (Table 1). The comparison illustrates that our method is uniquely effective at parameterizing transcriptional regulatory states within cell populations.

Direct Comparison of Single-Cell and 10-Cell Sampling Strategies. Maximum-likelihood inference reconstructs the single-cell expression distribution without the need to measure single cells. Ignoring the technical challenges of global single-cell methods (17, 20, 21, 27), it should also be theoretically possible to recreate the complete expression distribution by measuring many

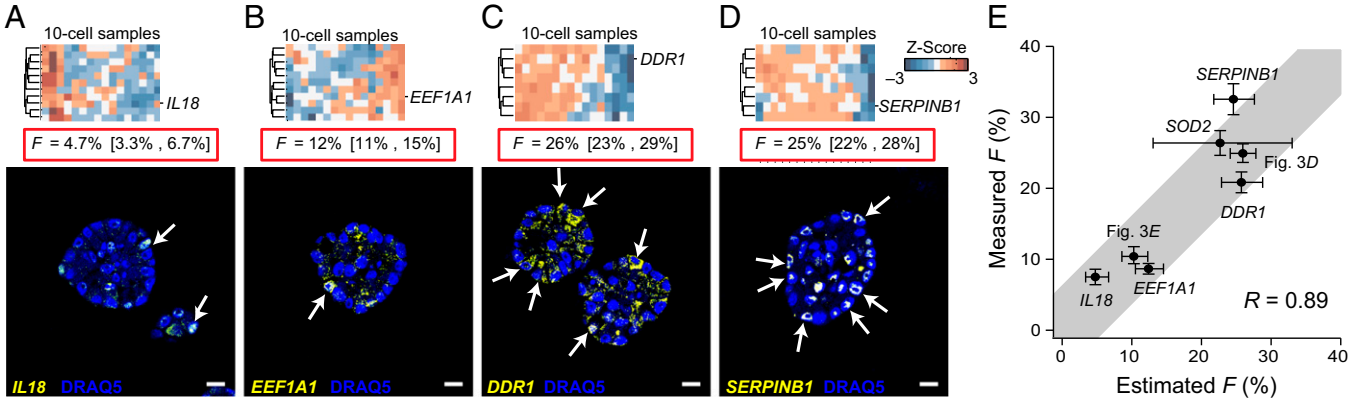

Fig. 4. Widespread parameterization of single-cell expression frequency by maximum-likelihood inference. (*A–D*, *Upper*) Clusters of 10-cell expression fluctuations among ECM-attached cells (27). A complete list of transcripts in each cluster is shown in *SI Appendix*, Fig. S7. Maximum-likelihood estimate of expression frequency (red box) is shown with 95% CI in brackets for each cluster. (*A–D*, *Lower*) RNA FISH images of a representative transcript in each cluster. Images are shown with DRAQ5 counterstain to visualize nuclei. Arrows show ECM-attached cells with high expression. (Scale bars, 10 μm.) (*E*) Percentage of cells scored for high expression by RNA FISH compared with the maximum-likelihood estimate of F. RNA FISH data are shown as the mean percentage with 95% CI of ECM-attached cells showing high expression. Maximum-likelihood predictions are shown as the parameter point estimate with 95% CI. The gray bar shows a one-to-one correspondence with 5% measurement tolerance. Estimates for *SOD2* are reprinted from Fig. 2*H*. Estimates for Fig. 3 *D* and *E* were calculated by pooling all scored transcripts within each cluster. Pearson correlation (R) between measured and inferred expression frequencies is shown.

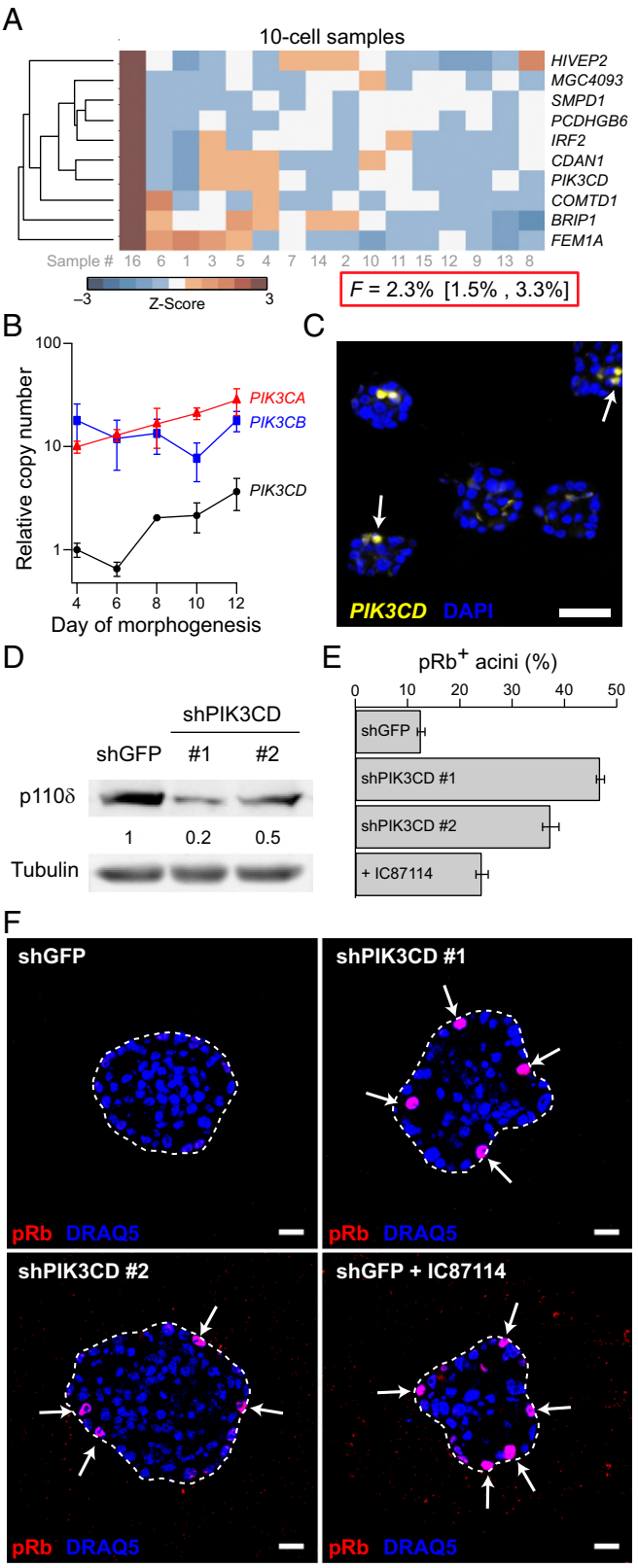

Fig. 5. A unique, very rare regulatory state is marked by *PIK3CD*, which is important for normal suppression of proliferation during breast-epithelial acinar morphogenesis. (*A*) Heat map of clustered 10-cell transcriptional profiles (27). Gray labels indicate the 10-cell sample numbers from Fig. 3*B*. Maximum-likelihood estimate of expression frequency (red box) is shown with 95% CI in brackets. (*B*) *PIK3CD* expression is up-regulated during 3D morphogenesis. Relative *PIK3CA* (red), *PIK3CB* (blue), and *PIK3CD* (black) expression was measured by qPCR at various time points during 3D morphogenesis. Data are shown as mean expression ± SEM normalized to the day-4 expression of *PIK3CD* of three independent experiments. *PIK3CG* was not expressed in MCF10A-5E cells (*SI Appendix*, Fig. S11). (*C*) Representative RNA FISH image of *PIK3CD* expression is shown with DRAQ5 counterstain to visualize nuclei. Arrow shows ECM-attached cells with high expression of *PIK3CD*. (Scale bar, 20 μm.) (*D*) Knockdown of p110δ by shRNA. MCF10A-5E cells were infected with either shGFP (lane 1) or with one of two shRNA sequences targeting p110δ (lanes 2 and 3). Lysates were analyzed by immunoblotting with tubulin used as the loading control. Densitometry of p110δ abundance is shown relative to the shGFP control. (*E* and *F*) Disruption of normal *PIK3CD* regulation elicits a hyperproliferative phenotype in 3D culture. shGFP, shPIK3CD #1, and shPIK3CD #2 cells or shGFP cells + 20 μM p110δ inhibitor IC87114 were fixed at day 15 of 3D morphogenesis, stained for pRb (red), and analyzed by confocal immunofluorescence. Cells were counterstained with DRAQ5 (blue) to label nuclei. Arrows in *F* highlight pRb-positive cells. (Scale bars, 20 μm.) Quantification of proliferating acini in each condition is shown in *E* as the mean ± SEM of eight independent experiments.

individual cells. However, it was not clear whether single-cell profiling would be as effective as stochastic profiling when reconstructing from a limited number of 1- or 10-cell samples. We anticipated that low expression frequencies would be particularly difficult for single-cell methods to characterize because of uncertainty in reliably capturing the rare regulatory state.

To compare single-cell profiling with stochastic profiling, we repeatedly simulated 1- or 10-cell measurements of gene clusters with similar characteristics to those previously examined (Figs. 3 *F* and *G* and 5*A* and *Methods*). The three 12-gene clusters varied in their expression fraction—infrequent (F = 25%), rare (F = 10%), and very rare (F = 5%)—and the very rare cluster was simulated as an LN–LN mixture or an EXP–LN mixture. When the number of observations was limited to 16 (as in the actual data), we found that maximum-likelihood inference provided superior estimates of F when using 10-cell groups (Table 2). Estimates from simulated observations of 16 single cells showed substantially higher mean squared error (MSE) for all gene clusters compared with 16 10-cell observations. The larger MSE of one-cell estimates was caused by increases in both the bias and variance of the estimate, whose magnitudes depended on the cluster characteristics and mixture model. These computational simulations provide an upper bound on performance, because experimental error from actual single-cell experiments (17, 19) should blur the data much more. By collecting a greater total number of cells when observations are limited, maximum-likelihood inference of stochastic 10-cell profiles provides a more accurate picture of the single-cell distribution than single-cell profiles.

Discussion

Maximum-likelihood inference of mixed regulatory states enables accurate single-cell expression characteristics to be gleaned from 10-cell measurements. For individual genes, the model requires a large number of samples to obtain precise estimates, and its advantage over explicit single-cell methods is debatable. However, by extending the approach to coregulated gene clusters (27, 39, 40), we can infer expression frequencies much more robustly than single-cell methods when the extent of sampling is limited. In fact, after identifying heterogeneously regulated genes at the transcriptome-wide level by stochastic profiling (20, 27), global inferences are achievable with the same number of random 10-cell samples. Maximum-likelihood inference can thus be immediately incorporated into stochastic profiling studies that seek a further understanding of single-cell regulation (20, 39).

Multiple studies have demonstrated that heterogeneous phenotypes are primed by earlier regulatory nonuniformities in gene expression (1–7). However, to date, these discoveries have relied on either predefined intracellular circuits or a mix of screening and serendipity. By combining stochastic profiling with maximum-

Table 1. Expression frequency estimates from alternative deconvolution methods

Method	Stochastic-profiling cluster		
	Infrequent	Rare	Very rare
Erkkilä et al. (48)*	20%	∼0%†	∼0%†
Repsilber et al. (49)	22%	60%	25%
Tolliver et al. (50)	18, 40, 23, 19%‡	59, 7.2, 11, 22%	30, 21, 19, 30%
Maximum-likelihood inference	25% [24%, 27%]§	10% [8.6%, 12%]	2.3% [1.5%, 3.3%]
RNA FISH	25% [24%, 26%]§	10% [9.4%, 12%]	5.6% [4.7%, 7.3%]

*Bayesian priors were set to 25%, 10%, and 5% for the infrequent, rare, and very rare clusters, respectively.
†The estimated frequency was 2×10^{-12}%.
‡A minimum of four subpopulations must be estimated with this deconvolution method.
§Bracket denotes 95% CI.

likelihood inference, one can now examine the single-cell transcriptome for expression frequencies or other regulatory patterns that correlate with a downstream phenotype of interest. Such programs are most likely to contain one or more triggers of the heterogeneous phenotype. For example, our follow-on work with *PIK3CD* suggests that it may enforce a quiescent phenotype in a subpopulation of cells that would otherwise enter the cell cycle.

One day, it may be possible to measure the genome, transcriptome, and proteome accurately and cheaply in single cells. While progress is being made toward this goal (9, 16, 17, 35, 51), in the meantime it is valuable to develop alternative methods with less-stringent sample requirements. Our study shows that a surprising amount of quantitative single-cell information can be deconvolved mathematically from measurements with 10-fold more starting material. The "average" cell might indeed be a myth (52), but that does not mean that small-sample averages of cells cannot point to the truth.

Methods

Single-Cell Model of Regulated Gene Expression. Distributions of transcripts per cell were generated under the steady-state approximation as previously described (30, 31). The basal lognormal regulatory state (Fig. 1*A*, blue) was defined with the following model parameters: $k_{binding}$ = 5, $k_{unbinding}$ = 10, $k_{elongation}$ = 50, and $k_{degradation}$ = 1.

The exponential regulatory state (Fig. 1*B*, blue) was defined with the following model parameters: $k_{binding}$ = 0.5, $k_{unbinding}$ = 10, $k_{elongation}$ = 50, and $k_{degradation}$ = 1. Basal regulatory states were perturbed by increasing $k_{binding}$ by 10-fold (lognormal) or 20-fold (exponential), decreasing $k_{degradation}$ by 3.3-fold (lognormal) or fivefold (exponential), or both. Probability densities were compared with the lognormal and exponential test distributions by integrating over integer copy numbers to generate a representative observation set. Observations and distributions were compared with the qqplot function in MATLAB (The MathWorks).

Table 2. Expression frequency estimates from repeated observations of 1 vs. 10 cells

True *F*	Mixture	Cells	Maximum-likelihood estimate of *F*		
			MSE × 10^{-2}	Bias × 10^{-2}	Variance × 10^{-2}
25%*	LN–LN	1	4.32	−20.76	0.01
		10	0.30	−3.40	0.19
10%†	LN–LN	1	2.35	−2.83	2.27
		10	0.19	1.37	0.17
5%‡	LN–LN	1	19.73	29.09	11.27
		10	0.16	1.73	0.13
5%§	EXP–LN	1	57.18	75.09	0.79
		10	4.50	0.80	1.77

MSE, bias, and variance were calculated across 100 simulations of 16 observations. *F* is defined from 0 to 100×10^{-2}. MSE = bias^2 + variance.
*Parameter set: μ_1 = [0.7–2.0], μ_2 = [-1.5 – -0.3], σ = 0.5.
†Parameter set: μ_1 = [-1.5 – -0.2], μ_2 = [-3.8 – -2.0], σ = 0.5.
‡Parameter set: μ_1 = [-2.0 – -0.9], μ_2 = [–3.8 – -3.1], σ = 0.5.
§Parameter set: μ_1 = [-0.9 – -2.0], λ = [11–145], σ = 2.3.

Simulations of Random 10-Cell Samples. Simulated 10-cell expression profiles were generated in MATLAB with the statistics toolbox or in R. The LN–LN model assumes a binomial distribution for the two regulatory states and a lognormal distribution of the transcripts within each state. For a random *n*-cell sampling (here *n* = 10), the number of cells drawn from the high regulatory state (*h*) was specified by a binomial distribution with parameters *n* and *F*. Next, *h* expression measurements were randomly drawn from a lognormal distribution with log-mean μ_1 and log-SD σ. The remaining $n - h$ expression measurements were also drawn from a lognormal distribution with log-mean μ_2 and log-SD σ. The sum of *n* measurements constituted one stochastic *n*-cell sample. In the EXP–LN model, transcripts from the basal regulatory state were drawn from an exponential distribution with rate parameter λ. This procedure was repeated for the indicated number of random samples.

Derivation of LN–LN Maximum Likelihood Estimator. To derive the LN–LN maximum-likelihood estimator, we began with a mixed population of cells occupying one of two regulatory states. The basal regulatory state expresses a transcript (*g*) at a low level with log-mean $\mu_2^{(g)}$ and log-SD σ. The induced regulatory state expresses the transcript at a higher level with log-mean $\mu_1^{(g)}$ and log-SD σ. The probability of drawing a single cell from the high regulatory state is characterized by the parameter *F*.

According to the two-state model, the single-cell expression for transcript *g* follows the pdf:

$$f_{mixture}^{(g)} = F \cdot f_1^{(g)} + (1-F) \cdot f_2^{(g)}, \quad [1]$$

where $f_1^{(g)}$ and $f_2^{(g)}$ are defined as

$$f_v^{(g)}(x) = \frac{1}{\sqrt{2\pi}\sigma x} \cdot \exp\left\{ -\frac{\left[\log(x) - \mu_v^{(g)}\right]^2}{2\sigma^2} \right\} \quad \text{for} \quad x > 0 \quad \text{and} \quad v \in \{1,2\}. \quad [2]$$

The i^{th} random sample of transcript *g*, $Y_i^{(g)}$, is the sum of *n* independent single-cell expression measurements (here, n = 10):

$$Y_i^{(g)} = \sum_{j=1}^{n} X_{ij}^{(g)}, \quad [3]$$

where $X_{ij}^{(g)}$ is the expression of transcript *g* in the j^{th} cell of the i^{th} random sample. Together, the random sample $Y_i^{(g)}$ describing the *n*-cell mixture has the pdf

$$f_n\left(y | F, \mu_1^{(g)}, \mu_2^{(g)}, \sigma\right) = \sum_{j=0}^{n} \binom{n}{j} F^j (1-F)^{n-j} f_{(j,n-j)}^{(g)}(y). \quad [4]$$

$\binom{n}{j} F^j (1-F)^{n-j}$ represents the binomial selection of cells from the basal or induced regulatory states with probabilities *F* and $1 - F$, respectively. $f_{(j,n-j)}^{(g)}$ is the density of a sum $Z_1 + \ldots + Z_n$ of independent random variables representing the *n*-cell draw from the mixture model:

$$Z_c \sim \begin{cases} LN(\mu_1^{(g)}, \sigma^2) & \text{if} \quad 1 \le c \le j \\ LN(\mu_2^{(g)}, \sigma^2) & \text{if} \quad j < c \le n \end{cases}. \quad [5]$$

The pdf for the sum of lognormally distributed random variables was approximated as previously described (53).

When expanded to a cluster of m transcripts, the log-likelihood function for the model parameters given k random n-cell samples is

$$\ell(F, \underline{\mu_1}, \underline{\mu_2}, \sigma) = \sum_{g=1}^{m}\sum_{i=1}^{k} \log\left[f_n\left(y_i^{(g)}|F, \mu_1^{(g)}, \mu_2^{(g)}, \sigma\right)\right], \quad [6]$$

where $\underline{\mu_1}$ and $\underline{\mu_2}$ are vectors containing the transcript-specific log-means for the two regulatory states: $\underline{\mu_1} = (\mu_1^{(1)}, \ \ldots \ , \mu_1^{(m)})$ and $\underline{\mu_2} = (\mu_2^{(1)}, \ \ldots \ , \mu_2^{(m)})$. The log-likelihood functions assume that the expression levels of each transcript are independent as defined by the specific mixture model and F. Derivation of all three maximum-likelihood estimators is included in *SI Appendix, Methods*.

Maximum-Likelihood Parameter Estimation and Model Selection. The derived log-likelihood functions in Eqs. **12**–**14** of *SI Appendix, Methods* are maximized by the most likely combination of parameters for the data $Y_i^{(g)}$. To estimate the parameters for the LN–LN mixtures, we required that $\mu_1^{(1)} > \mu_2^{(1)}$. This constraint ensured identifiability because $\ell(F, \underline{\mu_1}, \underline{\mu_2}, \sigma) = \ell(1-F, \underline{\mu_2}, \underline{\mu_1}, \sigma)$. We also transformed F with the logit function and $\underline{\lambda}$ and σ with the logarithm function to enable the use of faster, unconstrained optimization algorithms.

Because the log-likelihood function was multimodal, it precluded the straightforward use of gradient-based approaches to find globally optimal parameter combinations. We solved the high-dimensional nonconvex global optimization problem by combining genetic and simplex algorithms. First, the log-likelihood function was computed at randomly drawn parameter combinations to identify high-likelihood regions in parameter space at computationally low cost. In the regions of highest log likelihood, we then used the Nelder–Mead algorithm (54) to identify local maxima of the likelihood function. We further localized the global optimum by repeating a random search of parameter space around the optimum identified by the Nelder–Mead algorithm. The resulting high-likelihood regions were used to seed another Nelder–Mead optimization. The iterations of random search and Nelder–Mead optimization continued until convergence.

For estimating model parameters from transcriptional clusters, we first considered smaller subgroups of the cluster of interest. The best balance of computational time and stability of the resulting parameter estimates was achieved with four-gene subgroups (*SI Appendix*, Fig. S6). The log likelihood of each subgroup was optimized by the algorithm described above to identify the most-likely parameters for the transcripts in the subgroup. Based on the subgroup estimate, we then kept fixed $\underline{\mu_1}$ and $\underline{\mu_2}$ (for the LN–LN and relaxed LN–LN models) or $\underline{\mu}$ (for the EXP–LN model) and globally inferred F and σ (or F, σ_1, and σ_2 for the relaxed LN–LN model, or $\underline{\lambda}$, F, and σ for the EXP–LN model) by using the optimization algorithm described above. To confirm that the global optimum for each model had been identified, we pursued a constrained optimization in parallel, which required that the two regulatory states be sufficiently distinct from each other. Specifically, the density of the high regulatory state was constrained to be greater than the low regulatory state in the domain between the mode of the high state and the largest observation in the dataset. The likelihoods of the constrained and unconstrained optimizations were compared, and the higher likelihood inference was selected as the best parameterization for that mixture model. Last, the three mixture models were compared according to their BIC score:

$$BIC = -2\ell(\underline{\hat{\theta}}) + c\log(mk), \quad [7]$$

where $\underline{\hat{\theta}}$ is the vector of inferred parameters, c is the number of inferred parameters in the model, m is the number of transcripts in the cluster, and k is the number of n-cell random samples for each transcript. The best model predicted two distinct regulatory states with the lowest BIC score (*SI Appendix*, Table S1).

Approximate 95% CIs for the best model were estimated by numerically computing the inverse Hessian matrix of the negative log-likelihood function evaluated at the optimal parameter combination. Each i^{th} diagonal element (d_i) of this matrix leads to the confidence in the i^{th} inferred parameter ($\hat{\theta}_i$) as follows:

$$95\%\ CI_i = \hat{\theta}_i \pm 1.96\sqrt{d_i}. \quad [8]$$

Source code for the maximum-likelihood parameter estimation can be found at http://hmgu.de/icb/StochasticProfiling_ML.

Inference Comparisons of 1- and 10-Cell Random Samples. We simulated measurements for various gene clusters as described above with either $n = 1$ or $n = 10$, $m = 12$, and $k = 16$ with the mixture model and F specified in Table 1. Values of $\underline{\mu_1}$, $\underline{\mu_2}$, $\underline{\lambda}$, $\underline{\mu}$, and σ were drawn randomly from the individual transcripts comprising the inferences of Figs. 3 *F* and *G* and 4*A*. Model parameters were inferred as described above with the correct value of n in *SI Appendix*, Eqs. **12** and **14**. The inference procedure was repeated 100 times, yielding estimates $\hat{\theta}_i^j$ (j = 1, 2, ... 100) for each true parameter θ_i. This gives the following Monte Carlo estimates of bias, variance, and mean-squared error:

$$Bias(\hat{\theta}_i) = \frac{1}{100}\sum_{j=1}^{100}\hat{\theta}_i^j - \theta_i$$

$$Var(\hat{\theta}_i) = \frac{1}{99}\sum_{j=1}^{100}\left\{\hat{\theta}_i^j - \theta_i\right\}^2 \quad [9]$$

$$MSE(\hat{\theta}_i) = Bias(\hat{\theta}_i)^2 + Var(\hat{\theta}_i)\ .$$

Cell Lines and Culture Conditions. Cell lines and 3D culture conditions are described in *SI Appendix, Methods*.

Stochastic Sampling. Stochastic samples of *SOD2* were collected as previously described (20, 27, 39). Briefly, 3D cultures were snap-frozen and sectioned at day 10 of morphogenesis. Random 10-cell samples of ECM-attached acinar cells were achieved by laser-capture microdissection from cryosections. The RNA collected from these samples was amplified with a custom small-sample mRNA amplification procedure and quantified by qPCR or microarray (20, 27, 39). Microarray-based expression clusters were identified based on correlated expression fluctuations as described (27, 39).

Image Aquisition and Processing. RNA FISH, immunofluorescence, confocal microscopy, manual expression-frequency scoring, and image processing are described in *SI Appendix, Methods*.

Digital Scoring of Expression-Frequency Index. Multicolor RNA FISH images were acquired with wheat germ agglutinin (WGA), the riboprobe of interest, and the loading-control riboprobes as described in *SI Appendix, Methods*. Individual ECM-attached cells were manually segmented in ImageJ using the WGA, riboprobe, and loading-control stains to determine cell boundaries. The segmented regions of interest (ROIs) were saved as a single ZIP file in ImageJ. The pixels within each ROI were extracted and compared against a null pixel distribution composed of a random set of pixels from segmented cells within the same image. The 85th–95th percentiles of the cell ROI and the null distribution were compared after bootstrapping each distribution 300 times. A cell was scored in the high regulatory state if the bootstrapped 90% CI of the cell ROI was consistently greater than the bootstrapped 90% CI of the null distribution when evaluated from the 85th–95th percentile of pixels. Performing the same analysis on the loading-control riboprobes showed that fewer than 1% of all cells segmented showed detectable differences in total RNA expression. Therefore, the expression-frequency index for a field of view was quantified as the number of cells detected in the high regulatory state divided by the total number of cells segmented. At least 18 fields of view with at least 10 cells per field were acquired for each gene analyzed. Source code for image analysis can be found at http://hmgu.de/icb/StochasticProfiling_ML.

shRNA Cloning and Lentiviral RNAi. shRNA sequences against *PIK3CD* were cloned based on the targeting sequences suggested by the RNAi Consortium, except that the XhoI restriction site in the shRNA loop was changed to a PstI site for easier diagnosis during cloning. shGFP control was used as previously described (55). Primers were annealed at 95 °C in annealing buffer (10 mM Tris·HCl, 100 mM NaCl, and 1 mM EDTA) for 5 min on a thermocycler and cooled slowly to room temperature by unplugging the instrument. Annealed primers were phosphorylated in vitro with T4 polynucleotide kinase (New England Biolabs) and then cloned into pLKO.1 puro (56) that had been digested with EcoRI and AgeI. Lentiviruses were packaged and transduced into MCF10A-5E cells as previously described (39). Stable lines were screened for knockdown efficiency by immunoblotting.

Molecular Biology. Riboprobe synthesis, qPCR, immunoblotting, and validation of the IC87114 inhibitor are described in *SI Appendix, Methods*.

ACKNOWLEDGMENTS. We thank Jason Papin and Eric Greenwald for critically reviewing an early version of this manuscript. This work was supported by American Cancer Society Grant 120668-RSG-11-047-01-DMC (to K.A.J.), National Institutes of Health Director's New Innovator Award Program Grant 1-DP2-OD006464 (to K.A.J.), German Science Foundation Grants SPP 1395 and SPP 1356 (to F.J.T.), and German Academic Exchange Service Integrated Action 54367112 (to F.J.T.). K.A.J. is further supported by the Pew Scholars Program

in the Biomedical Sciences and the David and Lucile Packard Foundation. S.S.B. is supported by a Graduate Research Fellowship from the National Science Foundation. C.F. and F.J.T. are supported by the European Union within the European Research Council Grant LatentCauses.

1. Slack MD, Martinez ED, Wu LF, Altschuler SJ (2008) Characterizing heterogeneous cellular responses to perturbations. *Proc Natl Acad Sci USA* 105(49):19306–19311.
2. Raj A, Rifkin SA, Andersen E, van Oudenaarden A (2010) Variability in gene expression underlies incomplete penetrance. *Nature* 463(7283):913–918.
3. Singh DK, et al. (2010) Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol Syst Biol* 6:369.
4. Wernet MF, et al. (2006) Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* 440(7081):174–180.
5. Laslo P, et al. (2006) Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell* 126(4):755–766.
6. Gupta PB, et al. (2011) Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* 146(4):633–644.
7. Tyson DR, Garbett SP, Frick PL, Quaranta V (2012) Fractional proliferation: A method to deconvolve cell population dynamics from single-cell data. *Nat Methods* 9(9):923–928.
8. Loo LH, et al. (2009) An approach for extensibly profiling the molecular states of cellular subpopulations. *Nat Methods* 6(10):759–765.
9. Dalerba P, et al. (2011) Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 29(12):1120–1127.
10. Taniguchi K, Kajiyama T, Kambara H (2009) Quantitative analysis of gene expression in a single cell by qPCR. *Nat Methods* 6(7):503–506.
11. Bendall SC, et al. (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332(6030):687–696.
12. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* 5(10):877–879.
13. Lubeck E, Cai L (2012) Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Methods* 9(7):743–748.
14. Kurimoto K, et al. (2006) An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res* 34(5):e42.
15. Tietjen I, et al. (2003) Single-cell transcriptional analysis of neuronal progenitors. *Neuron* 38(2):161–175.
16. Hashimshony T, Wagner F, Sher N, Yanai I (2012) CEL-Seq: Single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* 2(3):666–673.
17. Ramsköld D, et al. (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 30(8):777–782.
18. Tang F, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6(5):377–382.
19. Shalek AK, et al. (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498(7453):236–240.
20. Wang L, Janes KA (2013) Stochastic profiling of transcriptional regulatory heterogeneities in tissues, tumors and cultured cells. *Nat Protoc* 8(2):282–301.
21. Reiter M, et al. (2011) Quantification noise in single cell experiments. *Nucleic Acids Res* 39(18):e124.
22. Hansen KD, Wu Z, Irizarry RA, Leek JT (2011) Sequencing technology does not eliminate biological variability. *Nat Biotechnol* 29(7):572–573.
23. Riedel N, Berg J (2013) Statistical mechanics approach to the sample deconvolution problem. *Phys Rev E Stat Nonlin Soft Matter Phys* 87(4):042715.
24. Shen-Orr SS, et al. (2010) Cell type-specific gene expression differences in complex tissues. *Nat Methods* 7(4):287–289.
25. Gong T, et al. (2011) Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE* 6(11):e27156.
26. Loo LH, et al. (2009) Heterogeneity in the physiological states and pharmacological responses of differentiating 3T3-L1 preadipocytes. *J Cell Biol* 187(3):375–384.
27. Janes KA, Wang CC, Holmberg KJ, Cabral K, Brugge JS (2010) Identifying single-cell molecular programs by stochastic profiling. *Nat Methods* 7(4):311–317.
28. Bengtsson M, Ståhlberg A, Rorsman P, Kubista M (2005) Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* 15(10):1388–1392.
29. Warren L, Bryder D, Weissman IL, Quake SR (2006) Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc Natl Acad Sci USA* 103 (47):17807–17812.
30. Peccoud J, Ycart B (1995) Markovian modeling of gene-product synthesis. *Theor Popul Biol* 48(2):222–234.
31. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4(10):e309.
32. Munsky B, Neuert G, van Oudenaarden A (2012) Using gene expression noise to understand gene regulation. *Science* 336(6078):183–187.
33. Gaestel M (2006) MAPKAP kinases - MKs - two's company, three's a crowd. *Nat Rev Mol Cell Biol* 7(2):120–130.
34. Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51(5):341–352.
35. Newman JR, et al. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* 441(7095):840–846.
36. Bar-Even A, et al. (2006) Noise in protein expression scales with natural protein abundance. *Nat Genet* 38(6):636–643.
37. Debnath J, Muthuswamy SK, Brugge JS (2003) Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* 30(3):256–268.
38. Debnath J, Brugge JS (2005) Modelling glandular epithelial cancers in three-dimensional cultures. *Nat Rev Cancer* 5(9):675–688.
39. Wang L, Brugge JS, Janes KA (2011) Intersection of FOXO- and RUNX1-mediated gene expression programs in single breast epithelial cells during morphogenesis and tumor progression. *Proc Natl Acad Sci USA* 108(40):E803–E812.
40. Stewart-Ornstein J, Weissman JS, El-Samad H (2012) Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. *Mol Cell* 45(4):483–493.
41. Wang CC, Jamal L, Janes KA (2012) Normal morphogenesis of epithelial tissues and progression of epithelial tumors. *Wiley Interdiscip Rev Syst Biol Med* 4(1):51–78.
42. Seal S, et al.; Breast Cancer Susceptibility Collaboration (UK) (2006) Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* 38(11):1239–1241.
43. Doherty GM, Boucher L, Sorenson K, Lowney J (2001) Interferon regulatory factor expression in human breast cancer. *Ann Surg* 233(5):623–629.
44. Shah SP, et al. (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 486(7403):395–399.
45. Fujii H, et al. (2005) Frequent down-regulation of HIVEP2 in human breast cancer. *Breast Cancer Res Treat* 91(2):103–112.
46. Foukas LC, Berenjeno IM, Gray A, Khwaja A, Vanhaesebroeck B (2010) Activity of any class IA PI3K isoform can sustain cell proliferation and survival. *Proc Natl Acad Sci USA* 107(25):11381–11386.
47. Knight ZA, et al. (2006) A pharmacological map of the PI3-K family defines a role for p110alpha in insulin signaling. *Cell* 125(4):733–747.
48. Erkkilä T, et al. (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics* 26(20):2571–2577.
49. Repsilber D, et al. (2010) Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* 11:27.
50. Tolliver D, et al. (2010) Robust unmixing of tumor states in array comparative genomic hybridization data. *Bioinformatics* 26(12):i106–i114.
51. Navin N, et al. (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472(7341):90–94.
52. Levsky JM, Singer RH (2003) Gene expression and the myth of the average cell. *Trends Cell Biol* 13(1):4–6.
53. Fenton L (1960) The sum of log-normal probability distributions in scatter transmission systems. *IRE Trans Commun Syst* 8(1):57–67.
54. Nelder JA, Mead R (1965) A simplex method for function minimization. *Comput J* 7(4): 308–313.
55. Orimo A, et al. (2005) Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. *Cell* 121(3):335–348.
56. Moffat J, et al. (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* 124(6):1283–1298.