Characterizing heterogeneous cellular responses to perturbations

Michael D. Slack\textsuperscript{a,b}, Elisabeth D. Martinez\textsuperscript{a,c}, Lani F. Wu\textsuperscript{a,b,1}, and Steven J. Altschuler\textsuperscript{a,b,1}

\textsuperscript{a}Department of Pharmacology, \textsuperscript{b}Green Center for Systems Biology, Simmons Cancer Center, and the \textsuperscript{c}Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center, Dallas, TX 75390

Edited by Alfred G. Gilman, University of Texas Southwestern Medical Center, Dallas, TX, and approved October 16, 2008 (received for review July 20, 2008)

Cellular populations have been widely observed to respond heterogeneously to perturbation. However, interpreting the observed heterogeneity is an extremely challenging problem because of the complexity of possible cellular phenotypes, the large dimension of potential perturbations, and the lack of methods for separating meaningful biological information from noise. Here, we develop an image-based approach to characterize cellular phenotypes based on patterns of signaling marker colocalization. Heterogeneous cellular populations are characterized as mixtures of phenotypically distinct subpopulations, and responses to perturbations are summarized succinctly as probabilistic redistributions of these mixtures. We apply our method to characterize the heterogeneous responses of cancer cells to a panel of drugs. We find that cells treated with drugs of (dis-)similar mechanism exhibit (dis-)similar patterns of heterogeneity. Despite the observed phenotypic diversity of cells observed within our data, low-complexity models of heterogeneity were sufficient to distinguish most classes of drug mechanism. Our approach offers a computational framework for assessing the complexity of cellular heterogeneity, investigating the degree to which perturbations induce redistributions of a limited, but nontrivial, repertoire of underlying states and revealing functional significance contained within distinct patterns of heterogeneous responses.

Automated microscopy | Cellular heterogeneity | Image analysis

Phenotypic heterogeneity in cellular populations has been observed in diverse physiological processes (1–9), pathophysiological conditions (10, 11), and responses to therapeutics (10, 12, 13). However, interpreting cellular heterogeneity—and its changes in response to perturbations—has been an extremely challenging problem. Does heterogeneity contain biologically (or clinically) important information? If so, what level of resolution is required to extract this information?

The complexity of cellular heterogeneity is not yet well understood: At one extreme, heterogeneity may reflect biological “noise” about a single “average” phenotypic state, whereas at the other extreme, it may reflect unbounded complexity, with each cell representing a distinct state. An intermediate alternative—that we investigate here—is that cellular populations can be well described as mixtures of a limited number of phenotypically distinct subpopulations. A practical challenge is to determine the complexity of observed cellular phenotypes and develop methods for identifying patterns of cellular heterogeneity.

Here, we developed methods for characterizing patterns of spatial heterogeneity observed within cellular populations. To capture cellular heterogeneity, we made use of high-content image-based assays, an increasingly powerful approach for capturing cellular responses to perturbations and facilitating the objective quantification of cellular phenotypes (14–28). [An interesting alternative—not considered here—is to study heterogeneity captured by fluorescence-activated cell sorting (FACS), which provides increased numbers of readouts per cell, but without spatial resolution (29)]. First, we applied image-processing algorithms to extract phenotypic measurements of the activation and colocalization patterns of cellular readouts from large numbers of cells in diverse conditions. Next, we used unsupervised clustering algorithms to identify phenotypic “stereotypes” within the overall population and assign probabilities to cells belonging to subpopulations modeled on these stereotypes. Subpopulation identification is based on application of a Gaussian mixture model (GMM), which has been applied recently in the context of online phenotype discovery (30). Finally, for each population (or condition), we estimated the fraction of cells in each subpopulation and summarized the results as a probability vector. The resulting “subpopulation profiles” are human- and machine-interpretable and enable the comparison of heterogeneous responses across multiple physiologic, pathological, or environmental perturbations.

To develop our approach, we analyzed existing and new image databases of drug-treated HeLa cells. Drug treatments provided reliable and well-characterized mechanistic classes of perturbations that could be varied in both time and dose. Furthermore, existing drug classifications provided a biologically motivated method for estimating a crucial property of cellular heterogeneity. Namely, whether increasing numbers of subpopulations provided improved ability to distinguish different mechanistic classes of perturbations. [Traditionally, the tradeoffs between model fit and complexity are addressed by a nonbiological criterion, such as by Bayesian Information Criterion (B.I.C.) or gap statistics (30).] Thus, drug perturbations and their classifications provided an unbiased criterion for determining the degree of complexity required to interpret patterns of cellular heterogeneity.

Results

To study the heterogeneous effects of perturbations on cells, we made use of an existing image database of HeLa cells treated with 100 drugs at 13 concentrations and stained by using a multiplexed collection of immunofluorescent markers (DNA/SC35/anillin) (25). However, this dataset did not contain temporal information and additionally lacked consistent numbers of representative drugs in several key categories of drug mechanism. Thus, we created a 25-drug dataset capturing the responses of HeLa cells to drugs at varying concentrations [low or high, as specified in Supporting Information (SI) Table S1] and time points (3, 6, 12, or 24 h). For this dataset, we selected 5 drugs in each of 5 known functional categories, affecting DNA replication (e.g., methotrexate), histone deacetylation (e.g., trichostatin A), microtubules (e.g., paclitaxel), the glucocorticoid receptors (e.g., dexamethasone), and topoisomerase (e.g., doxorubicin). We additionally included an “other” category containing 10 drugs of miscellaneous (e.g., dichloroacetic acid) or unspecified (e.g., green tea-extracted polyphenols) mechanistic classes.

Author contributions: M.D.S., E.D.M., L.F.W., and S.J.A. designed research; M.D.S. and E.D.M. performed research; M.D.S. analyzed data; and M.D.S., E.D.M., L.F.W., and S.J.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

To whom correspondence may be addressed: Department of Pharmacology, University of Texas Southwestern Medical Center, 6001 Forest Park Road, ND9.214, Dallas, TX 75390. E-mail: lani.wu@utsouthwestern.edu or steven.altschuler@utsouthwestern.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0807038105/DCSupplemental.

© 2008 by The National Academy of Sciences of the USA

www.pnas.org/cgi/doi/10.1073/pnas.0807038105

PNAS | December 9, 2008 | vol. 105 | no. 49 | 19305–19310
anism. After fixation, cells were stained with 1 of 2 sets of fluorescent markers (DNA/phospho-p38/phospho-ERK or DNA/actin/α-tubulin; see Table S2) selected as general readouts of cell signaling state (25).

For each image, we identified individual cellular regions and extracted phenotypic features (Fig. 1A and B). There are multiple approaches for extracting phenotypic information, including detecting either known biological phenotypes [e.g., apoptosis (31) or cell cycle phases (32) or general image properties, e.g., cell morphology and patterns of marker intensities (14, 18)]. Here, we found that a feature set designed to measure the degree of marker colocalization performed well in capturing phenotypic differences in signaling state among visually distinct cells. First, we measured ratios of marker intensities at each pixel to estimate the degree of marker colocalization. To avoid bias due to selection of a “favored” denominator, we measured all 3 possible pairs of marker ratios. Second, we quantized marker ratios into 16 bins. (Bin ranges were empirically chosen based on our data; the highest bin value was set to “infinity” to capture ratios when marker intensities in the denominator were close to background.) Third, we used 6 16 × 16 histograms to tally each of the three quantized marker ratio pairs for both cytoplasmic and nuclear subcellular regions. To down-weight low-intensity pixels, the contribution of each pixel to the histogram was weighted by its total intensity of markers. Finally, each histogram was normalized to have unit mass.

Taken together, these histograms resulted in a 1,536-dimensional (equal to 6 × 16 × 16) feature vector. These high-dimensional feature vectors were subsequently reduced to ~25 dimensions by principal-components analysis (PCA) (Materials and Methods). Although alternative choices of cellular features could certainly be substituted in the remaining analysis, we found that these ratio-metric features performed well at grouping cells with visually similar signaling states in similar regions of feature space.

To enable comparisons of cellular heterogeneity across experimental conditions, a reference training set of 10,000 cells was constructed for each marker set from a uniform sampling of control and drug-treated wells (Materials and Methods). This sample represented only a small fraction (<2%; see Materials and Methods) of our overall datasets. Probabilistic clustering was then applied to this training set to characterize cell heterogeneity as stochastic variation of feature vectors around a collection of distinct phenotypes (Fig. 1B and C). For a specified number of subpopulations $k$, expectation maximization (EM) clustering (33) characterized the total population as a weighted mixture of $k$ subpopulations, each defined by a Gaussian distribution around a distinct cellular stereotype and a (prior) overall probability of subpopulation occurrence (30) (Fig. 1C, stacked bar graph). (For convenience of visualization we illustrate our approach with $k = 4$; the effects of varying $k$ will be discussed subsequently.) The reference subpopulation model is thus defined by this collection of $k$ phenotype means, covariances, and prior probabilities of occurrence.

The heterogeneity of a cell population subjected to a specific experimental condition can be estimated by using the reference subpopulation model. For each cell, the (posterior) probability that it is a member of a given subpopulation can be computed from the subpopulation model by using Bayes’ rule. These probabilities for a single cell and all $k$ subpopulations can be represented as a $k$-dimensional vector whose entries sum to 1. By averaging these
drug-treated conditions to controls. A comparison of camptothecin (topoisomerase) to controls, for example, showed a significant shift to a dominant phenotype (Fig. 2A; e.g., dose 5, population 3). Comparison of colchicine and nocodazole (both targeting microtubules) to controls showed substantial redistributions of phenotypic heterogeneity after drug treatment (Fig. 2A; e.g., dose 9, population 2). We next compared the effects of drugs to one another. We found that drugs of similar mechanism often yielded similar profiles (e.g., Fig. 2A, colchicine vs. nocodazole). Conversely, mechanistically different drugs yielded distinct profiles (e.g., Fig. 2A, camptothecin vs. colchicine).

Dose responses typically revealed sharp transitions in subpopulation proportions as drug concentration increased, whereas untreated populations tended to remain in constant proportions (Fig. 2A). Multiphasic drug effects, such as seen for camptothecin, can be easily discerned as dramatic transitions in subpopulation proportions [Fig. 2A, second image, transitions at doses 5 and 10 in agreement with previous results (18, 25)]. The interpretation of results across time is complicated somewhat by the fact that controls did not maintain constant proportions, likely because of varying exposure times to the drug vehicle DMSO before fixation (Fig. 2B, first image). Although most drugs did not show significant differences from the controls at the 3- and 6-h time points, significant variation was observed at the 12- and 24-h time points (Fig. 2B). We note that analysis of automatically identified subpopulations revealed no significant enrichment for any specific cell cycle stage, as might be expected from our choice of pixel-based features (SI Text and Fig. S2). These observations suggest that subpopulation profiles contain information that may offer insights into drug modes of action.

We next wondered whether the significant redistributions of subpopulations were consistent across drug perturbations of similar mechanism. To assess the “information content” of observed patterns of heterogeneity, we computed subpopulation profiles for our 25 annotated drugs at the 24-h time point and grouped the results by category of mechanism (Fig. 2C). (Here, we visualized these results using a heatmap, rather than stacked bar graphs, to provide easier visual comparison among larger numbers of drugs.) Whereas DNA replication-inhibiting drugs showed high variability in their subpopulation proportions, the responses of drugs within all other categories generally showed high concordance. Rigorous assessment of the similarities and differences between drug effects requires a quantitative measure of “distance” between subpopulation profiles. When considered as probability vectors, subpopulation profiles may be viewed as information theoretic descriptions of cell population heterogeneity. Similarities among profiles can be quantitatively compared by using the Kullback–Leibler (KL) divergence (34) and visually displayed by using multidimensional scaling (MDS) plots (SI Text). We observed that profiles showed significant grouping by drug mechanism (Fig. 3A, k = 4). Thus, drugs of similar mechanism appear to induce similar patterns of cellular heterogeneity.

A natural question arose from our analysis: Would increasing the number of subpopulations, k, in the reference model discern finer distinctions between drug effects and thus yield a more accurate classification of drugs? In theory, increasing k allows finer resolution at which to measure cellular heterogeneity, although classification accuracy for very large values of k may fail to improve—or even degrade—because of the introduction of nonbiological separation of cellular phenotypes. We found for the DNA/p-p38/p-ERK and DNA/actin/α-tubulin marker sets that categorization performance for most drug categories sharply improved as k increased from 1 (no predictive power) to ~5. Afterward, no significant improvement in classification was observed (Fig. 3B). In a few cases, categorization performance was poor. For example, the DNA/actin/α-tubulin marker set showed poor performance for the microtubule category (Fig. 3B Middle). However, as previously shown, this poor performance was due to the fact that the micro-

$k$-dimensional vectors over a population of cells, a subpopulation profile is obtained as the expected overall proportion of each subpopulation for a given experimental condition. Replicates produced reproducible subpopulation profiles (Fig. S1), which were averaged to obtain 1 final profile per condition. Thus, the heterogeneity observed within any experimental condition can be concisely encoded by a probability distribution over the reference subpopulations.

We next examined the degree to which subpopulation profiles can be used to distinguish heterogeneous responses to different drugs (Fig. 2A). We first compared subpopulation profiles for...
A small number of phenotypic subpopulations can distinguish drug effects by mechanism. (A) As the number of subpopulations in the reference model increases (k = 2, 4, 10, and 100), MDS plots (40) reveal relative similarities among drugs (computed by using KL divergence) within 3 mechanism categories (H, blue diamonds; M, green triangles; G, orange stars). Axes are labeled in arbitrary units. (B) Drug classification performance is computed for each value of k (and drug category; see also Table S5). Performance is measured by −log10 of the P value for a rank-sum statistic [Mann–Whitney U test (39)] obtained by comparing distributions of intra- and intercategory subpopulation profile similarity scores. Profile similarity is computed for 12- and 24-h time points of low drug concentration and combined by rank addition. Dotted lines indicate significance threshold corresponding to P = 0.05. Black arrowheads (Left) (k = 2, 4, 10, 100) correspond to MDS plot in A. Black arrowheads (Right) (k = 7) correspond to MDS plot at bottom right and uses reference model shown in Fig. S1D.

Discussion
Phenotypic heterogeneity is often ignored or viewed as an impediment to understanding the effects of perturbations on cellular populations. Traditional measurements of a mean and a variance typically do not capture the complexity of cellular responses to perturbations often observed in microscopy. On the other hand, cell-to-cell variation suggests a potential for limitless cellular complexity. Analysis of our datasets suggested that the effects of perturbations may be characterized largely as redistributions of a limited, but nontrivial, repertoire of underlying states. Subpopulation profiles provided an intuitive and computationally tractable approach to understanding the effects of perturbations on cellular populations.

Fig. 3. A small number of phenotypic subpopulations can distinguish drug effects by mechanism. (A) As the number of subpopulations in the reference model increases (k = 2, 4, 10, and 100), MDS plots (40) reveal relative similarities among drugs (computed by using KL divergence) within 3 mechanism categories (H, blue diamonds; M, green triangles; G, orange stars). Axes are labeled in arbitrary units. (B) Drug classification performance is computed for each value of k (and drug category; see also Table S5). Performance is measured by −log10 of the P value for a rank-sum statistic [Mann–Whitney U test (39)] obtained by comparing distributions of intra- and intercategory subpopulation profile similarity scores. Profile similarity is computed for 12- and 24-h time points of low drug concentration and combined by rank addition. Dotted lines indicate significance threshold corresponding to P = 0.05. Black arrowheads (Left) (k = 2, 4, 10, 100) correspond to MDS plot in A. Black arrowheads (Right) (k = 7) correspond to MDS plot at bottom right and uses reference model shown in Fig. S1D.
method for identifying functional significance from patterns of heterogeneity. Analysis of subpopulation profiles revealed that drugs of (dis)similar mechanism induced (dis)similar patterns of subpopulation redistributions. Thus, our work suggests that patterns of heterogeneity can be quantified and used to extract meaningful biological information.

Our approach for grouping cells is unsupervised, and the identification of subpopulations depends on the composition of the dataset, the markers, the extracted features, and the number of subpopulations. In some cases, subpopulation phenotypes and their redistributions may be interpretable. For example, the microtubule-destabilizing drugs vinorelbine and nocodazole induced a decrease in subpopulations characterized by α-tubulin-positive staining. Our approach also identifies subpopulations whose phenotypes cannot be easily interpreted by current biological knowledge. For example, it is unclear how aldosterone induces an increase in subpopulation 3, characterized by weak diffused p-ERK staining coincident with clear p-p38 nuclear staining. In general, the best method for interpreting subpopulations and their redistributions may be through relative comparisons with profiles of perturbations with known mechanisms.

GMMs provide a reasonable first approximation for grouping cells in high-dimensional feature space. However, certain subpopulations may still display some degree of phenotypic heterogeneity (e.g., Fig. 1C Upper, subpopulation 2). This may be due to several factors, including: Not all subpopulations will be well modeled by Gaussian distributions; some subpopulations may cover large regions of feature space and need further subdivisions; and automated grouping based on phenotypic features may not agree with visual grouping. Model parameters in future studies of heterogeneity may be optimized to improve visual tightness of identified subpopulation phenotypes.

Our analysis suggests that observed heterogeneity may reflect variation around a discrete set of phenotypic states. Our work provides a framework for future studies with larger sample sizes, greater numbers of perturbations, and higher numbers of markers to rigorously test questions about the nature of heterogeneity, including: Do unperturbed cells occupy essentially the same phenotypic space as cells with significantly perturbed pathways? And, do perturbations induce redistributions among a limited repertoire of underlying stereotypical states? We envision that the ability to identify and interpret information contained in patterns of cellular heterogeneity will provide insights into physiology and disease that may be missed by traditional population-averaged or small-cell-number studies.

Materials and Methods

Drugs and Concentrations. The drug-dose data of 100 compounds in 13 3-fold serial dilution are as previously reported (25). The new time-course data consisted of 35 compounds at 2 concentrations (see Table S1).

Cell Culture. For the time-course data, HeLa cells were diluted to 5,000 cells per 50 μl of media and plated onto 384-well glass-bottom plates. Plates were incubated at room temperature for 30 min to minimize edge effects in wells (36), then placed in a 37 °C5% CO2 incubator overnight. Cells were fixed at 3, 6, 12, and 24 h after drug treatment. See SI Text for details about cell culture, drug treatment, and plate layout.
Fluorescent Markers. The marker set for the drug-dose data (DNA/SC35/anillin) are as previously reported (25). For the time-series data, 2 sets of fluorescent markers were used: DNA/EP (p38p/ERK) and DNA/anillin/tubulin. See SI Text for details about staining protocol.

**Image Background Correction and Cell Segmentation.** Image background correction was done by using the ImageJ background subtraction rolling-ball algorithm (37), with rolling-ball size set to 500 pixels. Original image dimensions were 1,392 × 1,040 pixels. Cell segmentation methods used to define DNA and non-DNA (cytoplasmic) regions are as in ref. 18 and are summarized in SI Text.

**Feature Extraction.** For each cell, a 1,536 dimensional feature vector was computed based on histograms of the ratios of marker intensities per pixel, in both cytoplasmic and nuclear cellular regions. See Results and SI Text.

**Feature Data Sampling.** Feature data from multiple wells was drawn by using random sampling. The final clustering was performed with 1,000 samplings for each replicate well. Each replicate well was used for training data. The total number of cells (training data) sampled from all plates for computing PCA coordinate proportionate number of total cells detected within the well. The total number of cytoplasmic and nuclear cellular regions. See Feature Extraction.

For the DNA/SC35/anillin marker set). For the DNA/SC35/anillin marker set, all wells were sampled.

**PCA Coordinate Transformations.** Principal-components analysis (PCA) was used to reduce the dimensionality of feature data from 1,536 to 27 for the DNA/p38p/ERK marker set, and to 26 for the DNA/anillin/tubulin marker set. The choice of dimension was made for each marker set by randomizing the order of feature dimensions for each sampled cell and computing the eigenvalues of the resulting (randomized feature) covariance matrix. Subsequent modeling and profiling computations all used the PCA reduced data.

**Reference Models.** Subpopulation reference GMM models were computed from sampled feature data by using the EM algorithm (33). Code was based on the publicly-available Netlab Toolbox. For each model, EM clustering was run 10 times, starting from a K-means clustering (38) using randomly chosen means. The final clustering with the best log-likelihood value was chosen as the subpopulation reference model. To eliminate the influence of convergence failures, each run was attempted up to 5 times with new initial conditions until convergence was reached. For each marker set, subpopulation models were computed for each number of subpopulations clusters in the set (2, 3, 4,..., 18, 19, 20, 30, 40, 50, 100).

**Subpopulation Profiles.** Subpopulation profiles for each drug-treatment replicate (single well) were computed by averaging the posterior probabilities for the corresponding subpopulation model for 1,000 samplings (with replacement) consisting of PCA reduced feature data from 1,000 cells per sample. Profiles for each drug-treatment condition were computed by an average of profiles for each replicate well, averaged by the number of cells observed in each well. When exactly 1 of the 2 replicate wells had <1,000 cells in total, the profile for the other well was used.

**Similarity Comparison.** Individual profiles were compared by using a symmetrized version of Kulback-Leibler divergence (34). In particular, given subpopulation profiles P and Q for a model with k subpopulations, the similarity measure S(P,Q) is defined as

\[ S(P, Q) = \frac{KL(P, Q) + KL(Q, P)}{2}, \]

where

\[ KL(P, Q) = \sum_{i=1}^{k} \log_{2}(P_i/Q_i). \]

**MDS Plots.** The Matlab mdscale function was used to compute data for multidimensional scaling plots (40). Kruskal’s stress-1 metric (total error in similarity representation) was computed to <0.05 for each plot in Fig. 3A, and <0.12 for the plot in Fig. 3B.

**Identification of Exceptional Phenotypes.** For a given drug treatment condition, the model fit score was computed as the ratio of the model log likelihoods for drug treatment data vs. reference training data. Samplings of 1,000 cells were used to compute log likelihoods.

**ACKNOWLEDGMENTS.** We thank Michael Johnson (Nikon) and Jurg Rohrer (BD Biosciences), all members of the S.J.A. and L.F.W. laboratory for extremely helpful discussions and feedback, and the anonymous reviewers for their helpful suggestions. This work was supported by National Institutes of Health Grants R01 GM081549 (to L.F.W.), R01 GM071794 (to S.J.A.), K22 CA118717-01 (to E.D.M.), and P50 CA70907 UT SPORE in Lung Cancer; the Welch Foundation (I-1619 and I-1644 to L.F.W. and S.J.A.); and the University of Texas Southwestern Endowment for Scholars in Biomedical Research (to L.F.W. and to S.J.A.).