Capturing single-cell heterogeneity via data fusion improves image-based profiling

Mohammad H. Rohban[1,2], Hamdah S. Abbasi[1], Shantanu Singh [1] & Anne E. Carpenter [1]

Single-cell resolution technologies warrant computational methods that capture cell heterogeneity while allowing efficient comparisons of populations. Here, we summarize cell populations by adding features' dispersion and covariances to population averages, in the context of image-based profiling. We find that data fusion is critical for these metrics to improve results over the prior alternatives, providing at least ~20% better performance in predicting a compound's mechanism of action (MoA) and a gene's pathway.

[1] Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [2] Computer Engineering Department, Sharif University of Technology, Tehran 11155/1639, Iran. Correspondence and requests for materials should be addressed to M.H.R. (email: rohban@sharif.edu)

As a very early large-scale, high-dimensional, single-cell-resolution data type, high-throughput microscopy experiments have presented one of the first exemplars of the challenges in summarizing and comparing cell populations.

One of the key challenges is creating a profile of each cell population. A profile is a summary of many features of a population that enables efficient comparison with other populations while simultaneously capturing their natural variations and possible subpopulations. Recent studies have yielded many insights into cellular heterogeneity and its importance[1–4].

Although anecdotal evidence of the value of capturing heterogeneity abounds, it has remained puzzling that so-called average profiling, the practice of feature-averaging all single-cell measurements together using measures-of-center (mean or median), has remained the top-ranked approach in the field of image-based (or morphological) profiling, whether those features are raw or pre-processed using unsupervised learning, or whether they derive from classical image processing or deep learning.

In image-based profiling, average profiling has been used as a straightforward way of summarizing a cell population (a sample) into a fixed length vector (a sample's profile), with one value per feature per sample. Various metrics of similarity between profiles of two samples can then be used to infer whether they show similar phenotypic responses to their respective treatments. Average profiling typically results in a thousand-fold decrease in data size (because there are typically around a thousand cells per well in image-based profiling experiments conducted in multi-well plates), which makes downstream processing both computationally manageable and potentially statistically more robust.

However, average profiling results in loss of information about a population's heterogeneity. The information loss can manifest in different forms. For example, multiple configurations of distinct subpopulations of cells could yield identical average profiles. Or, if two subpopulations with opposite phenotypes exist in a sample, they might cancel out yielding a profile indistinguishable from that of a sample that contains neither subpopulation. In addition to information loss, averaging can result in misleading interpretation of feature associations, e.g., Simpson's paradox[5]. Finally, averaging makes the implicit assumption that the joint distribution of the underlying features is unimodal, which if violated can lead to artifacts. In this paper we investigate whether including heterogeneity measures in the profiles of cells undergoing various treatment conditions can improve upon prior methods that do not capture heterogeneity well.

Several methods have been developed in an attempt to capture cell population heterogeneity while still allowing efficient comparisons between different populations. A simple solution is to compute the cell population's dispersion (e.g., standard deviation or median absolute deviation, MAD) for each feature and concatenate these values with the average profile. Although feature normalization brings features to comparable scales, features in average profiles generally follow a probability distribution different from that of the features in dispersion-based profiles. This discrepancy may lead to the correlation between profiles being biased toward either only features of the average or the dispersion. Concatenation can also dilute the signal-to-noise ratio (SNR) if one type of profile already has a low SNR[6], i.e., the SNR of concatenated data would be lower than the maximum of SNRs across data types. In practice, concatenation of median and MAD profiles has been shown to provide only a minor improvement over median profiling alone[7].

Measures of dispersion might only capture a small fraction of the heterogeneity in the data, i.e., they disregard subpopulation structures, because they involve processing each feature separately. Instead of capturing dispersion for each individual feature, one can alternatively model the heterogeneity by clustering cells using all features simultaneously. In this approach, a subset of data is used to estimate clusters of cells (representing subpopulations) and profiles are calculated as the feature averages within subpopulations[8]. Alternatively, cells can be classified into pre-determined phenotype classes using a supervised approach, and the profile is then defined as the fraction of cells in each phenotype class[9]. However, many cell phenotypes are better considered as a continuum of varying morphologies rather than discrete populations. Further, there may exist some rare subpopulations that are unique to a small portion of the data that may be overlooked in the clustering step. As a remedy, if we instead try to cluster each sample separately, the subpopulations may not be appropriately matched across samples, which makes the profiles uncomparable across the samples. Unfortunately, despite their intuitive appeal, none of these ideas have proven to significantly improve upon the baseline average profiling, at least, on the single public dataset with available ground truth, which are annotations of the mechanism of action (MOA) of a small set of compounds. In a comparison of profiling methods, average profiling (after dimensionality reduction) outperformed methods that attempted to capture heterogeneity in the data[7]. More recent work demonstrated a deep learning approach to feature extraction that yielded the highest performance yet, but nonetheless relied on average profiling[10].

Here, we test fusing information from the dispersion profiles with the average profiles at the level of profiles' similarity matrices. This avoids inclusion of features with inherently different probability densities in the final profile. Modeling profile similarity matrices from disparate data types using a graph has been shown to be effective in handling heterogeneous data sources such as DNA methylation, miRNA expression, and mRNA expression[6].

We also consider alternate heterogeneity representations that do not explicitly model subpopulation information, but nevertheless capture heterogeneity. Higher order moments, which consider combinations of features (as opposed to univariate moments of single features such as mean/median or standard deviations) are excellent candidates. As shown schematically (Fig. 1), two cell populations may differ dramatically but have identical means and standard deviations. However, there is a substantial difference in the covariance (a second moment) of two features between the control (on the left) and treatment (on the right) cell populations, making this information useful to include in the populations' profiles.

We also motivate the use of higher order joint statistical moments for profiling from a more theoretical standpoint. In the terminology of estimation theory, we aim to find a sufficient statistic for the unknown subpopulations to serve as the sample profile. A sufficient statistic is a summary of data that provides maximal information about the unknown parameters of a model that is used to explain the data. Previous work has shown that under certain assumptions, the first, second, and third order moments, collectively, are approximately a sufficient statistic for modeling subpopulations given a large enough sample size[11] (Methods). Unfortunately, for typical single-cell datasets, sample sizes are too small, and computational requirements are too high, for estimating third order moments.

Here we find that even sparse random projections[12] of covariances (second-order moments) can provide a substantial improvement in the ability to accurately compare cell populations for phenotypic similarity, when combined with median and MAD profiles via data fusion.

Testing profiling methods against each other is not a trivial exercise, given that the true similarities and differences among
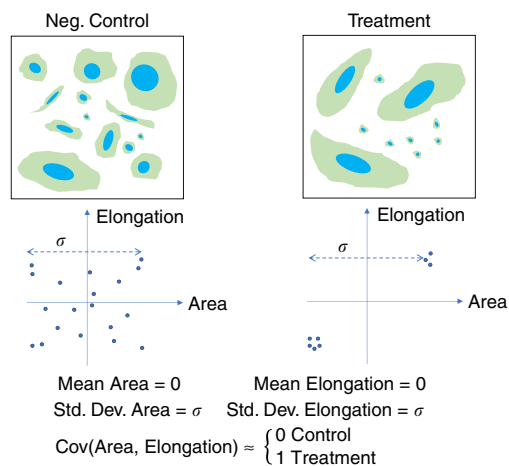
Fig. 1 Features' covariance can capture cell phenotypes better than feature averages or dispersion. In this synthetic example, the negative control sample (on the left) consists of cells displaying heterogeneous morphologies. The treatment, on the other hand, shows two distinct subpopulations. In both cases, the scatter plot helps to see that the mean and standard deviation of both measured cell features (area and elongation) are equivalent in the two cases. However, the two features positively correlate in the treatment condition as opposed to the control. In such a case, the covariance can distinguish the phenotypes better than simple averages (e.g., means and medians) and measures of dispersion (e.g., standard deviations and median absolute deviations)

large sets of cell populations is rarely known. We therefore tested the approach on three different publicly available datasets where some ground truth (i.e., expected results), albeit imperfect, is known. Cell measurements in the datasets are based on Cell Painting, which is an image-based assay designed to capture cell morphology[13]. For these benchmark datasets, our laboratory had released the image data[14,15] but for this study we collected ground truth to create a proper testing scenario. We used datasets that had a sufficient number of perturbations for the data fusion technique to work (Methods), and therefore did not include the dataset reported in a previously published study[7].

To summarize, to account for the single-cell heterogeneity in comparing cell populations, we propose to fuse information from average, dispersion, and randomly projected covariance profiles at the level of profiles' similarity matrices. On three new benchmark datasets, the resulting fused similarity matrix shows significant improvement of concordance to the known ground truth compared to similarity matrices that are based on each type of profile alone, and some methods that were previously proposed to capture population heterogeneity.

Results

Fused profile similarities improves performance. To evaluate each profiling method, we tested whether pairs of cell populations that look most alike, according to the computed image-based profiles, have been treated with perturbations that are annotated as having the same mechanism of action (for compounds) or the same pathway (for gene overexpressions). Similarity between pairs of image-based profiles are established based on the profiles' correlation (Methods).

We find that the enrichment of top-correlated perturbation pairs, whether genetic or chemical, in having the same mechanism of action or the same pathway (Methods), is improved when median absolute deviation and/or covariances are combined with the median profiles through Similarity

Network Fusion (SNF) (Methods) (Fig. 2a and Supplementary Tables 1 and 2). Covariances are summarized by sparse random projections to avoid the curse of dimensionality. The enrichment score we use here is appropriate for applications where the aim is to identify the most highly correlated treatment pairs rather than to classify every treatment into a class[7]. Examples in drug discovery include lead-hopping, identifying gene function, and matching genetic and chemical perturbations. An increase in enrichment score would directly translate to a proportional increase in the chances of finding a novel connections between genes, compounds, and diseases.

The improvements are observed even when MOAs/pathways with few (fewer than five) compound/gene pairs are removed, confirming that the improvement was not driven by a few outlier groups. We also evaluated another method that implicitly accounts for heterogeneity (Factor Analysis on single-cell data[7]), which performed poorer than the proposed method (but had been the best performing in a previously published, smaller dataset in a different application[7]), as well as methods that do not account for heterogeneity (Supplementary Fig. 1). The improved intra-MOA similarities, especially in certain MOAs (Fig. 2b, c, Supplementary Tables 3–5), indicates that median, MAD, and sparse projections of covariances are complementary sources of information.

The improvement in enrichment of top-correlated perturbation pairs is observed across the three datasets, which involve different experimental conditions and perturbation/annotation types. Although trivial concatenation of dispersion measures with the median profile has shown marginal improvement in the past[7], combining MAD and covariance with medians via data fusion provides consistent and substantial improvement (typically around 20%) in the mentioned enrichment score.

We conclude that capturing cell-to-cell heterogeneity is of value in image-based profiling of cell populations. Mean-averaged metrics from deep learning-based feature extractors have shown promise in image-based profiling; these metrics would likely also be improved by the method presented here. These and other strategies can now be tested and the results shared, because we provide the datasets publicly. Source code, image processing pipelines, and gene/compound annotation data to reproduce these results are available (https://github.com/carpenterlab/2018_rohban_natcomm).

Methods
Annotations. We used the Repurposing hub[16], https://clue.io/repurposing-app, to annotate compounds with their mechanism of action (accessed on 28 February 2018). For missing annotations, we used other resources such as https://www.drugbank.ca. The gene overexpression dataset contains biological pathway annotations, generated by domain experts at our institution. For pathways marked as having canonical and non-canonical members, we merged all members.

Rationale. The theoretical basis we present assumes that the cellular phenotypes can be modelled as a mixture of gaussians. This model has been shown to be effective in capturing subpopulations in imaging data[8,17]. In this model, the subpopulations and their proportions correspond to mixture centers and mixture prior probabilities, respectively. Both of these quantities are considered as unknown parameters.

It has been shown that, under mild assumptions, these unknown parameters can be estimated using the first, second, and third moments of data[11]. More specifically, if the mixtures are spherical Gaussians, and their centers are linearly independent, all the unknown parameters can be estimated with high precision given a sufficiently large number of data points (see Theorems 2 and 3[11]). In other words, the first, second, and third moments of the data constitute an approximate sufficient statistic for the unknown parameters in GMM when the sample size is sufficiently large. Average profiling uses only a small portion of this sufficient statistic–the first moment–to represent the sample. We can make this representation richer by also including the second and third moment profiles. Going beyond the third moment does not add any additional information with regard to the GMM (Theorems 2 and 3[11]).
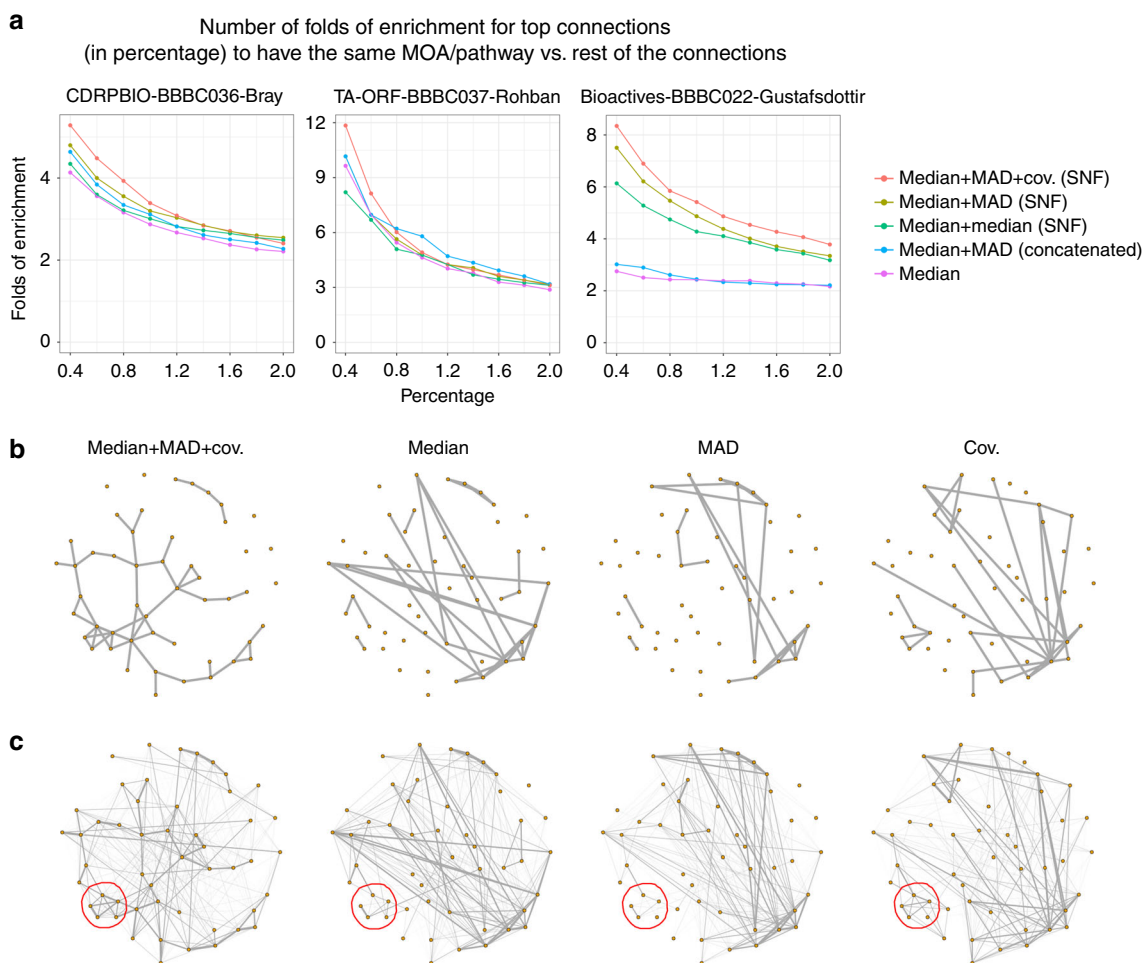
a

b

c

Fig. 2 Fusing metrics of cell heterogeneity increases the percentage of validated connections. **a** When median, MAD and random projections of covariance profiles are combined through SNF (red line), the enrichment in having same MOA/pathway annotations is improved, especially for the strongest, most relevant connections above 0.5%. This is shown in three separate experiments involving small molecules (left, right) and gene overexpression (middle). Enrichment is versus a null distribution, which is based on the remainder of the connections. **b** Similarity graphs for the mechanism of action (MOA) class Adrenergic receptor antagonists, using different types of profiles in CDRPBIO-BBBC036-Bray. This MOA was chosen because it showed the highest improvement upon combining different profiles. The goal is a qualitative view on how data fusion improves within-MOA connectivities. Each node represents a compound, and two nodes are connected if the similarity of their corresponding profiles is ranked among the top 5% most-similar pairs. Median, MAD, and random projections of covariance profiles seem to be complementary for this MOA, as they cover mostly non-overlapping compound connections. The overall connectivity of compounds in this MOA is improved once these profiles are combined through SNF. Graph layouts are the same across data types and are based on the similarities in median + MAD + cov. (SNF); note that this causes the left-most graph to appear less cluttered and less connected, but the main purpose of the visualization is to observe the structure of connections, not the number of connections (which is quantified systematically in part **a**). **c** Weighted similarity graph as in the previous plot except that edge thicknesses are based on an exponential weighting of the ranked similarity values. Sub-clusters that are moderately present in two or three profile types (such as the one marked in red in bottom left) became stronger after applying data fusion using SNF

We did not test the third moment because our datasets contain in the order of a few thousand cells per sample, whereas millions of cells would be needed to robustly compute third moments ($O(d^3)$), where d is the dimensionality of the feature space; on the order of 100 in our case). As well, the dimensionality of the final profiles rapidly grows as d increases. Although computing second-order moments is more feasible, it nonetheless requires dimensionality reduction to be of practical use: for 500 features, the second moment is nearly 125,000-dimensional, which is both computationally and statistically difficult to work with in forming the treatments similarity matrix. We use sparse random projections[12] of the vectorized covariance matrices to reduce the dimensionality to 3000 while approximately preserving pairwise profile distances.

Combining first and second-order moments. Because the statistical distributions of mean, MAD, and covariance profiles can be different in general (Supplementary Fig. 2), we combined them at the sample similarities level, rather than simply concatenating the profiles. We use Similarity Network Fusion (SNF)[6], which

operates on a graph representation of the dataset in each data type (in our case, three: medians, MADs, and covariances). A graph diffusion process is then used to combine the graph for each data type into the final network, which encodes the pairwise similarity values. SNF has shown great promise in fusing biological readouts when the number of samples is on the order of few hundreds[6]. The method is expected to be less effective when sample size falls below a threshold as local neighbors start to become less similar, violating the assumptions made in the method[18]. For this reason, we did not use the prior benchmark dataset in ref. [7].

Parameter settings. We used the SNFtool R package (Ver. 2.2.1) for data fusion to combine data types (median, MAD, and covariance profiles), and set the neighborhood size k = 7 in the similarity graph, gaussian weight function bandwidth μ = 0.5, and number of iterations T = 10 (for two data types) and T = 15 (for three data types) in SNF, which are typical choices for the algorithm. To avoid over-fitting, we did not test alternative values of these parameters. Prior to applying SNF, similarity matrices are z-scored based on median and MAD and then linearly

scaled to map 99.9th percentile to 0.999. This helps to make sure that the similarity values are on the same scale across data types.

We used 3000 sparse random projections with the density of $p = 0.1$ (the probability of an entry in the random projection matrix of being non-zero) to reduce the dimensionality of the covariance profiles in all the datasets. We observed reasonable consistency against randomness in the treatment correlation matrices when using around 3000 random projections (Supplementary Fig. 3). Pearson correlation of profiles is used to form similarity matrices, which are used as the inputs to SNF.

Factor analysis. We performed exploratory factor analysis using factanal function in R and a model is trained on a random sample of 25,000 control cells of each dataset. Fifty factors were chosen, which accounted for ~76% of total variance. Factor score coefficients (loadings) were computed using regression and Varimax rotation of orthogonally uncorrelated factors. Loadings of standardized control cells were used to estimate factor scores.

Principal component analysis. We performed principal component analysis using prcomp function in R (Ver. 3.3.3), and a model is trained on a random sample of 25,000 control cells of each dataset. Fifty principal components were chosen, which accounted for ~76% of total variance. Loadings of standardized control cells were used to compute principal component scores by multiplying it with the normalized data matrix.

Evaluation tasks. We evaluated different profiling strategies in this paper (Fig. 2) based on whether the most-similar treatment pairs (above a given cutoff) are enriched for having the same MOA/pathway annotation, after removing un-annotated compounds. To ensure that strong profile similarities are not driven by systematic effects that might make samples on the same plate look more similar to each other than to those in other plates, all same-plate pairs were excluded in this analysis. To avoid introducing additional parameters, we did not filter for per-turbations with strong phenotypes, although in practice, this could be a valuable step to further improve similarity predictions.

We rejected an alternative evaluation approach, accuracy in MOA/pathway classification[7], which only works well if MOAs are all well/equally represented in the dataset. The approach we took is better suited for the MOA class imbalance situation (as is the case for the datasets analyzed in this paper), as the enrichment is calculated based on a null distribution that tends to normalize MOA class sizes implicitly. Otherwise, treatments belonging to larger MOA classes tend to dominate the classification accuracy. Note that the chemical datasets we have presented reflect a huge variety of structures rather than a small number of compounds hand-picked to belong to particular classes; furthermore, annotations are sparse as many small molecules' mechanisms are unknown. As a result, while the number of samples are large, they are spread across many classes, resulting in many classes with very few samples. Although alternate metrics such as F-score/ precision/recall can help to mitigate class imbalances, they cannot overcome the small sizes for most classes in this dataset.

Enrichment score. We define enrichment score as the odds ratio in a one-sided Fisher's exact test, which tests whether having high profile similarity for a treat-ment pair is independent of the treatments sharing an MOA/pathway. To perform the test, we form the 2×2 contingency table by dividing treatment pairs into four categories, based on whether they have high profile correlations, determined by a specified threshold (in rows) and whether they share an MOA/pathway (in col-umns). The odds ratio is then defined as the ratio of elements in the first row divided by that of the second row in the contingency table. This roughly measures how likely it is to observe same MOA/pathway treatment pairs in highly correlated vs. non-highly correlated treatment pairs.

Data availability

We used three datasets to evaluate the profiling methods:

CDRPBIO-BBBC036-Bray: 2200 known bioactive compounds in U2OS cells. This dataset is the bioactive subset of a publicly available dataset[15]. Raw images are available at https://idr.openmicroscopy.org/webclient/?show = screen-1251.

Bioactives-BBBC022-Gustafsdottir: 1600 known bioactive compounds in U2OS cells. This is the image set BBBC022v1[14], available from the Broad Bioimage Benchmark Collection[7]. Raw images are available at https://data.broadinstitute.org/bbbc/BBBC022 and https://idr.openmicroscopy.org/webclient/?show = screen-1952. The compounds in this dataset have some overlap with CDRPBIO.

TA-ORF-BBBC037-Rohban: ~200 genes in various signaling pathways are over-expressed in U2OS cells[19]. Raw images are publicly available at https://idr. openmicroscopy.org/webclient/?show = screen-1751.

In all three datasets, around 1700 single-cell image-based readouts were obtained by running the Cell Painting assay[13] and an image processing pipeline in CellProfiler[20] software. The features are z-scored platewise in all datasets based on the negative controls.

Extracted image-based features are publicly available in the following s3 bucket s3:// cellpainting-datasets under folders corresponding to the respective names of the datasets.

Code availability

Source code, image processing pipelines, and gene/compound annotation data to reproduce these results are available (https://github.com/carpenterlab/2018_rohban_natcomm).

References

1. Janes, K. A. Single-cell states versus single-cell atlases—two classes of heterogeneity that differ in meaning and method. *Curr. Opin. Biotechnol.* **39**, 120–125 (2016).
2. Altschuler, S. J. & Wu, L. F. Cellular heterogeneity: do differences make a difference? *Cell* **141**, 559–563 (2010).
3. Pelkmans, L. Cell Biology. Using cell-to-cell variability—a new era in molecular biology. *Science* **336**, 425–426 (2012).
4. Deb, D. et al. Combination therapy targeting BCL6 and phospho-STAT3 defeats intratumor heterogeneity in a subset of non–small cell lung cancers. *Cancer Res.* **77**, 3070–3081 (2017).
5. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
6. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
7. Ljosa, V. et al. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.* **18**, 1321–1329 (2013).
8. Loo, L.-H. et al. An approach for extensibly profiling the molecular states of cellular subpopulations. *Nat. Methods* **6**, 759–765 (2009).
9. Fuchs, F. et al. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Mol. Syst. Biol.* **6**, 370 (2010).
10. Michael Ando, D., McLean, C. & Berndl, M. Improving phenotypic measurements in high-content imaging screens. bioRxiv 161422 preprint at https://www.biorxiv.org/content/10.1101/161422v1 (2017).
11. Hsu, D. & Kakade, S. M. Learning Mixtures of Spherical Gaussians: Moment Methods and Spectral Decompositions. In *Proc. 4th Conference on Innovations in Theoretical Computer Science* 11–20 (ACM, Berkeley, CA, USA 2013).
12. Li, P., Hastie, T. J. & Church, K. W. Very Sparse Random Projections. In *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 287–296 (ACM, Philadelphia, PA, USA 2006).
13. Bray, M.-A. et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
14. Gustafsdottir, S. M. et al. Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* **8**, e80999 (2013).
15. Bray, M.-A. et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6**, 1–5 (2017).
16. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
17. Slack, M. D., Martinez, E. D., Wu, L. F. & Altschuler, S. J. Characterizing heterogeneous cellular responses to perturbations. *Proc. Natl Acad. Sci. USA* **105**, 19306–19311 (2008).
18. Wang, B., Jiang, J., Wang, W., Zhou, Z. H. & Tu, Z. Unsupervised metric fusion by cross diffusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* 2997–3004 (IEEE, Providence, RI, USA 2012).
19. Rohban, M. H. et al. Systematic morphological profiling of human gene and allele function via Cell Painting. *elife* **6**, e24060 (2017).
20. Carpenter, A. E. et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).

Author contributions

M.H.R., H.S.A., S.S., and A.E.C. all contributed to drafting the manuscript. M.H.R. proposed the novel profiling method, gathered the ground truth annotation, and ana-lyzed the data with inputs from the other coauthors. H.S.A. contributed to the data analysis and benchmarked some previously proposed profiling methods, including Factor Analysis and Principal Component Analysis. S.S. helped to make the datasets publicly available.

Additional information

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.