Sampling strategies to capture single-cell heterogeneity

Satwik Rajaram1, Louise E Heinrich1, John D Gordan2,3, Jayant Avva4, Kathy M Bonness4, Agnieszka K Witkiewicz5, James S Malter6, Chloe E Atreya2,3, Robert S Warren3,7, Lani F Wu1,3 & Steven J Altschuler1,3

Advances in single-cell technologies have highlighted the prevalence and biological significance of cellular heterogeneity. A critical question researchers face is how to design experiments that faithfully capture the true range of heterogeneity from samples of cellular populations. Here we develop a data-driven approach, illustrated in the context of image data, that estimates the sampling depth required for prospective investigations of single-cell heterogeneity from an existing collection of samples.

Cellular populations often exhibit heterogeneity in morphology, signaling state and genotype, which can all play important roles in normal tissue function as well as in disease progression and drug resistance1,2. Experimental and analytical advances now allow researchers to probe and characterize individual cells3. As studies assess an increasing number of experimental conditions, a critical question researchers face is how to use sampling approaches to faithfully capture the true range of heterogeneity in each condition using the fewest cells possible (Fig. 1a). For example, by extracting small amounts of tissue (‘cores’; Fig. 1b) from multiple tissue specimens and placing these tissues on the same slide, tissue microarrays (TMAs)4 have helped standardize tissue analysis, reduce cost and substantially improve efficiency and throughput5. Similarly, high-throughput, multiwell-based screens capture each experimental condition with a small number of replicate wells. Here we ask what determines the number of samples required within or across conditions to capture heterogeneity in single-cell studies (Fig. 1a).

One might expect that the answer to this question depends on the degree of heterogeneity within a population (the central limit theorem, e.g., suggests that the number of samples required depends only on the s.d. of the phenotypic distribution). However, individual samples may not reflect the heterogeneity of the whole population. In practice, the number of samples required may be heavily influenced by nonhomogeneity of the sampling. Experimentally, cells are sampled not independently but in subsampled batches (e.g., cores or wells); cells within one batch may be strongly correlated and therefore not represent independent samplings of the population as a whole. For example, a single TMA core draws cells from the same tissue region. If tissue phenotypes vary on spatial scales larger than a core, then these sampled cells will be more phenotypically similar—and reveal far less about properties of the overall tissue—than a comparable number of cells sampled in a spatially random fashion (Fig. 1c, random versus one core). Similarly, in cell culture, cells within a well may be more similar to one another than cells from replicate wells6; or cells from areas within the same well may behave differently on account of local cellular density variations7. Thus, the number of samples required to capture population heterogeneity depends on the phenotypic variation observed across samples drawn from the same experimental condition.

The question of how many samples to take has been studied in various contexts, including for TMA cores8,9, needle biopsies10 and so on. In many of these contexts, a consensus on appropriate sampling has emerged (e.g., three 0.6-mm-diameter TMA cores11). Crucially, these studies focus on recovering population-averaged properties, such as the mean or number of biomarker ‘positive’ cells12-14 (although the ability to recapture correlations with biological and/or functional readouts is sometimes used8,9). However, agreement in such bulk metrics does not ensure similar phenotypic probability distributions (Fig. 1b bottom), and thus these approaches provide little guidance for studying heterogeneity. Here we develop an approach to estimate the number of samples required so that the probability distributions of cellular phenotypes in a sample (Fig. 1c, blue–yellow curves) match that of the whole population (Fig. 1c, green curve). This allows us not only to ensure that sampled phenotypes are present in similar proportions as they are in the whole population, but also to guarantee agreement in commonly used population quantifiers (including those of central tendency and variability), since these are derived from the probability distribution.

Our approach uses three steps to identify the number of samples needed to capture population heterogeneity (Fig. 1d). First, we designed a measure, referred to as a KS’ score, that behaves like the standard Kolmogorov–Smirnov (KS) test statistic15 for comparing phenotypic distributions of samples to the whole but has improved sensitivity (see Online Methods, Supplementary Fig. 1, and Supplementary Software) for detecting enrichment of extreme phenotype values. While we use the KS’ for our

1Department of Pharmaceutical Chemistry, University of California, San Francisco, California, USA. 2Department of Medicine, University of California, San Francisco, California, USA. 3Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, California, USA. 4Green Center for Systems Biology, University of Texas Southwestern Medical Center, Dallas, Texas, USA. 5Department of Pathology, University of Arizona, Tucson, Arizona, USA. 6Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas, USA. 7Department of Surgery, University of California, San Francisco, California, USA. Correspondence should be addressed to S.J.A. (steven.altschuler@ucsf.edu) or L.F.W. (lanifwu@ucsf.edu).

Received 21 March; Accepted 11 August; Published online 4 September 2017; DOI:10.1038/nmeth.4427
Figure 1 | Sampling strategy to capture single-cell heterogeneity. (a) Approach to determine how many samples (cores, replicates, draws) per condition are required for studies of heterogeneity. Top right, cellular phenotype distributions for different numbers of pooled samples or whole population (green); bottom right, heterogeneity captured with different numbers of samples. (b) Samples that recover population averages may not capture heterogeneity. Papillary adenocarcinoma tumor (green outline) shows extensive TTF1 staining heterogeneity that individual cores (circles) may not fully capture. Bottom, cores and whole tissue have similar mean TTF1 nuclear intensities (triangles), yet differ greatly in their phenotypic distributions. AU, arbitrary units (log scale). (c) Capture of tissue heterogeneity depends on the number of cells sampled and the nature of the sampling. TTF1 distributions are generated by repeated cell sampling; yellow and blue represent low and high KS’ dissimilarity, respectively, between core and whole-tumor (green curve) distributions. A single virtual core (~1,000 cells on average) is unreliable, but spatially random draws of the same number of cells (bottom plot) capture heterogeneity as reliably as do ten virtual cores. (d) Method for determining the number of samples needed to capture whole-population heterogeneity within a specified distribution tolerance (triangle 1) at a desired level of confidence (triangle 2). Users can select these levels (intersection of dotted lines). Top, comparison of whole tissue and pooled samples (point colors correspond to box outline colors in c based on distribution KS’ statistic versus median differences (see Online Methods). Bottom, confidence curves for achieving a desired KS’ tolerance as a function of sampling depth (number of cores) or type (core versus random).

demonstrations, the approach outlined here is compatible with other measures of distribution similarity, such as KS, Anderson–Darling or chi-square test statistics. Second, we developed theoretical bounds to relate KS’ scores to familiar population-averaged metrics such as the median or percentage of positive cells (see Online Methods; diagonal boundary in Fig. 1d, top). Third, we developed a framework to estimate how many samples will be required in future experiments to represent the phenotypic heterogeneity of the whole population (Fig. 1d, bottom).

In more detail, we start with a ‘representative’ whole population of cells and a phenotype of interest. We then randomly ‘draw’ a specified number of samples (e.g., TMA cores, wells, etc.) and compute the KS’ score to quantify the similarity of the phenotypic distribution of the pooled cellular population to that of the whole. To capture the inherent stochasticity of sampling, we repeat this process a large number of times (N = 1,000) to generate a distribution of KS’ scores for a given number of samples (Fig. 1d, top). As expected, if we repeat this procedure for increasingly larger numbers of samples (Fig. 1d colors from black to red), the distribution of KS’ scores will tend toward zero. This procedure provides a strategy to assign a confidence level (Fig. 1d, bottom, triangle 2) that the whole and pooled distributions from a given number of samples will be close (in the sense of the KS’ score or a population-averaged score; Fig. 1d, triangles 1 and 3). Our approach provides a general and quantitative starting point for assessing the tradeoff between extracting more samples and obtaining better estimates of whole-population heterogeneity.

We applied our approach to explore sampling strategies for studies of heterogeneity in microscopy. The ability to simultaneously capture spatial context, morphology and biomarker expression of a population makes microscopy an ideal platform for studying heterogeneity. As with any assay, observed heterogeneity depends on the choice of assay readouts as well as nonbiological (‘technical’) variability introduced by the assay itself. Accordingly, we here sought to provide guidance on how sampling is affected by image generation, biomarkers and cellular features.

First, we investigated the impact of the image generation processes—known to have strong quantitative effects on microscopy-based
measurements\textsuperscript{16,17}—in the context of designing TMAs to profile tissue heterogeneity. We defined the whole population as all cells present in the imaged tissue section, and we performed sampling by computationally extracting cells within randomly placed TMA-sized core regions (\(~0.6\text{ mm};\) see Online Methods). We applied our approach to a panel of 38 liver cancer specimens, of which 25 had two serial sections stained for the same antibody (YAP). Crucially, YAP staining was performed 5 months apart, and images were acquired using different microscopes. Despite these differences in the image generation process (Supplementary Fig. 2), we found broad agreement (Fig. 2a, roughly diagonal distribution of points) in the number of TMA cores needed to capture the heterogeneity in YAP nuclear intensity (on the subset of 25 cases that had two consecutive sections stained for YAP). These trends were robust for a wide range of KS’ threshold and confidence values (Supplementary Fig. 3). (We expect these trends to hold for other measures of distribution similarity, such as KS, though KS’ is particularly sensitive to identifying poorly represented tails in the distribution; Supplementary Fig. 4.) Thus, in the context of this data set, the number of samples determined by our method is largely a property of the specimen and is robust to differences in image generation.

Second, we investigated the impact of biomarker choice on TMA design. For biomarkers costained on the same tissue section, we compared the numbers of cores required to capture heterogeneity. YAP consistently requires more TMA cores than DAPI (Fig. 2b), LKB1 or \(\beta\)-catenin (Supplementary Fig. 5), and this perhaps reflects intrinsic differences in the regulation of these markers. However, our results also revealed significant diversity across the patient cohort for the same biomarker; two cores are adequate for some patients, while others required as many as ten. To understand the feasibility of using a single TMA design to study the heterogeneity of a large patient cohort, we examined how varying the number of cores affects the proportion of patients...
whose heterogeneity is well captured. Although all biomarkers require ~10 cores to capture the heterogeneity of every patient (Fig. 2c), the tradeoff between using fewer cores versus representing more patients well is biomarker dependent. For example, TMAs designed to sample phenotypic heterogeneity of LKB1 or β-catenin might poorly sample heterogeneity of YAP. Our sampling strategies can inform experimental design and biomarker selection of larger scale studies of heterogeneity.

Finally, we investigated the impact of cellular image feature choice in the context of designing high-throughput cell-culture-based experiments for profiling heterogeneity. We asked how many replicate wells should be performed per condition to sample heterogeneity. We made use of A549 cells containing three genetically encoded fluorescent markers to mark the nucleus, cytosol and a DNA-repair gene XRCC5

We analyzed seven 384-well imaging plates, each containing 28 replicate ‘control’ wells. For each cell, we extracted 215 single-cell features belonging to one of three classes (intensity, texture, and morphology; Supplementary Fig. 6). Measurements were pooled across the 28 replicate control wells to define 215 whole-population feature distributions. We then used our approach to estimate how many replicate wells are required to recover the whole-population distribution (Fig. 2d). The heterogeneity of some features, such as those relating to morphology, can be recovered with just one or two wells. In contrast, intensity features tend to be far more affected by well-to-well variation and require more wells to be sampled. Intensity features are a highly diverse set; features quantifying biomarker intensities near background levels are particularly hard to sample (e.g., cytoplasmic levels of a biomarker that is largely localized to the nucleus) and thus require large numbers of wells. Nevertheless, these analyses predict that three replicate wells are sufficient to capture heterogeneity for many, but not all, features. Thus, feature and biomarker selection both play roles in determining the number of samples required for studies of heterogeneity.

We have provided a general approach for estimating how many samples are required to represent distributions of heterogeneous phenotypes, a question typically considered only in the context of population-averaged quantifiers. Our work highlights the importance of within-sample correlations in answering this question. These correlations can be complex and experiment or specimen specific, which makes their quantitative effects difficult to predict from first principles. We established a data-driven framework that quantifies mismatch in heterogeneity between sample and whole population, relates this mismatch to effects on familiar population-average quantifiers, and allows the researcher to balance the tradeoff between number of samples and desired confidence for small heterogeneity mismatch.

Our approach makes use of representative single-cell data that capture the full range of phenotypic heterogeneity to estimate sampling depth of prospective studies. In many cases, such data may be readily available, though in other cases gathering such data may require upfront effort on account of specimen size

or rarity. Of course, it is not possible to determine via any analytical technique whether a set of specimens is truly representative. However, it is possible to ask whether fewer specimens would provide similar results; analysis of our data sets suggests that a smaller collection would have provided similar confidence estimates for numbers of samples taken (Supplementary Fig. 7). In practice, sample size may also be reduced by using assay-specific procedures to minimize nonbiological effects (e.g., image correction) or by generating samples more efficiently (e.g., using stereology

or H&E information). Beyond the contexts of immunofluorescence microscopy, which is highlighted in our case studies, our methodology also applies to other imaging modalities (e.g., IHC images, quantified by percent positive counts; Supplementary Fig. 8) to alternate sampling frameworks (e.g., choosing an appropriate placement of cores or designing ‘heterogeneity-TMAs’ for samples too large to fit on a slide) as well as to other single-cell assay technologies. Taken together, our methodology provides a rational approach to the design of experiments targeting phenotypic heterogeneity.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Calvert, T.D. Tisty and P.B. Stark for helpful discussions. This work was supported by the NCI K08CA175143 (C.E.A.), P01HL088594 (J.S.M.), a Conquer Cancer Foundation Young Investigator Award from the Scopus Foundation (J.D.G.), a gift from the Edmund Wattis Littlefield Foundation (R.S.W.), NSF PHY-1545915 (S.J.A.), Stand Up To Cancer (S.J.A.), NCI R01 CA133253 (S.J.A.), NCI R01 CA185404 (L.F.W.) and NCI R01 CA184984 (L.F.W.), and the Institute of Computational Health Sciences (ICHS) at UCSF (S.J.A. and L.F.W.).

AUTHOR CONTRIBUTIONS

S.R., S.J.A. and L.F.W. conceived of and designed the study. L.E.H., J.D.G., K.M.B. and R.S.W. performed the experiments, analyzed the data and contributed to sample collection. S.R., S.J.A. and L.F.W. performed the experiments and/or provided data. S.R., S.J.A., and L.F.W. developed the algorithms; and S.R. performed the analysis. R.S.W. and A.K.W. contributed samples. The manuscript was written by S.J.A., S.J.A. and L.F.W. with contributions from C.E.A. and J.S.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Almendro, V., Marusyk, A. & Polyak, K. Annu. Rev. Pathol. 8, 277–302 (2013).
2. Altschuler, S.J. & Wu, L.F. Cell 141, 559–563 (2010).
3. Yuan, G.C. et al. Genome Biol. 18, 84 (2017).
4. Wan, W.H., Fortuna, M.B. & Furmanski, P. J. Immunol. Methods 103, 121–129 (1987).
5. Camp, R.L., Neumeister, V. & Rimm, D.L. J. Clin. Oncol. 26, 5630–5637 (2008).
6. Bray, M.A. & Carpenter, A.E. in Assay Guidance Manual (eds. Sittampalam, G.S. et al.) 617–650 (Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2013).
7. Smidje, B. et al. Nature 461, 520–523 (2009).
8. Eckel-Passow, J.E. et al. Diagn. Pathol. 5, 48 (2010).
9. Tennstedt, P. et al. Int. J. Oncol. 40, 261–268 (2012).
10. Jiang, J., Colli, J. & El-Galley, R. J. Endourol. 24, 143–147 (2010).
11. Rimm, D.L. et al. J. Clin. Oncol. 29, 2282–2290 (2011).
12. Wampfler, J.A. J. Cancer Sci. Ther. 3, 120–124 (2011).
13. Goethals, L. et al. J. Pathol. 208, 607–614 (2006).
14. Khan, A.M. & Yuan, Y. J. Clin. Oncol. 29, 66–78 (1951).
15. North, A.J. J. Cell Biol. 172, 9–18 (2006).
16. Pawley, J. Biotechniques 26, 884–886, 888 (2000).
17. Kang, J. et al. Nat. Biotechnol. 34, 70–77 (2016).
18. Minner, S. et al. Mod. Pathol. 26, 106–116 (2013).
19. Weibel, E.R., Hsia, C.C. & Ochs, M. J. Appl. Physiol. 102, 459–467 (2007).
Supplementary Fig. 2a (Supplementary Fig. 8). A collection of fresh-frozen hepatocellular carcinoma (HCC, n = 38) samples was used in the present study. HCC specimens were collected at the University of California, San Francisco (San Francisco, California). Institutional Review Board approval was obtained from the UCSF IRB (Committee on Human Research), and informed consent was obtained from all subjects. See Gordan et al., manuscript in preparation, for further details of this sample set. Immunofluorescence staining and imaging were performed at the Gladstone Institutes Histology and Light Microscopy core using fluorescently labeled primary antibodies (all from Cell Signaling, Danvers, Massachusetts) to CTNNB1 (Mouse mAb L54E2, conjugated to Alexa Fluor 555, 1:200, cat. no. 5612) YAP (Rabbit mAb D8H1X, conjugated to Alexa Fluor 488, 1:200, cat. no. 14729) and LKB1 (Rabbit mAb, 1:250, cat. no. 13031, followed by goat anti-rabbit conjugated to Alexa Fluor 633, 1:200). Staining and imaging artifacts. Additionally, for the liver cancer data, staining artifacts and out-of-focus regions were automatically identified based on the lack of local intensity variations in these regions. For the adenocarcinoma (Fig. 1) and breast cancer (Supplementary Fig. 8) data sets, analysis was performed on the nonedge tumor regions (pathologist-identified green curves in Fig. 1 and Supplementary Fig. 8 respectively). For the liver cancer data (Fig. 2a–c), analysis was performed on the full tissue image, excluding regions that were at the edge or that displayed imaging artifacts.

Calculation of biomarker expression distributions. 1. Tissue. The starting point is a tissue image with nuclear-specific biomarker (DAPI levels for IF and deconvoluted haematoxylin intensity in IHC). A multilevel Otsu thresholding was applied to the nuclear image to identify nuclear pixels. A Laplacian of Gaussian filter, combined with an extended minima transform was applied to the nuclear biomarker image to identify well-separated nuclear centers. These centers served as seeds for a watershed transform that was used to partition the nuclear pixels into distinct, spatially disconnected nuclear regions. Regions too large to be nuclei were successively divided, while those too small to be nuclei were dropped. Nuclei in regions identified as exhibiting artificial staining or outside the region of interest (as identified above) were dropped. For each nucleus, the intensity of biomarkers was quantified by the mean intensity of the biomarker over pixels in that nucleus. Given a region of interest (e.g., whole tumor or pooled cores), the nuclear distribution of a biomarker was calculated by pooling the nuclear intensities across all nuclei whose centroids fell within that region.

2. Cell culture. Image background subtraction was performed using the ImageJ Rolling Ball Background Subtraction algorithm21. Cells in an image were automatically identified using our in-house watershed-based algorithm22, which identifies nuclear regions based on a nuclear marker and subsequently uses these identified nuclei as seeds for the identification of cell boundaries based on a cytoplasmic marker. For each cell, 215 different image features were calculated based on the intensities of the three biomarkers (H2B/cytoplasmic/XRCC5) in the pixels belonging to the cell. These features include (i) intensity features that are summaries of the intensities of the biomarkers in different cellular compartments (nucleus/cytoplasm/whole cell), (ii) texture features (Haralick/Zernike) that capture local variations of the three biomarkers (H2B/cytoplasmic/XRCC5) in the pixels belonging to the cell. These features include (i) intensity features that are summaries of the intensities of the biomarkers in different cellular compartments (nucleus/cytoplasm/whole cell), (ii) texture features (Haralick/Zernike) that capture local

Online Methods
Sample preparation, staining and imaging. 1. Adenocarcinoma tissue (Fig. 1). Formalin-fixed paraffin-embedded (FFPE) human non-small-cell lung cancer (papillary adenocarcinoma) tissue blocks were purchased from ILSBio, re-embedded and sliced in 5 µm sections by the Molecular Pathology Core Facility at UTSW. Sudan Black B blocking was used to reduce autofluorescence. TTF1 staining was performed using polyclonal rabbit antibodies (cat. no. sc-13040, dilution 1:100, Santa Cruz Biotech). Digital images of stained tissue sections were obtained using a ScanScope Digital slide scanner at 20× (Aperio ePathology, Leica Biosystems).

2. Liver cancer tissue (Fig. 2a–c). A collection of fresh-frozen hepatocellular carcinoma (HCC, n = 38) samples was used in the present study. HCC specimens were collected at the University of California, San Francisco (San Francisco, California). Institutional Review Board approval was obtained from the UCSF IRB (Committee on Human Research), and informed consent was obtained from all subjects. See Gordan et al., manuscript in preparation, for further details of this sample set. Immunofluorescence staining and imaging were performed at the Gladstone Institutes Histology and Light Microscopy core using fluorescently labeled primary antibodies (all from Cell Signaling, Danvers, Massachusetts) to CTNNB1 (Mouse mAb L54E2, conjugated to Alexa Fluor 555, 1:200, cat. no. 5612) YAP (Rabbit mAb D8H1X, conjugated to Alexa Fluor 488, 1:200, cat. no. 14729) and LKB1 (Rabbit mAb, 1:250, cat. no. 13031, followed by goat anti-rabbit conjugated to Alexa Fluor 633, 1:200). Staining and imaging was controlled by IN Cell Analyzer software (GE). One image was acquired per well. Images with obvious anomalies (e.g., out of focus, abnormal fluorescent patterns caused by dust, scratches on the plate) were discarded after manual inspection.

4. Immunohistochemistry (IHC) sample preparation and staining (Supplementary Fig. 8). An FFPE human breast cancer specimen was obtained from UT Southwestern Shared Tissue resource. Ki67 immunostaining was performed using primary monoclonal rabbit anti- Ki67 antibody (cat. no. 790–4286, clone 30-9, dilution 1:100, Ventana Medical Systems) on an automated BenchMark stainer (Ventana Medical Systems). Digital images of Ki67 stained tissue sections were obtained using a ScanScope Digital slide scanner at 20× (Aperio ePathology, Leica Biosystems).

Identification of regions of interest in tissue. We developed an image analysis approach to identify and exclude cells at the edge of the tissue (as these potentially display edge-staining artifacts). Additionally, for the liver cancer data, staining artifacts and out-of-focus regions were automatically identified based on the lack of local intensity variations in these regions. For the adenocarcinoma (Fig. 1) and breast cancer (Supplementary Fig. 8) data sets, analysis was performed on the nonedge tumor regions (pathologist-identified green curves in Fig. 1 and Supplementary Fig. 8 respectively). For the liver cancer data (Fig. 2a–c), analysis was performed on the full tissue image, excluding regions that were at the edge or that displayed imaging artifacts.

Calculation of biomarker expression distributions. 1. Tissue. The starting point is a tissue image with nuclear-specific biomarker (DPI levels for IF and deconvoluted haematoxylin intensity in IHC). A multilevel Otsu thresholding was applied to the nuclear image to identify nuclear pixels. A Laplacian of Gaussian filter, combined with an extended minima transform was applied to the nuclear biomarker image to identify well-separated nuclear centers. These centers served as seeds for a watershed transform that was used to partition the nuclear pixels into distinct, spatially disconnected nuclear regions. Regions too large to be nuclei were successively divided, while those too small to be nuclei were dropped. Nuclei in regions identified as exhibiting artificial staining or outside the region of interest (as identified above) were dropped. For each nucleus, the intensity of biomarkers was quantified by the mean intensity of the biomarker over pixels in that nucleus. Given a region of interest (e.g., whole tumor or pooled cores), the nuclear distribution of a biomarker was calculated by pooling the nuclear intensities across all nuclei whose centroids fell within that region.

2. Cell culture. Image background subtraction was performed using the ImageJ Rolling Ball Background Subtraction algorithm21. Cells in an image were automatically identified using our in-house watershed-based algorithm22, which identifies nuclear regions based on a nuclear marker and subsequently uses these identified nuclei as seeds for the identification of cell boundaries based on a cytoplasmic marker. For each cell, 215 different image features were calculated based on the intensities of the three biomarkers (H2B/cytoplasmic/XRCC5) in the pixels belonging to the cell. These features include (i) intensity features that are summaries of the intensities of the biomarkers in different cellular compartments (nucleus/cytoplasm/whole cell), (ii) texture features (Haralick/Zernike) that capture local
biomarker intensity variations, or (iii) features that describe cellular morphology. See Supplementary Figure 6 for a more detailed summary. Intensity features that included the cytoplasmic intensity of the nuclear-localized biomarkers H2B and XRCC5 were classified as low contrast. The whole population distribution for a feature was calculated by pooling feature values for cells belonging to all 28 replicate wells in a plate, while sample distributions were generated by pooling (cells from) a subset of wells together.

Virtual sampling of cores in tissue. For each virtual sampling, N nonoverlapping cores were placed randomly on the image such that (i) the entire core (circle of 0.6 mm diameter, between 600 and 1,000 pixels in the 20× images analyzed here) was within the tissue; (ii) the core was centered within the ‘good’ tissue area (which was identified as described above); and (iii) more than 70% of the core area was covered by cells of interest. Cells whose centroids were within 0.3 mm of the core center were considered as belonging to the core. Biomarker intensities of cells from all cores were pooled together to construct the core intensity distribution for each virtual sampling.

KS’, a new measure to compare phenotypic heterogeneity. A widely used measure to compare distributions is the Kolmogorov–Smirnov (KS) statistic13, which measures the maximum difference between two CDFs (Supplementary Fig. 1a, bottom, double-sided arrows). More precisely, if (I) and (W) are the CDFs for the core and whole tissue, the corresponding KS score is given by  max{|(I) − (W)|}. The KS statistic has several virtues, one of which is that the KS value is insensitive to the nature of the distributions (i.e., normal, Poisson and so on), yet it can still detect changes to both the location and shape of the distributions.

However, a disadvantage of the KS statistic is that it tends to be less sensitive to changes near the tails of distributions23. This can be undesirable in contexts where subpopulations of the most or least stained cells have important biological meaning. The reason for the loss of sensitivity is that the KS statistic measures the largest difference between CDFs regardless of where this difference occurs. Yet, a large difference in CDFs is far rarer at the tails (differences between sample and whole CDFs tend to zero at the tails; Supplementary Fig. 1c, top) and should be considered more significant than when the same difference occurs away from the tails. For our purposes, it is desirable to use a test that is more sensitive to changes across the whole range of intensities, including the tails.

We chose to modify the KS statistic to increase sensitivity at the intensity tails by normalizing differences of CDFs by the magnitude of expected deviations when sampling from the whole-tissue distribution W. Observed deviations near the tail would then become more apparent when divided by a small expected deviation, working in the same way as a z-score. Towards this end, we started with an explicit formula computed by Anderson and Darling (Supplementary Fig. 1c, bottom), |(I) − (W)|/√(W(I − W(I)), which computes the expected standard deviation at each intensity value I between the CDFs of the whole distribution W. Our desired normalized CDF difference between a core (I) and whole tissue (W(I)) is then given by |(I) − (W(I))|/√(W(I − W(I))).

Our modified KS statistic, which we refer to as the KS’ statistic, is then simply \( KS' = \max_I \frac{0.5 \times |(I) - (W(I))|}{\sqrt{W(I)(1 - W(I))}}. \) Thus, the KS’ allows us to compare the distributions of cores to that of whole tissue with particular sensitivity to extreme phenotypes. We note that the KS’ was designed as a measure of distribution similarity rather than a test statistic in the current work.

A particular strength of the KS’ is its ability to theoretically bound differences (between whole population and samples) in various statistical quantifiers of the distribution. For example, if \( I_{50} \) is the median of the whole-specimen distribution (i.e., \( W(I_{50}) = 0.5 \)) then we have the relationship:

\[
KS'(W,C) = \max_I \left| \frac{0.5 \times |(I) - (W(I))|}{\sqrt{W(I)(1 - W(I))}} \right| \geq \left| \frac{0.5 \times |(I_{50}) - W(I_{50})|}{\sqrt{W(I_{50})(1 - W(I_{50}))}} \right| = |(I_{50}) - 0.5|
\]

Since \( |(I_{50})| \) is the fraction of the population less than the whole-specimen median, it would be equal to 0.5 if the whole and sample medians matched. Thus, if the KS’ between the whole and its sample distributions is 0.1, the whole median must lie between the 40th and 60th percentile of the sample. (We note that in cases where (I) is discontinuous exactly at \( I_{50} \), e.g. integrates one or more events at \( I_{50} \), then we interpret the right-hand side of the inequality to be the smallest absolute difference of the range of values for the jump to 0.5.)

Similar bounds can be established for whole versus sample mean or percentage positive cells (Supplementary Fig. 8). For example, if the percentage of positive cells \( p \) is defined as the fraction of cells above a threshold intensity \( I_t \), then the error between whole \( p_W = 1 - W(I_t) \) and sample \( p_C = 1 - C(I_t) \) is bounded as:

\[
|p_W - p_C| = |W(I_t) - C(I_t)| = \left| \frac{W(I_t) - C(I_t)}{\sqrt{W(I_t)(1 - W(I_t))}} \right| \times \sqrt{W(I_t)(1 - W(I_t))} \leq \sup_I \left| \frac{W(I) - C(I)}{\sqrt{W(I)(1 - W(I))}} \right| \times \sqrt{W(I_t)(1 - W(I_t))} \leq KS' (W,C) \times \sqrt{W(I_t)(1 - W(I_t))}.
\]

Determining the number of samples for a single whole specimen. The heterogeneity observed in a sample can be influenced by inhomogeneity introduced by sampling (i.e., the fact that not all cells within the whole have the same probability of belonging to the same sample). We assume here that a computational means has been established to simulate experimental sampling; in our case, this was done through construction of virtual cores of the same size as a true TMA core for tissue or selecting cells by well in cell culture.

1. For a given number \( N \) of samples, extract the single-cell data and calculate the corresponding KS’ score by comparing the distribution of \( N \) pooled samples to the whole.

2. To model the inherent stochasticity of sampling, repeat step 1 multiple times (we used 1,000 repeats for our results) to construct a distribution (quantified by its CDF) of KS’ scores for \( N \) samples.

3. Repeat steps 1 and 2 while varying \( N \) over the experimentally reasonable range of samples. This will give us CDFs

DOI: 10.1038/nmeth.4427
for the KS' for different values of \( N \) (this is essentially Fig. 1d bottom, with the KS' scores shifting to the left as more samples are extracted).

4. Determine the maximum value of KS' for sampling to be considered 'good'. This maximum value can be calibrated using the inequalities relating the KS' to bounds on the median, mean, percentage positive, etc. For example, a KS' < 0.1 means the median of the whole population will lie between the 40th and 60th percentile of the sample.

5. Determine the desired confidence level that sampling needs to be 'good' (e.g., at least 80% of the times we generate samples we want the KS' < 0.1).

6. By comparing the CDFs generated in step 3, find the minimum value of \( N \) such that the CDF at the KS' tolerance exceeds the desired confidence level.

This procedure selects the minimum number of samples required to provide a desired level of confidence that the difference between sample and whole distributions is within a specified KS' tolerance.

**Design of experimental sampling using a pilot panel.** The scenario we envisioned is to design a large experiment profiling heterogeneity based on whole-specimen data observed within a pilot set of specimens. In the paper (Fig. 2c) we illustrate the design of a TMA (potentially consisting of hundreds of patient tissue specimens) based on a pilot set of ~30 whole-tissue liver cancer specimens. As with any statistical estimator, we treat the pilot set as representative of data for the large-scale experiment. The first step is to calculate the specimen’s KS’ CDF curves for each of the \( N_{\text{PS}} \) pilot specimens across varying numbers of samples (as outlined above). For a pilot specimen \( i \), let \( C_i(K,n) \) denote the probability of \( n \) pooled samples yielding a distribution that differs from the whole by KS’ score less than \( K \). Then, as we average across specimens, the expected fraction of specimens for which of \( n \) pooled samples capture the corresponding whole-population distribution well (as defined by a KS’ score less than \( K \)) is given by:

\[
C(K,n) = \frac{\sum_{i=1}^{N_{\text{PS}}} C_i(K,n)}{N_{\text{PS}}}
\]

This process may be repeated for different biomarkers, etc., as demonstrated in Figure 2c.

**Code availability.** The MATLAB code used to generate the main figures is provided as supporting code. An R implementation of the KS’ is also included. The latest version of the code will be available at https://github.com/AltschulerWu-Lab/SamplingForHeterogeneity.

**Data availability statement.** The raw data used to generate Figure 1 are available at http://dx.doi.org/10.5281/zenodo.841835. The raw data that support the findings of Figure 2 are available from the corresponding author upon reasonable request. The processed source data used to generate the main figures are provided along with the manuscript as supporting information.

A Life Sciences Reporting Summary for this paper is available.

21. Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. Nat. Methods 9, 671–675 (2012).
22. Loo, L.H., Wu, L.F. & Altschuler, S.J. Nat. Methods 4, 445–453 (2007).
23. Mason, D.M. & Schuenemeyer, J.H. Ann. Stat. 11, 933–946 (1983).
Experimental design

1. Sample size
   Describe how sample size was determined.
   Our paper explores the consequences of varying sample size for quantifying heterogeneity, thus sample size is a variable for our study.

2. Data exclusions
   Describe any data exclusions.
   Fig 2A. One point is outside the plotted range, but is provided in the source data. A section from tumor-adjacent liver was inadvertently used and was omitted from the final analysis of Fig 2A-C which focus on tumors.

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   Fig 2A. We compared serial section tumor samples imaged from two different microscopes. Fig 2D was obtained from 7 replicated plates.

5. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   N/A

6. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

- [x] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- [x] A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- [x] A statement indicating how many times each experiment was replicated
- [x] The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- [x] A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- [x] The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
- [x] A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- [x] Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.
Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Custom code written and tested in Matlab v2016b. An R function is provided for computing the KS' value.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

N/A

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

TTF1 polyclonal rabbit antibodies (H-190) (Cat.# sc-13040, dilution 1:100 Santa Cruz Biotech) followed by anti-rabbit Alexa Fluor® 568 (1:1000).
Santa Cruz recommend this antibody for immunohistochemistry studies, observing nuclear localization in formalin fixed paraffin embedded (FFPE) human lung tissue.

Ki67 rabbit monoclonal antibody (IgG) (Cat.# 790-4286, clone 30-9, dilution 1:100, Ventana Medical Systems)
Ventana recommend this antibody for immunohistochemistry studies in FFPE tissue - observing nuclear localization in human tonsil tissue.

LKB1 rabbit monoclonal antibody (1:250, Cell Signaling Technology, Cat.#13031) followed by goat anti-rabbit conjugated to Alexa Fluor® 633(1:200, Cell Signaling Technology)
Cell Signaling Technology recommend this antibody for immunohistochemistry in FFPE tissue, observing cytoplasmic localization in lung, ovarian and prostate carcinoma samples, and in LKB1 positive and LKB1 negative FFPE cell pellets.

Fluorescent conjugated antibodies;

CTNNB1 (L54E2) mouse monoclonal antibody (conjugated to Alexa Fluor® 555, 1:200, Cell Signaling Technology, Cat.# 5612)
Cell Signaling Technology recommend this antibody for immunofluorescence studies, observing presence or absence of cytoplasmic and cell membrane localization in high- and low- CTNNB1 expressing human cell lines.

YAP rabbit monoclonal antibody (D8H1X, conjugated to Alexa Fluor® 488, 1:200, Cat. #14729)
Cell Signaling Technology recommend this antibody for immunofluorescence studies, observing presence or absence of nuclear signal in high- and low- YAP1 expressing human cell lines, in a cell-density-dependent manner.
10. Eukaryotic cell lines
   a. State the source of each eukaryotic cell line used. The human lung epithelial carcinoma cell line A549 (ATCC® CCL-185™) was obtained from ATCC in 2013. A549-XRCC5 consists of a clonal subpopulation of A549 #CCL-185 that express H2B-CFP, cytoplasmic mCherry, and XRCC5-YFP, generated as described in Kang et al, 2015 (doi:10.1038/nbt.3419).
   b. Describe the method of cell line authentication used. A549 has not been authenticated.
   c. Report whether the cell lines were tested for mycoplasma contamination. Initial stocks of A549-XRCC5 were expanded over 2 passages and frozen into 50 aliquots used to generate data in Figure 2D. A sample of this passage tested negative for mycoplasma by PCR upon freezing (August 2016), and at the time of publication (August 2017).
   d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. A549 does not appear in the ICLAC table (V8) of misidentified cell lines published 1/12/2016.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

1. Description of research animals
   Provide details on animals and/or animal-derived materials used in the study.
   N/A

Policy information about studies involving human research participants

12. Description of human research participants
   Describe the covariate-relevant population characteristics of the human research participants.
   N/A