SUPPLEMENTARY INFORMATION

Selection of clonal cell lines. Clones were made from H460 (Carney *et al*, 1985) and HBEC (Vaughan *et al*, 2006) cell lines (gifts of Dr. John Minna at UT Southwestern Medical Center). Cells were plated in serial dilution at ~0.25 cells per well in a 96-well plate and inspected manually to ensure growth occurred from a single site. 49 H460 and 75 HBEC clones were randomly chosen along with the parental population for the experiments.

Selection of NCI-60 cell lines (Suppl. Table 2). The 5 most sensitive and 5 most resistant cell lines to paclitaxel within the NCI-60 panel were selected using the latest (August 2008) publicly available GI50 values (downloaded October 6, 2008 from the NCI depository website (NCI/NIH); values based on the highest repeat numbers were used (in our case 29)). The leukemia cell lines were excluded from the initial selection process as they were mostly non-adherent and thus unsuitable for our imaging assays. The identities of all selected cell lines were checked by DNA fingerprinting using Powerplex sequencing (performed at the UT Southwestern Medical Center Core Facilities). The CAKI-1 cell line was discarded since its identity could not be confirmed by fingerprinting, leaving 9 final NCI cell lines for our study.

Cell culture and drug sensitivity assay. Cell lines were grown in RPMI 1640 medium supplemented with either 10% (H460 and HBEC clones) or 5% (NCI-60 cell lines) fetal bovine serum (FBS), 2mM L-glutamine and 1x penicillin-streptomycin in a 37°C / 5% CO_2 incubator. Cells were plated at a density of 10,000 cells per well on Nunc 96-well glass-bottomed imaging plates in triplicate wells, and incubated overnight (16 hours) to allow cells to adhere (Suppl. Table 3). For the drug-sensitivity assays, cells were treated with paclitaxel (10nM) or doxorubicin (1μM) for 48 hours. Cells were fixed with 4% paraformaldehyde (PFA) in PBS for 5 minutes. Each collection of H460, HBEC, or NCI

cell line experiments was performed on a single day to eliminate issues of day-to-day variability (per time point).

Marker selection and immunostaining. Six marker sets (MS1-MS6) were selected for the heterogeneity assays, and one apoptosis marker set (Annexin-V, cleaved Caspase3 and PARP) was used for the drug assay (Suppl. Table 1). Hoechst 33342 was used in all marker sets to identify nuclear regions. Cells were fixed with 4% paraformaldehyde for 5 minutes, permeabilized with 0.2% Triton X-100 solution in TBS for three minutes, washed with TBST, blocked with 5% BSA solution in TBST at room temperature for two hours, and washed with TBST three times. 5% BSA in TBST was used for primary and secondary antibody dilutions. Finally, plates stained with antibodies were incubated in the dark at room temperature for two hours and then washed again with TBST three times. After the final washing step, 100µl of TBST containing 0.1% sodium azide was added to each well.

Image acquisition and processing. All fluorescence images were acquired using a TE-2000 E2 epifluorescence microscope (Nikon) equipped with integrated Perfect-Focus System (PFS), Nikon Plan Apochromat 20x objective lens and Photometrics CoolSNAP HQ camera using 1x1 camera binning. Image acquisition was controlled by Metamorph software (Universal Imaging). Image background correction was done using the National Institute of Health ImageJ rolling-ball background subtraction software (Rasband *et al*, 1997-2009). Cellular regions were determined using a watershed-based segmentation algorithm (Loo *et al*, 2007) which first retrieves nuclear regions using DNA staining then combines multiple cytosolic region markers to identify cellular boundaries. Images were visually inspected, and images with severe focus, staining, or cell-segmentation artifacts were discarded. We identified ~4,000 cellular regions per marker set and clone after applying automated cell segmentation to our image data.

Image quality control. We manually inspected all fluorescence images (~69,000) and discarded those presenting obvious anomalies (e.g. focus issues and abnormal fluorescence staining). To account for systematic overlap between adjacent 4x4 images, cells situated within the right-hand side margin (10% of the image width) and bottom margin (10% of image height) of those image frames sharing common margins with neighbour frames were removed from our dataset. Finally, images with poorly segmented cells were re-segmented with manually optimized segmentation parameters.

Measure of growth rate, total cellular count, and local cell density (Suppl. Fig. 2). The growth rate of a clone was computed based on the fold increase of cell density five days after plate seeding (no drug). Total observed cell count was computed from the total number of cells detected in the images after 16 hours. Local cell density of a clone, referring to the "clumpiness" of cells after plate seeding, was estimated for each cell by counting the number of neighboring cells whose geometric center is located within r pixels of its own center. The median number of such neighbors of all cells in the clone was taken as the measure of local cell density. We used r = 100 in our analysis (cell size is 60-70 pixels long).

Fluorescence intensity normalization across plates. To account for plate-to-plate fluctuation of fluorescence intensity, greyscale values of each image's pixels were normalized against the parental clone within each plate. For a given plate p and fluorescence channel m, the distribution of median intensity per cellular region was collected across all replicate wells of the parental clone. The median value of this intensity distribution, defined as $J_m^{(p)}$, was used to transform the pixel intensity $I_m^{(p)}$ of all images from channel m in plate p to a new value $I'^{(p)}_m$ by a simple rescaling operation using a fixed reference parameter I_0: $I_m^{(p)} \rightarrow I_m'^{(p)} = I_m^{(p)} \cdot \frac{I_0}{J_m^{(p)}}$. This normalization procedure resets the median pixel intensity per cellular region in the parental clone to the same value I_0 across

all plates. In our analysis, the parameter I_0 was set to 500 (the dynamic range of pixel value is between 0 and 4,095). For the NCI-60 dataset, due to weaker staining of several fluorescence markers, we used the 75th percentile pixel intensity as $J_m^{(p)}$ instead of the median value to ensure that all pixel values remained within a reasonable numerical range after rescaling.

Feature reduction by PCA transformation. We applied the pixel intensity-based features developed in (Slack *et al*, 2008) to capture cellular signaling phenotypes. In our analysis, we used a grid with edges at 0, 1/3, 2/3, 1, 4/3, 5/3, 2, 7/3, 8/3 and ∞ (infinity) to discretize the intensity ratios for each pixel. We simplified the feature computation by keeping only one copy of inverse intensity ratios since they provide redundant information. For instance, in Marker Set-1 we used the ratios DNA/pPTEN, pSTAT3/pPTEN and DNA/pSTAT3 (without using pPTEN/DNA, pPTEN/pSTAT3 and pSTAT3/DNA) when evaluating the features. To reduce computational workload (there were approximately 200,000 cells per marker set), 10% of the cells in the non drug-treated dataset were randomly sampled from all wells (with replacement) for each marker set *m* to form an initial sample set $S^{(m)}$. The weights assigned to individual wells were determined proportionally to the total number of cells detected per well. Feature data associated to these randomly sampled cells were used to normalize feature values and create mathematical models to characterize phenotypic heterogeneity as described below. The feature values were transformed to *z*-scores, each with respect to its mean and standard deviation computed from the sample set $S^{(m)}$, before reduction to their most prominent principal components (PCs). The feature dimension after principal component analysis (PCA)-based reduction was 9 for Marker Set-1, 6 for Marker Set-2, 2 for Marker Set-3 and 5 for Marker Set-4. The choice of dimension was made for each marker set by computing an eigenvalue noise threshold for the covariance matrix of feature data. The threshold was determined by randomly scrambling the order of feature dimensions for each sampled cell, and computing the eigenvalues of the resulting

(randomized feature) covariance matrix. The noise threshold was chosen to be double the largest such eigenvalue. Any dimensions whose eigenvalues from the (non-randomized) feature covariance matrix did not meet this threshold were discarded. For convenience, we denote by $F^{(m)}$ the PCA-reduced feature data of the cells selected in $S^{(m)}$ for marker set *m*.

Reference models of heterogeneity. All subpopulation reference model computations used the sampled PCA-reduced feature data $F^{(m)}$ from all non drug-treated H460 clonal populations and their parent population as described above (Slack *et al*, 2008). Subpopulation reference models were derived using Gaussian Mixture Models (GMM). The GMM parameters were fitted based on the Expectation-Maximization (EM) algorithm (Dempster *et al*, 1977). For each model, EM clustering was executed ten times, starting from a *K*-means clustering (Kaufman and Rousseeuw, 1990) using randomly chosen means. The final clustering with the best log likelihood value was chosen as the subpopulation reference model. Each run was attempted up to five times with new initial conditions until convergence was reached. Bayesian information-theoretical criterion (*BIC*) (Schwarz, 1978) and the Gap statistics (*Gap*) (Hastie *et al*, 2001) were used to evaluate the optimal number of subpopulations (*K*). The BIC seeks to maximize the (log) likelihood of the observed data samples given the model parameters while minimizing the complexity of the model to avoid overfitting. On the other hand, the Gap statistics determined the optimal number of clusters (subpopulations) by comparing the change in dispersion within clusters to that expected under a uniform null distribution. Due to large sample size, the *BIC* tends to continue growing as *K* increases. In these situations the best choice typically occurs when the *BIC*-versus-*K* curve encounters an inflexion. We tested models with different values of *K* ranging from 3 to 20 and found that models with *K* between 3 and 7 reasonably captured the overall cellular heterogeneity using both *BIC* and *Gap* criteria (Suppl. Fig. 4). In our analysis, a GMM with $K = 5$ was used for each of the four marker sets unless stated otherwise in a text or figure.

Subpopulation profiles and enrichment profiles. For each replicate well associated to a given clone, PCA-reduced feature data from 1,000 randomly selected cells (with replacement) were used to determine a probability distribution (profile) of subpopulation assignment. The computed GMM was used to assign to each cell in a population a (posterior) probability vector of belonging to each subpopulation. The averaged probability vectors over all cells within a population produced a subpopulation profile. The weighted average of such profiles across replicate wells, based on the relative total number of cells per well, was generated 1,000 times by repeated cell sampling to yield an average subpopulation profile for each clone per marker set. To assess the variation in changes to subpopulation profiles, we transformed the profile to reflect the log-fold change of subpopulations as follow: we define

- $[p_1^{(C)} \ldots p_K^{(C)}]$ as the subpopulation profile of clone *C*,
- $[p_1^{(P)} \ldots p_K^{(P)}]$ as the subpopulation profile of the parent *P* on the same imaging plate.

Then the component-wise log ratio vector

$$[ln(p_1^{(C)}/p_1^{(P)}) \ldots ln(p_K^{(C)}/p_K^{(P)})]$$

is defined as the subpopulation enrichment of clone *C* versus parent *P*. In practice, the enrichment value was truncated between -4 and 4 to avoid numerical outliers (Fig. 2).

Hierarchical clustering and ordering of clones in the hierarchical tree. Average linkage hierarchical clustering of subpopulation profiles and enrichment profiles was performed using Matlab built-in functions. We used the symmetrized Kullback-Leibler dissimilarity measure (Kullback and Leibler, 1951) or the Euclidean distance as distance measure to cluster clones when they were represented by their subpopulation profiles or enrichment profiles, respectively. Both representations yield comparable results in our study. However, the enrichment profiles have the ability to amplify small differences in the composition of

subpopulations, which could be beneficial when dealing with small subpopulations. Both normalized and non-normalized profiles yield comparable conclusions in our study. For visualization, non-normalized profiles are useful for seeing absolute composition of subpopulations, while normalized profiles are helpful for comparing subpopulation composition across multiple populations and lend themselves better to standard heatmap representations. In order to illustrate separation between paclitaxel sensitive and resistant clones within the hierarchical clustering, we recursively pivoted each branch of the tree from the top to the bottom and reordered its two child nodes so that the average paclitaxel sensitivity in the sub-tree spanned by its left-hand side child was always smaller than or equal to the one in the sub-tree spanned by its right-hand side child. The pivoting affected the linear ordering of clones but preserved the original hierarchical clustering.

Subpopulation profiles of HBEC clones and selected NCI-60 cell lines. We computed the subpopulation profiles for any additional cell populations based on the original H460 model of heterogeneity. The fluorescence intensities in the background subtracted images were first normalized in the same way as described above. Cellular feature vectors were also normalized and reduced using the same parameters as the ones associated to the H460 model. The subpopulation profiles were then computed as described above.

Similarity comparison of subpopulation profiles and multidimensional scaling plots. Similarity of phenotypic heterogeneity between two clones (or cell lines) was computed using the symmetrized Kullback-Leibler dissimilarity between their subpopulation profiles (Kullback *et al*, 1951). For a collection of clones (or cell lines), multidimensional scaling (MDS) was performed on the pairwise dissimilarity matrix associated to their K-dimensional subpopulation profiles (Borg and Groenen, 1997) to yield a configuration of points in an s-dimensional space (typically s is much smaller than K) such that the Euclidean distances between these points approximate the degree of subpopulation profile similarity among the clones (or cell lines). MDS was performed with $s = 2$ using Matlab

software (version 7.4.0). In principle, the resulting MDS plots places clones with similar subpopulation profiles closer together, and clones with dissimilar subpopulation profiles further apart.

Measuring and comparing the spread of phenotypic heterogeneity. To estimate the spread of phenotypic heterogeneity across the 49 H460 clones and across the 75 HBEC clones, we pooled both sets of data to build an overall GMM model, and computed subpopulations profiles for all clones with the number of clusters K varying from 3 to 14. The 5th and 95th percentile values of the median pairwise KL dissimilarity measure served as the confidence interval for the spread of heterogeneity. Next, we randomly sampled (without replacement) 40 clones from each collection of clone populations and computed their median pairwise Kullback-Leibler (KL) dissimilarity measure. We repeated the random sampling 1,000 times to obtain an empirical distribution of the median pairwise KL dissimilarity both within our H460 clones ($\{d_{H460}\}$) and within our HBEC clones ($\{d_{HBEC}\}$). The median values of the median pairwise KL dissimilarity measure were used to quantify the spread of phenotypic heterogeneity within each collection of clones across varying numbers of subpopulations. A one-sided two-sample Kolmogorov-Smirnov test was applied to the two empirical distributions $\{d_{H460}\}$, $\{d_{HBEC}\}$ to assess whether the H460 cell line exhibited significantly larger spread of phenotypic heterogeneity than the HBEC cell line (Suppl. Table 4).

Drug sensitivity assignment. Drug sensitivity of each clone was calculated based on the log ratio between the numbers of non-apoptotic cells in the drug-treated case over non drug-treated case compared to the parental clone. For each given clone C, the relative drug sensitivity of clone C versus parent P was defined by the log ratio

$$DS(C) = -\ln\left(\frac{(\frac{n_C^+ - a_C^+}{n_C^- - a_C^-})}{(\frac{n_P^+ - a_P^+}{n_P^- - a_P^-})}\right)$$

where:

n_C^+, n_P^+: number of drug-treated cells observed in *C* and *P;*

a_C^+, a_P^+: number of apoptotic, drug-treated cells observed in *C* and *P;*

n_C^-, n_P^-: number of only DMSO-treated cells observed in *C* and *P;*

a_C^-, a_P^-: number of apoptotic, only DMSO-treated cells observed in *C* and *P*.

A clone *C* is considered relatively sensitive (S) if the index *DS(C)* is positive and relatively resistant (R) if the index is negative. In the H460 drug sensitivity assay, we encountered an image focus issue on one plate and did not have the values of n_P^+, a_P^+ *and* n_C^+ for clones 33 and 35. Hence clones 33 and 35 were discarded from all analysis involving drug sensitivity. To estimate the drug sensitivity of five other clones on that plate (32, 34, 36, 41, 49) for which only the variables n_P^+ and a_P^+ were missing, we recovered n_P^+ by manual counting using the brightfield images (which were in reasonably good focus) and estimated a_P^+ using the average apoptotic rate of the parental clone from the six other plates, defined as r_P, so that $a_P^+ = r_P \cdot n_P^+$.

Measuring drug sensitivity separation accuracy based on subpopulation profiles. We measured the extent to which drug-sensitive and resistant cell populations could be separated based on their subpopulation profiles. We considered collections of cell populations (e.g. H460 clones), with each member: 1) represented by a subpopulation profile (i.e. a vector); and 2) assigned either as drug resistant (R) or sensitive (S) according to their drug sensitivity measure. The hyperplane that "best" separated the sensitive and resistant cell populations, based on their subpopulation profiles, was computed using the support vector machine (SVM) algorithm implemented in Matlab version 7.4.0 (A linear

kernel was used to avoid data over-fitting). Separation accuracy was computed by counting the percentage of clones (or cell lines) that were correctly classified by the SVM. We assessed the statistical significance p-value of the separation accuracy against a background distribution associated to random permutations of drug sensitivity assignment among all cell populations. The background distribution of the separation accuracy was estimated based on 10^6 iterations of random permutations of drug sensitivity assignments.

Measuring drug sensitivity separation among small numbers of cellular populations. When the number of clones (cell lines) n is relatively small versus the dimension of subpopulation profiles K (e.g. $K = 5$ and $n \leq 10$), estimation of the separation significance becomes less precise since many configurations of random drug sensitivity (R/S) assignment can lead to high separation accuracy. Instead, we chose to assess the extent of drug sensitivity separability as follows: 1) apply linear SVM to the dataset and measure the separation accuracy; 2) remove r populations from the original dataset and measure the separation accuracy; 3) repeat step 2 by considering all possible ways of removing r populations from the original dataset, then compute the overall average separation accuracy; 4) repeat steps 2-3 for increasing values of r (from 0 to 4 in our study). To maintain balance between the numbers of R/S labels, we discarded configurations where the numbers of sensitive and resistant populations differ by more than one. The extent of R/S separation in the dataset, for each value of r, was compared to the average separation accuracy obtained over all distinct permutations of the R/S labels in the dataset (Suppl. Fig. 11, 13). Such a comparison provides an indication as to whether the actual configuration shows higher level of drug sensitivity separability than random.

Determination of cell cycle state (Suppl. Fig. 5). To determine the cell cycle state, we used an empirical approach based on the distribution of total DNA intensity per cell across the H460 clone dataset. A two-class Gaussian mixture model was automatically fitted to the distribution of total DNA intensity per cell. The means m_1, m_2 (with $m_1 < m_2$) and

standard deviations s_1, s_2 of the mixture model were used to derive the following classification rule: a cell with total DNA intensity I_{tot} is in

G_1 state if $I_{tot} \leq m_1 + s_1$,

S state if $m_1+s_1 < I_{tot} \leq m_2-s_2$, and

G_2/M state if $I_{tot} > m_2-s_2$.

Preservation of subpopulation profile similarity across marker sets (Suppl. Fig. 10). To measure the extent to which similarity of subpopulation profiles among clones (or cell lines) is preserved across marker sets (e.g. from marker set *P* to *Q*), we proceeded as follows: 1) identify the *k* nearest neighbor clones of clone *i* based on the KL dissimilarity between their profiles in marker set *P*; 2) compute the average pairwise dissimilarity measure between clone *i* and these *k* clones in marker set *Q*, $D_Q^{(k)}(i)$; 3) build a background distribution for $D_Q^{(k)}(i)$, $[D_{BG,Q}^{(k)}(i)]$, by repeatedly computing the average pairwise dissimilarity $D_{BG,Q}^{(k)}(i)$ between clone *i* and *k* other randomly selected clones in marker set *Q*; 4) repeat steps 1-3 for all clones and count the total number of clones for which the significance p-value of $D_Q^{(k)}(i)$ versus $[D_{BG,Q}^{(k)}(i)]$ is less than a fixed threshold *t* ($t = 0.05$ in our analysis). In the "random" configuration, we generated for each clone *i* another realization of $D_{BG,Q}^{(k)}(i)$, $d_{BG,Q}^{(k)}(i)$, estimated its significance p-value against $[D_{BG,Q}^{(k)}(i)]$, and finally counted the number of clones for which $d_{BG,Q}^{(k)}(i)$ is significant against $[D_{BG,Q}^{(k)}(i)]$.

Assessing the accuracy of drug sensitivity prediction of all H460 clones using a Leave-One-Out strategy (Suppl. Fig. 12). To test how well our heterogeneity models can predict drug sensitivity, we used a "leave-one-out" cross-validation approach. Each clone was systematically excluded from the training set when building the heterogeneity model and then used as the test set for classification. This process was repeated independently for each of the 50 clones and 6 marker sets.

Measuring the robustness of heterogeneity models in revealing drug sensitivity separation (Suppl. Fig. 15). We assessed the degree to which the number of subpopulations used in our GMM-based heterogeneity model affects the accuracy of separating the H460 clones by their drug sensitivities. Specifically, a ten-fold cross-validation was performed for different numbers of subpopulations, *K* (we tested *K* between 2 and 14).

Measuring reproducibility of drug sensitivity separation accuracy over time. A subset of 10 H460 clones and the 8 selected NCI-60 cell lines were re-assayed from replicate freeze-downs. For the time course experiments, cells were kept in continuous culture for extended period after thawing. Results are shown in Suppl. Table 5.

Measuring reproducibility of subpopulation profiles among replicates wells (Suppl. Fig. 8). To test for reproducibility of subpopulation profiles from the three replicate wells of each clone, we first generated profiles from 3 replicate wells for each of the 49 clones or 7 parental controls of the H460 cell line (total number: 168 = 3x(49+7)). We computed the cumulative distribution functions (cdf) for the pairwise Kullback-Leibler (KL) dissimilarity measures, averaged over different choices of triplets of subpopulation profiles. The triplets of profiles were chosen to be replicates of the same clones ("*intraC"*); distinct clones belonging to the same cluster in the hierarchical trees as illustrated in Suppl. Fig. 7 ("*intraK"*); or clones from distinct clusters in the hierarchical trees ("*interK").* We measured the distributions of *intraC, intraK* and *interK* by applying four different threshold values (*th* = 0.1, 0.2, 0.3, 0.4) to the linkage distances between linked nodes to split the hierarchical tree (the linkage distances were provided by the Matlab function *linkage.m).* Threshold *th=0* assigns each profile to a separate cluster, while *th=1* performs no cut, i.e. one cluster contains all profiles. When replicate wells yield reproducible subpopulation profiles, the cdf's for *intraC* and *intraK* should lie above that of *interK* (that is, for *intraC* and *intraK* should have much smaller values than *interK*).

Data availability. Due to the large data size, post-processed data along with README.txt can be downloaded at:

http://www4.utsouthwestern.edu/altschulerwulab/papers/msb2010/default.html.

Raw images of a subset (extreme 10 + parent, MS1) (~3 GB) or the complete set of H460 clones (~17 GB per marker set) are available upon request.

Supplementary Table S1. Antibodies and dyes used in immunofluorescent staining.

Marker Set	Marker	Active/Inactive	Catalog #	Lot #	Dilution
DNA	Hoechst 33342		Invitrogen H1399	-	10 μg/ml
Marker Set 1	phospho-STAT3 (S727)	Active	BD Transduction Laboratories 612543	97087	1:100
	phosphor-PTEN (pSpTpS380/382/385)	Inactive	Biosource 44-1066G	103	1:100
Marker Set 2	phospho-ERK1/2 (pTpY185/187)	Active	Biosource 44680A1	1388699A	1:100
	phospho-P38 (pTpY180/182)	Active	Sigma M8177	104K4788	1:100
Marker Set 3	E-cadherin FITC	-	BD Transduction Laboratories 612131	45433	1:100
	β-catenin	-	BD Transduction Laboratories 610154	76283	1:100
	phospho-GSK3-β (S9)	Inactive	Biosource 44-600G	2601	1:100
Marker Set 4	phospho-Akt (pS473)	Active	Biosource 44-621G	502	1:100
	Histone 3 Lysine-9 acetylated (H3K9-Ac)	Active	Abcam ab12179	648913	1:500
Marker Set 5	Actin (Phalloidin Alexa 488)	-	Invitrogen A12379	23896W	1:40
	β–tubulin	-	BD Transduction Laboratories 558608	68808 B	1:50
Marker Set 6	Glyceraldehyde 3-Phosphate dedhydrogenase (GAPDH)	-	Abcam ab9485	448196	1:500
	Pericentrin	-	Abcam ab4448	26969	1:500
Apoptotic Marker (Drug experiment)	Annexin-V-FITC	Active	BD Pharmingen 51-65874X (556420)	88205	1:100
	Cleaved Caspase-3	Active	BD Transduction Laboratories 559565	180	1:100
	Poly ADP Ribose Phosphate (PARP)	Active	BD Transduction Laboratories 550781	97484	1:100
Secondary Antibodies	Anti-mouse IgG-Alexa 488	-	Molecular Probes A11001	56881A	1:1000
	Anti-mouse IgG-Alexa 647	-	Molecular Probes A21235	51782A	1:1000
	Anti-mouse IgG-Alexa 546	-	Molecular Probes A11003	53045A	1:1000
	Anti-rabbit IgG-Alexa 488	-	Molecular Probes A11008	54155A	1:1000
	Anti-rabbit IgG-Alexa 546	-	Molecular Probes A11010	435414	1:1000
	Anti-rabbit IgG-Alexa 647	-	Molecular Probes A21244	459547	1:1000

Supplementary Table S2. List of NCI-60 cell lines with extreme paclitaxel sensitivity tested for correlation between patterns of heterogeneity and drug sensitivity (http://dtp.nci.nih.gov/docs/cancer/cancer_data.html).

	Cell line	Origin	Paclitaxel GI50 index
Sensitive	HCT-116	Colon	-8.573
	HT29	Colon	-8.57
	MCF7	Breast	-8.545
	NCI-H460	Lung	-8.53
	HS 578T	Breast	-8.506
Resistant	CAKI-1[†]	Renal	-6.686
	ACHN	Renal	-6.384
	OVCAR-4	Ovarian	-6.189
	UO-31	Renal	-5.96
	NCI/ADR-RES	Ovarian	-5.515

(†) CAKI-1 was discarded since its identity could not be confirmed by DNA fingerprinting.

Supplementary Table S3. Layout of 96-well imaging plate for assaying non-drug treated (top) and drug-treated (bottom) cell populations (R*j* = replicate well *j*, *j=1…3*).

Plate layout for assaying non-treated cells

	Marker Set1			Marker Set 2			Marker Set 3			Marker Set 4		
Clone 1	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Clone 2	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Clone 3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Clone 4	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Clone 5	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Clone 6	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Clone 7	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Parent	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3

Plate layout for assaying drug sensitivity

	DMSO			DMSO+Doxorubicin						DMSO+Paclitaxel		
Clone 1	R1	R2	R3	R1	R2	R3				R1	R2	R3
Clone 2	R1	R2	R3	R1	R2	R3				R1	R2	R3
Clone 3	R1	R2	R3	R1	R2	R3				R1	R2	R3
Clone 4	R1	R2	R3	R1	R2	R3				R1	R2	R3
Clone 5	R1	R2	R3	R1	R2	R3				R1	R2	R3
Clone 6	R1	R2	R3	R1	R2	R3				R1	R2	R3
Clone 7	R1	R2	R3	R1	R2	R3				R1	R2	R3
Parent	R1	R2	R3	R1	R2	R3				R1	R2	R3

Supplementary Table S4. Kolmogorov-Smirnov (KS) two-sample, one-sided test showed that the collection of H460 clones exhibits significantly larger spread of phenotypic heterogeneity than the collection of HBEC clones ((†)p-value < 10^{-2}; p-value < 10^{-3} for all other cases).

Number of clusters	KS score	
	Marker Set 1	Marker Set 4
3	0.2297	0.0917
4	0.2678	0.0767
5	0.2305	0.0868
6	0.1822	0.0736
7	0.1950	0.0573†
8	0.1339	0.0678
9	0.1241	0.0822
10	0.1370	0.0928

Supplementary Table S5. Reassessment of a subset of 10 H460 clones and the 8 selected NCI-60 cell lines from replicate freeze-downs showed that the drug sensitivity separation is reproducible.

Cell Type	Time	Marker Set	
		1	4
Extreme 10 H460	Original	100	80†
	Repeat	100	100
	Week 1	70†	80†
	Week 2	80†	70†
	Week 4	60†	80†
Extreme 8 NCI-60	Original	87.5†	100
	Repeat	75†	100
	Week 8	87.5†	100

(†) Separation accuracy not one standard deviation above the average accuracy over all possible permutations of drug sensitivity assignment

Supplemental References

Borg I, Groenen P (1997) *Modern Multidimensional Scaling: theory and applications*. New York: Springer-Verlag.

Carney DN, Gazdar AF, Bepler G, Guccion JG, Marangos PJ, Moody TW, Zweig MH, Minna JD (1985) Establishment and identification of small cell lung cancer cell lines having classic and variant features. *Cancer Res* **45:** 2913-2923.

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc* **39:** 1-38.

Hastie T, Tibshirani R, Walther G (2001) Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63:** 411-423.

Kaufman L, Rousseeuw PJ (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.

Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* **22:** 79-86.

Loo LH, Wu LF, Altschuler SJ (2007) Image-based multivariate profiling of drug responses from single cells. *Nat Methods* **4:** 445-453.

NCI/NIH DTP http://dtp.nci.nih.gov/docs/misc/common_files/cell_list.html. *http://dtpncinihgov/docs/misc/common_files/cell_listhtml*.

Rasband WS, ImageJ, National Institutes of Health B, Maryland, USA, http://rsb.info.nih.gov/ij/ (1997-2009).

Schwarz G (1978) Estimating the Dimension of a Model. *Annals of Statistics* **6:** 461-464.

Slack MD, Martinez ED, Wu LF, Altschuler SJ (2008) Characterizing heterogeneous cellular responses to perturbations. *Proc Natl Acad Sci U S A* **105:** 19306-19311.

Vaughan MB, Ramirez RD, Wright WE, Minna JD, Shay JW (2006) A three-dimensional model of differentiation of immortalized human bronchial epithelial cells. *Differentiation* **74:** 141-148.

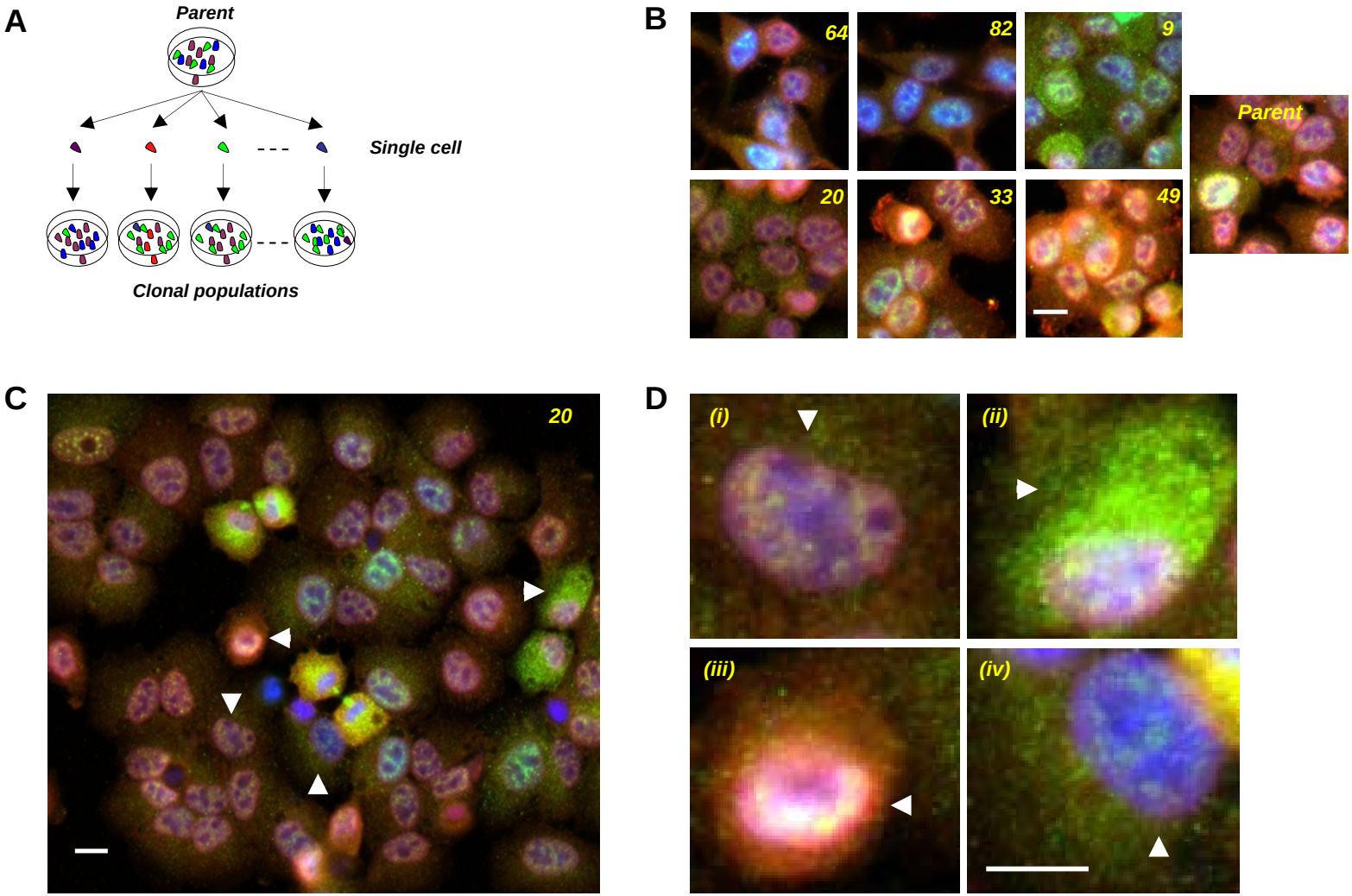


Figure S1. Heterogeneous signaling states are observed within and among a panel of non small cell lung cancer (NSCLC) H460 clones. **A-B.** A panel of 49 H460 clones (A) display phenotypically diverse signaling states as measured by activation and colocalization patterns of pSTAT3 and pPTEN immunostaining (B). While some clones are phenotypically similar to the parent (e.g. clones 20 and 33), others are dramatically dissimilar to the parent but similar to each other (e.g. clones 64 and 82). **C.** Heterogeneous cellular signaling states are observed within each clone. An expanded view of clone 20 reveals the presence of distinct, stereotyped cellular signaling states. Arrowheads indicate cells shown in (D). **D.** Distinct cell states present in one clone may be found in varying proportions within other clones. Shown are four example cells in distinct signaling states from clone 20. Cells with phenotypes similar to (i-iv) are seen in high proportions within the parent culture, clone 9, clone 49 and clone 82, respectively. Pseudocolors for images in (B-D) are: DNA-blue, pSTAT3-green, pPTEN-red. Scale bars: 20μm in (B-C) and 10μm in (D).

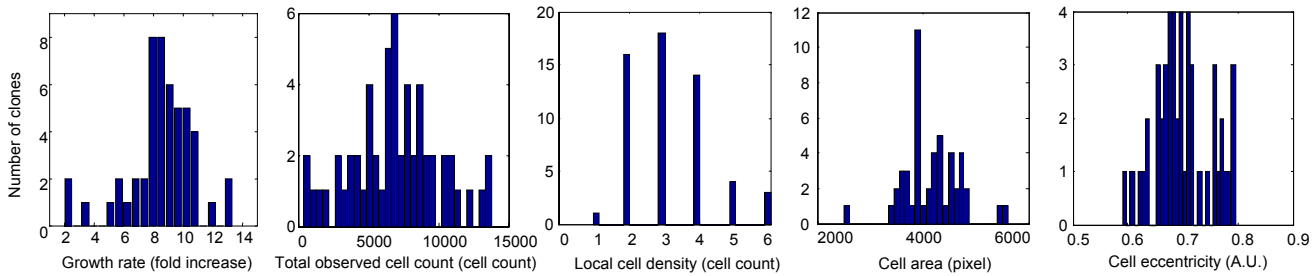


Figure S2. Non small cell lung cancer (NSCLC) H460 clones exhibit phenotypic variability. The H460 clones show a high degree of variability for growth rate, total observed cell count, local cell density (clumpiness), and morphology (as measured by cell area and eccentricity).

Figure S3. Clones of similar drug sensitivities have similar phenotypes across all marker sets. Shown are thumbnail images of all clones (columns) sorted in the same order as in Figure 2, from all four marker sets (rows). Image pseudocolors are as in Figure 1A. Relative drug sensitivities to paclitaxel are displayed under the thumbnail images according to the color bar above (red: resistant, black: intermediate, green: sensitive). Scale bar: 20μm.

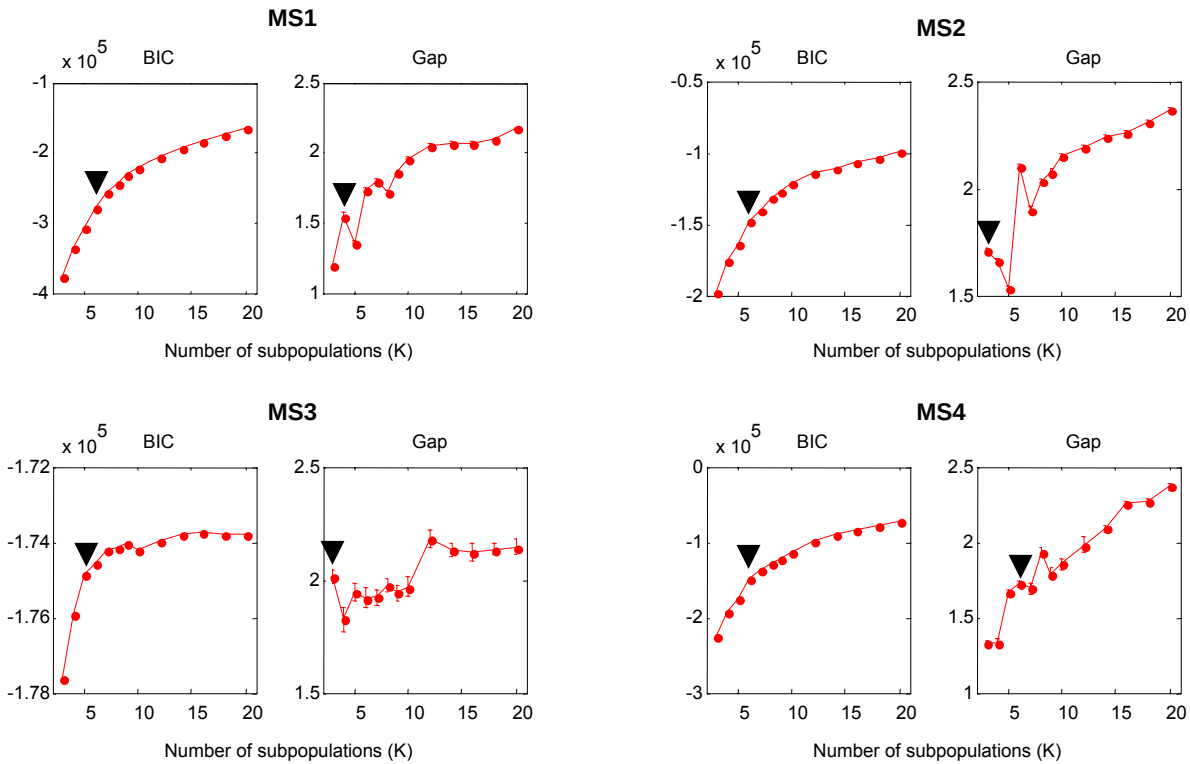


Figure S4. Optimal number of subpopulations in the reference model is suggested based on the Bayesian information-theoretical criterion (BIC) and the gap statistic (Gap). Traditionally the optimal number of subpopulation is found at the maximum of BIC value or before the gap statistic shows a significant decrease. However, with our large sample size (18,000) the BIC curve grows with increasing number of subpopulations. Therefore, we choose to identify the optimal number of subpopulations where the slope of the BIC curve starts to decrease (i.e. where the BIC curve shows an "elbow"). For all four marker sets, we found that the suggested optimal number of subpopulations consistently ranges between three and seven (black arrows). For simplicity, we use reference models with five subpopulations for all four marker sets.

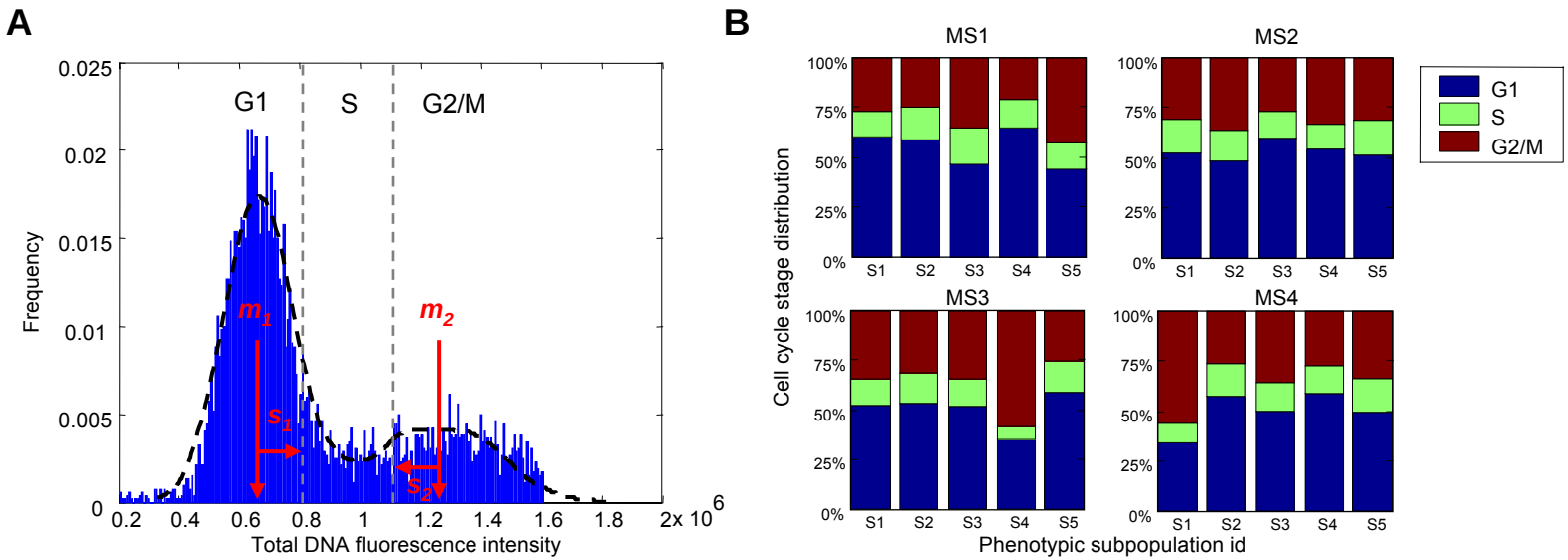


Figure S5. Phenotypic subpopulations identified by our Gaussian mixture models contain similar proportion of cells across different stages of cell cycle. A. Two-class Gaussian mixture model with means m_1,m_2 and standard deviations s_1,s_2 was automatically fitted to the histogram of total DNA intensity in each clonal population (black dashed line: density function of the fitted Gaussian mixture distribution, scaled so as to fit the contour of the histogram). Two threshold values m_1+s_1, m_2-s_2 (gray dashed lines) were used to classify the cell cycle stage of each cell either as G1, S or G2/M. **B.** The distribution of cell cycle stages within each phenotypic subpopulation was obtained by averaging over all H460 clonal populations.

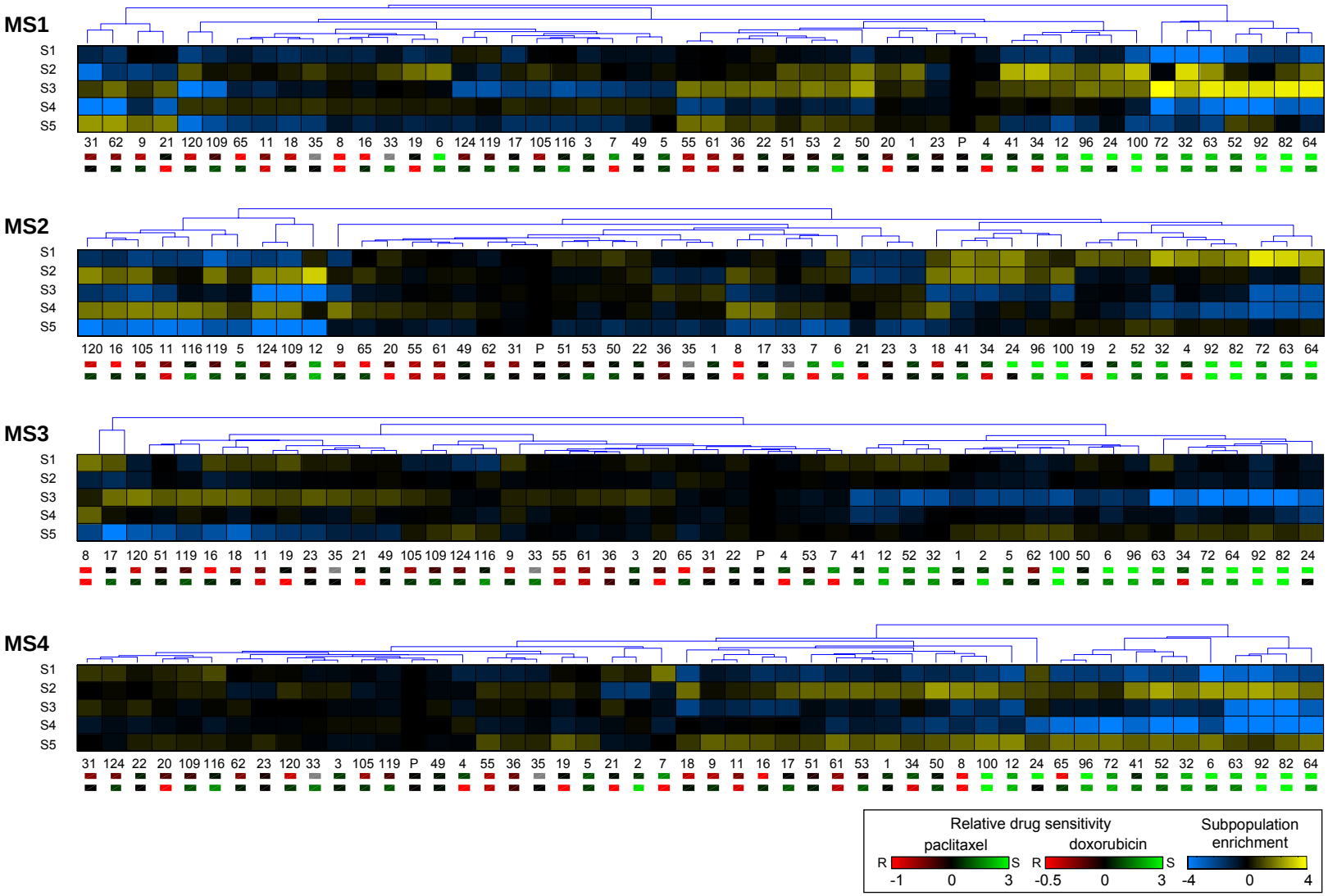


Figure S6. Clones with similar patterns of subpopulation enrichments tend to exhibit similar sensitivities to paclitaxel and doxorubicin. Subpopulation enrichment profiles are computed for each marker set as the log ratio of clone subpopulation proportions relative to the H460 parent population. An original clone ordering is determined by hierarchical clustering based on the similarity of their subpopulation enrichment profiles (using Euclidean distance). Tree nodes were then pivoted so that the average drug sensitivity of all clones under the left node of each branch is smaller or equal to the one under the right node (dendrogram at top). Drug sensitivities of each clone to paclitaxel are displayed in red (resistant), black (intermediate) and green (sensitive) according to the red-black-green scale bars above. Top row: paclitaxel; bottom row: doxorubicin. (Gray: paclitaxel sensitivity scores of clones 33 and 35 are unreliable due to an image-focus problem.)

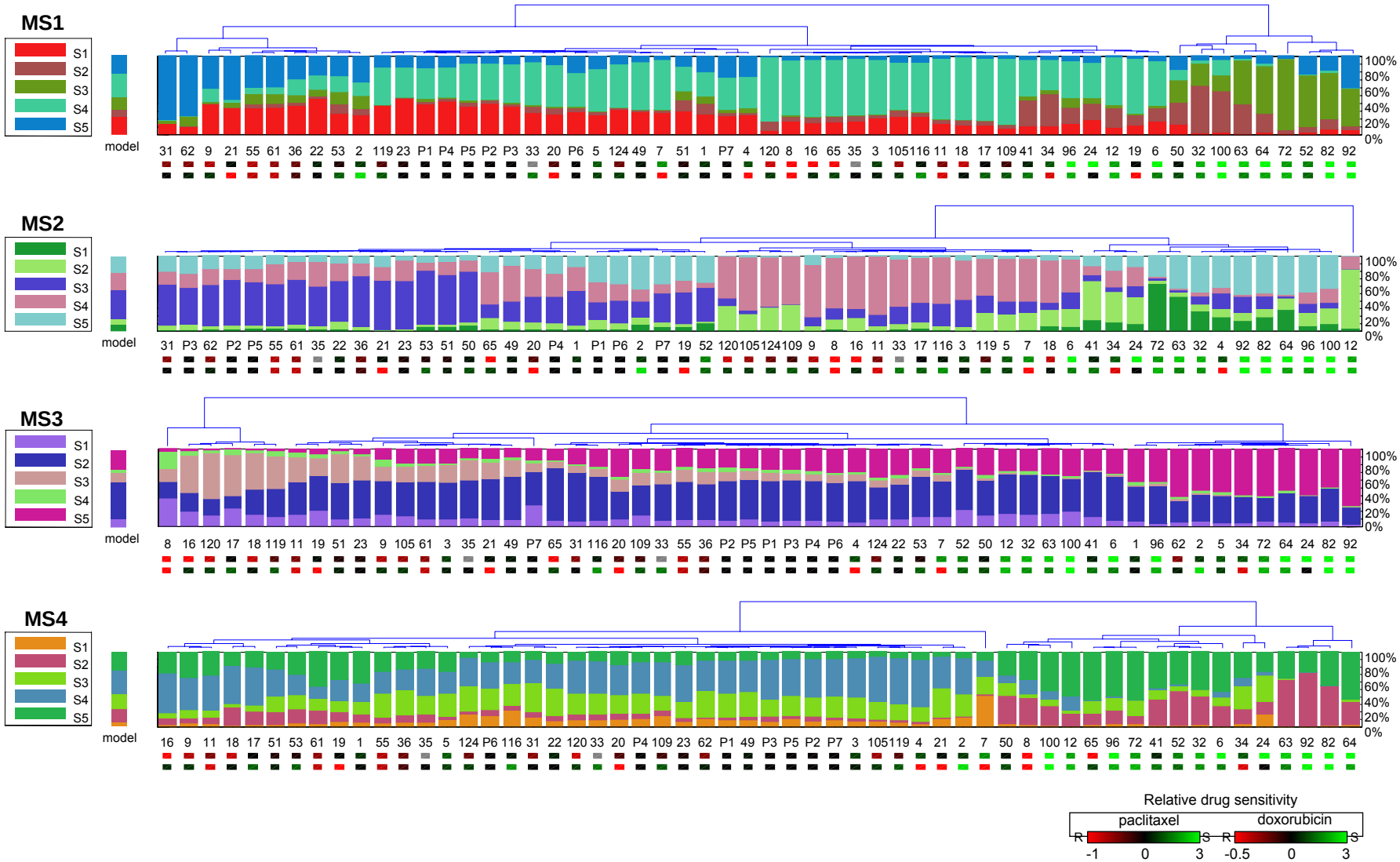


Figure S7. Clones with similar patterns of subpopulation profiles tend to exhibit similar sensitivities to paclitaxel and doxorubicin. Subpopulation profiles are computed for each marker set. An original clone ordering is determined by hierarchical clustering based on the similarity of their subpopulation profiles (using Kullback-Leibler dissimilarity measures). Trees were created using the "average" method in Matlab. Tree nodes were then pivoted so that the average drug sensitivity of all clones under the left node of each branch is smaller or equal to the one under the right node (dendrogram at top). Relative drug sensitivities are displayed under the clone indices according to the red-black-green scale bars above (red: resistant, black: intermediate, green: sensitive). Top row: paclitaxel; bottom row: doxorubicin. (Gray: paclitaxel sensitivity scores of clones 33 and 35 are unreliable due to an image-focus problem.)

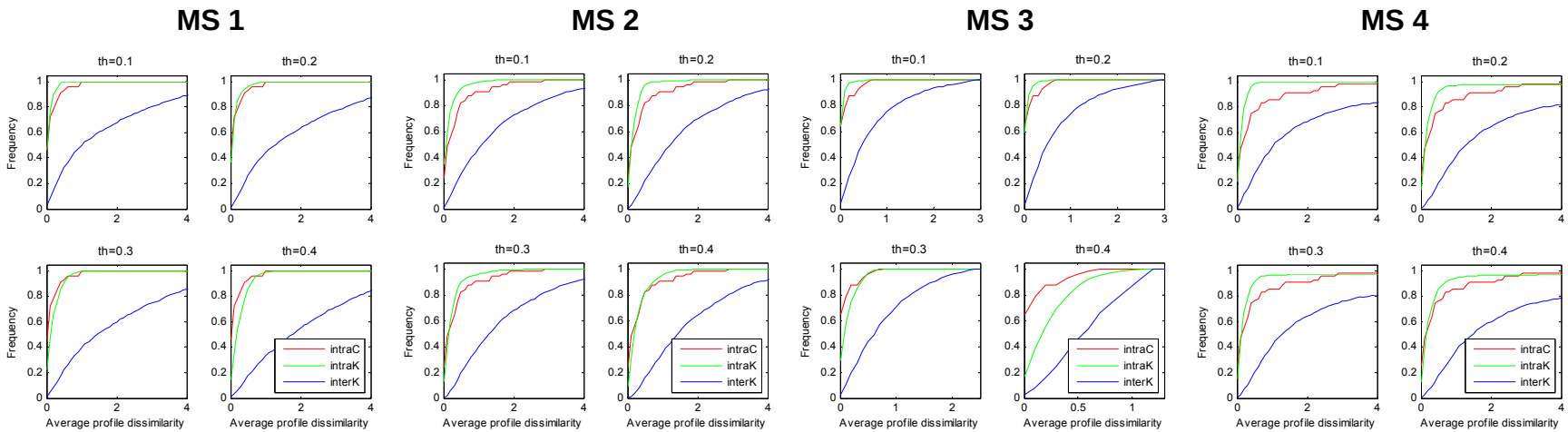


Figure S8. Replicate wells possess reproducible subpopulation profiles. Shown are the cumulative distribution functions for pairwise Kullback-Leibler (KL) dissimilarity measures, averaged over different choices of triplets of subpopulation profiles. The triplets of profiles were chosen to be replicates of the same clones (intraC, red); distinct clones belonging to the same cluster in the hierarchical trees as illustrated in Supplementary Fig. 8 (intraK, green); or clones from distinct clusters in the hierarchical trees (interK, blue). Shown are results for clustering by applying four different threshold values (*th*) to the linkage distances between linked nodes in the hierarchical tree (provided by the Matlab function *linkage.m)*; *th* = 0 yields no cuts (each profile is a separate cluster), *th* = 1 yields one cluster (one cluster contains all profiles). Profiles were obtained from 3 replicate wells from each of the 49 clones or 7 parental controls of the H460 cell line (total number: 168 = 3x(49+7)).

A

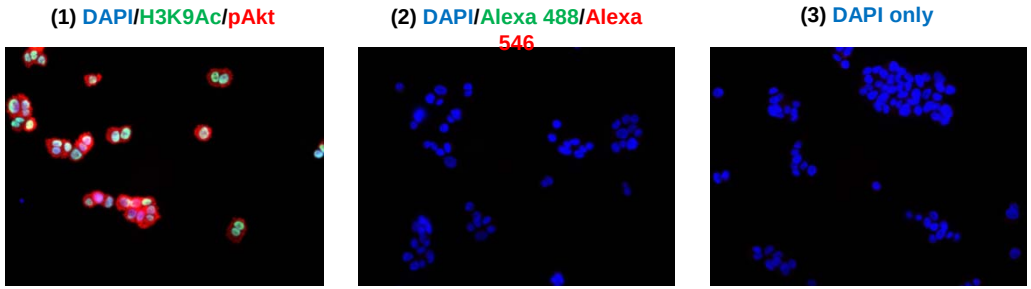
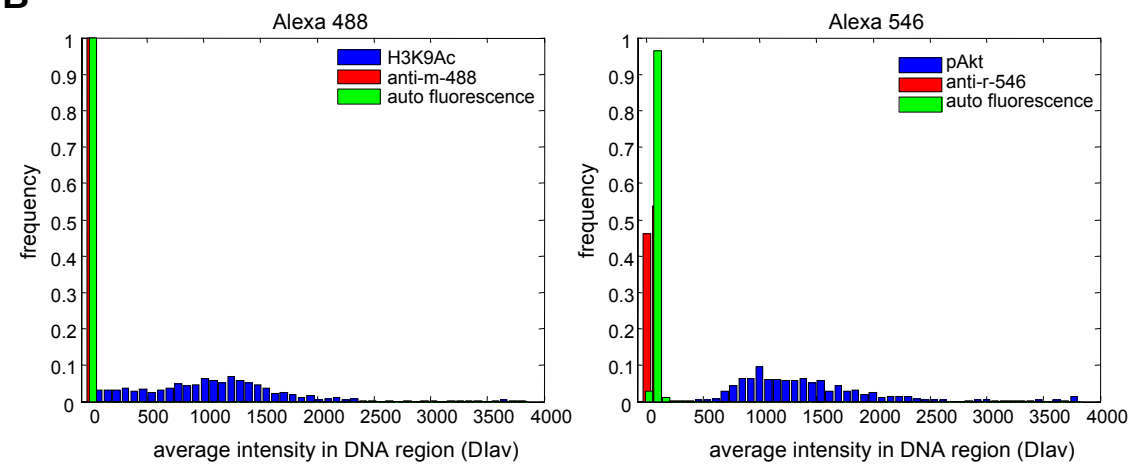


Image saturation range: DAPI (0-4000) Alexa 488 (0-2000) Alexa 546 (0-2000)

Intensity Range	(1)		(2)		(3)	
	Min	Max	Min	Max	Min	Max
Hoechst	0	4000	0	4000	0	4000
Alexa_488/520	0	3000	0	50	0	50
Alexa_546/617	0	3000	0	200	0	200

B



C

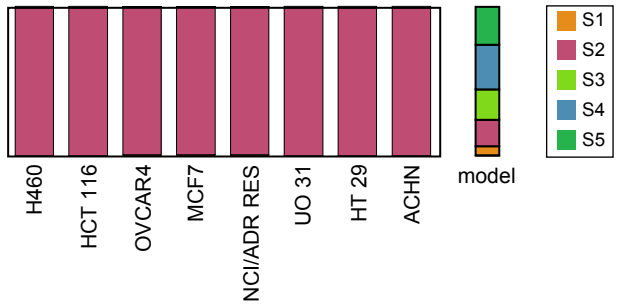


Figure S9. Cells with non-specific staining of secondary antibodies (without first antibodies) have background-like fluorescence intensity. A. RGB images of the H460 parental clone stained with 1) MS4 (Hoechst + primary and secondary antibodies) , 2) Hoechst + secondary antibodies, and 3) Hoechst alone. **B.** Average intensities in the DNA region obtained from secondary staining (red) and DAPI staining plus auto fluorescence (green) have very low value in the 488 and 546 channels (near background level) as opposed to those obtained from MS4 (blue). **C.** Subpopulation profiles of 8 NCI cell lines selected from NCI 60 panel (Supplementary Information) stained with secondary antibodies were dominated by a single subpopulation (S2).

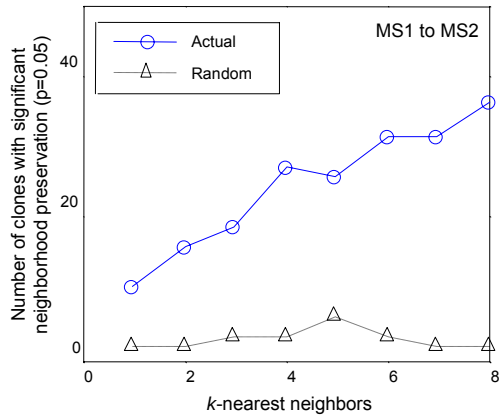


Figure S10. Similarity of phenotypic subpopulation profiles among the H460 clones is significantly preserved across different marker sets. Shown are the numbers of H460 clones that preserve significantly higher average similarity (in marker set 2) to their *k* nearest neighbors derived from marker set 1. The significance *p*-value was set to 0.05. The random case was obtained by replacing the *k* nearest neighbors by *k* randomly drawn distinct clones from the entire collection of clones (Supplementary Information).

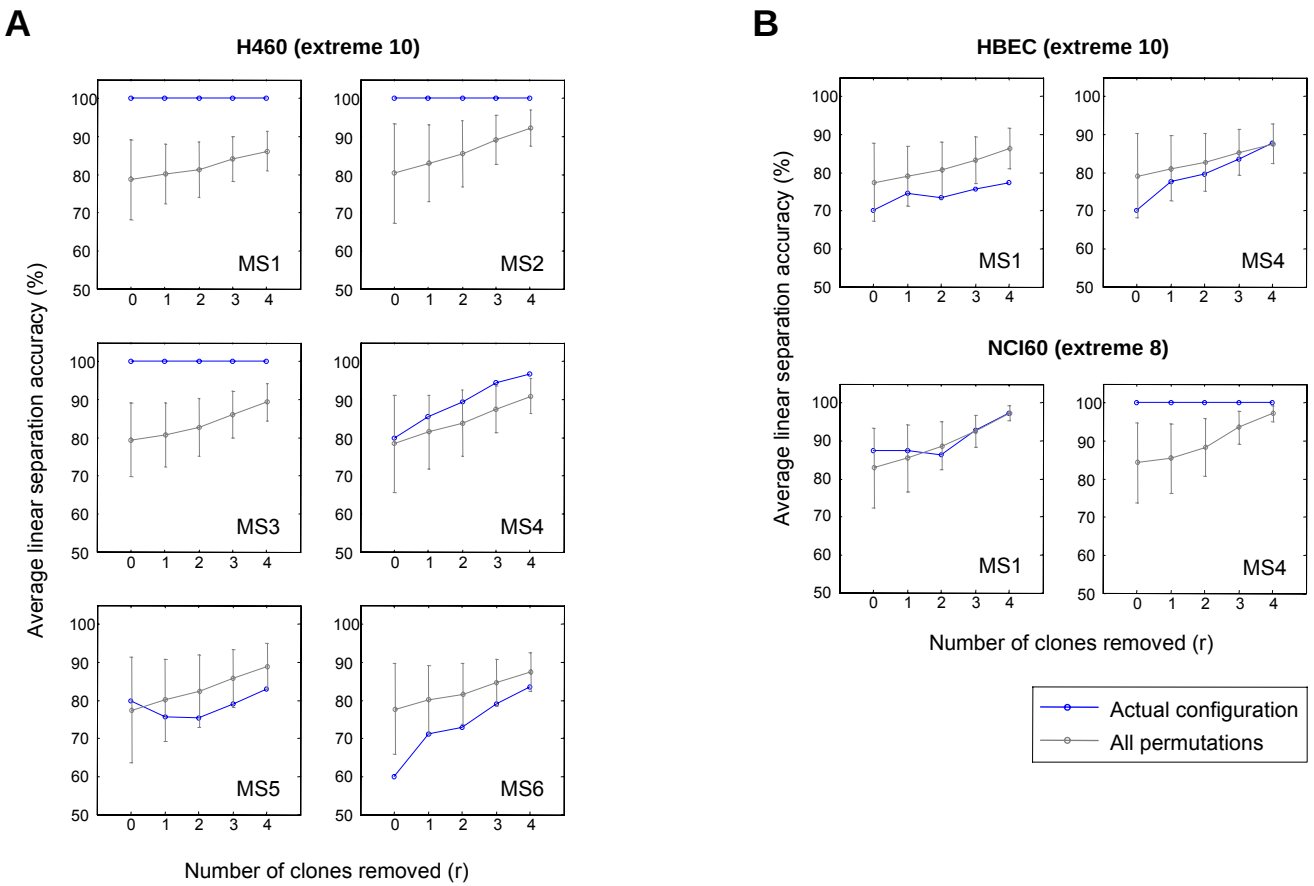
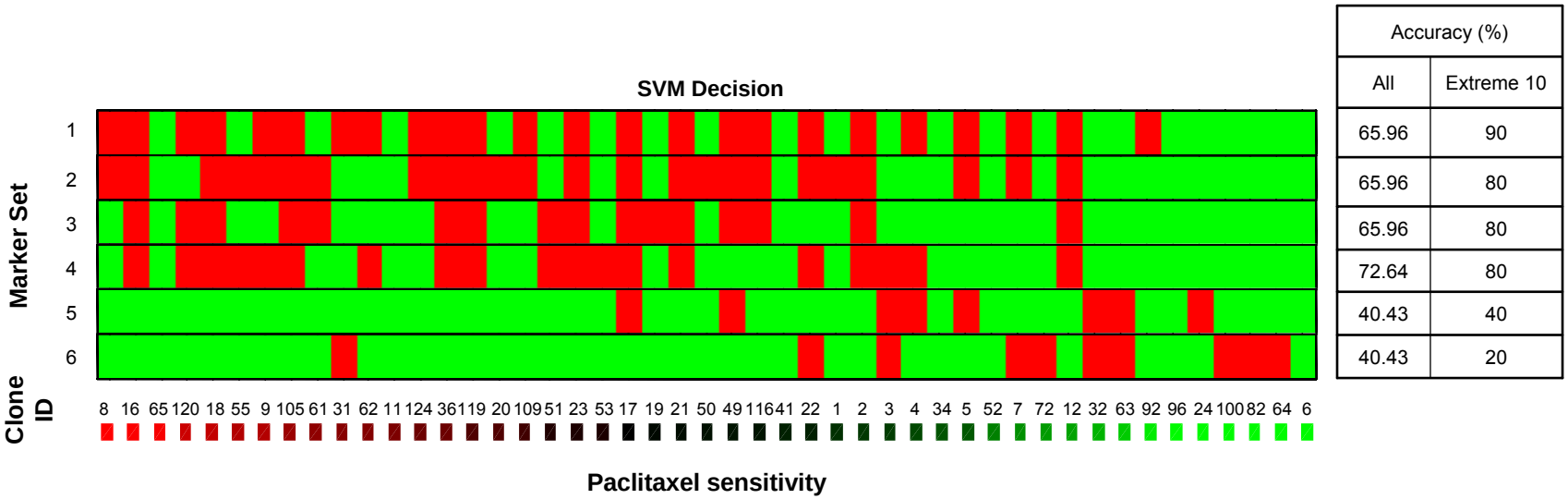


Figure S11. Measure of the separation between a small number of extreme sensitive (S) and resistant (R) cellular populations to paclitaxel. This procedure was applied when only a small number of data points were available (10 or less) and measure of the statistical significance of separation by drug sensitivity became less precise. Significance of the separation is, instead, determined by whether or not the separation accuracy of the actual assignment (blue circle) is higher than that of the random configuration (blue circle with an additional of one standard deviation). We measured the extent of separability between R/S clones by considering subsets of configurations where *r* clones (cell lines) were removed from the dataset. Blue circle: average separation accuracy by linear SVM. Gray circle: average separation accuracy by linear SVM over random permutation of drug sensitivity assignments. Gray error bars: standard deviation of the separation accuracy across random permutations of drug sensitivity assignments.



Accuracy (%)	
All	Extreme 10
65.96	90
65.96	80
65.96	80
72.64	80
40.43	40
40.43	20

Figure S12. The H460 heterogeneity model showed robust prediction accuracy of drug sensitivity across all four sets of signaling markers but not for "neutral" markers. The prediction accuracy results were obtained using a leave-one-out cross-validation, where each clone was excluded once from the training set when building the heterogeneity model and used as the test set. Left panel: red/green squares indicate SVM decisions of resistance/sensitivity for each clone in each marker set (top) and its actual drug sensitivity (bottom). Right panel: overall average prediction accuracies for all clones and for the extreme 10 (5 most sensitive and 5 most resistant).

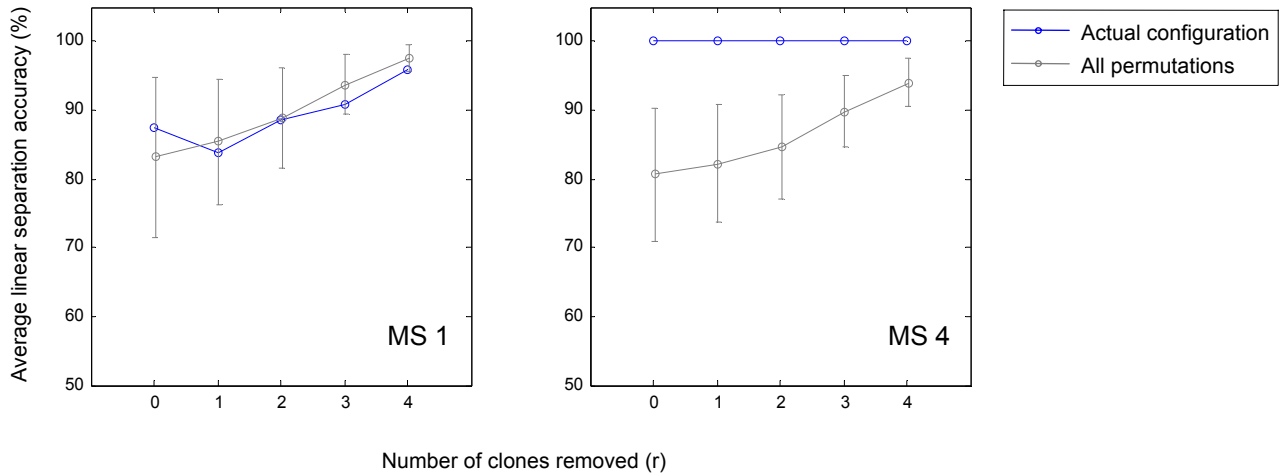


Figure S13. Separation accuracy for the paclitaxel resistant and sensitive cell lines is similar between a Gaussian Mixture Model (GMM) obtained from the NCI-60 (extreme 8) cell lines and the original H460 model. A new heterogeneity model was generated based solely on the NCI-60 (extreme 8) cell lines both for marker set 1 and for marker set 4. We removed the least resistant cell line from NCI9 to build a balanced data set (4 resistant and 4 sensitive cell lines). This model was used to derive heterogeneity profiles and calculate separation accuracies between the 4 most sensitive and 4 most resistant cell lines. This result is consistent with the bottom two panels of Supplementary Figure 11B which was obtained from the H460 model.

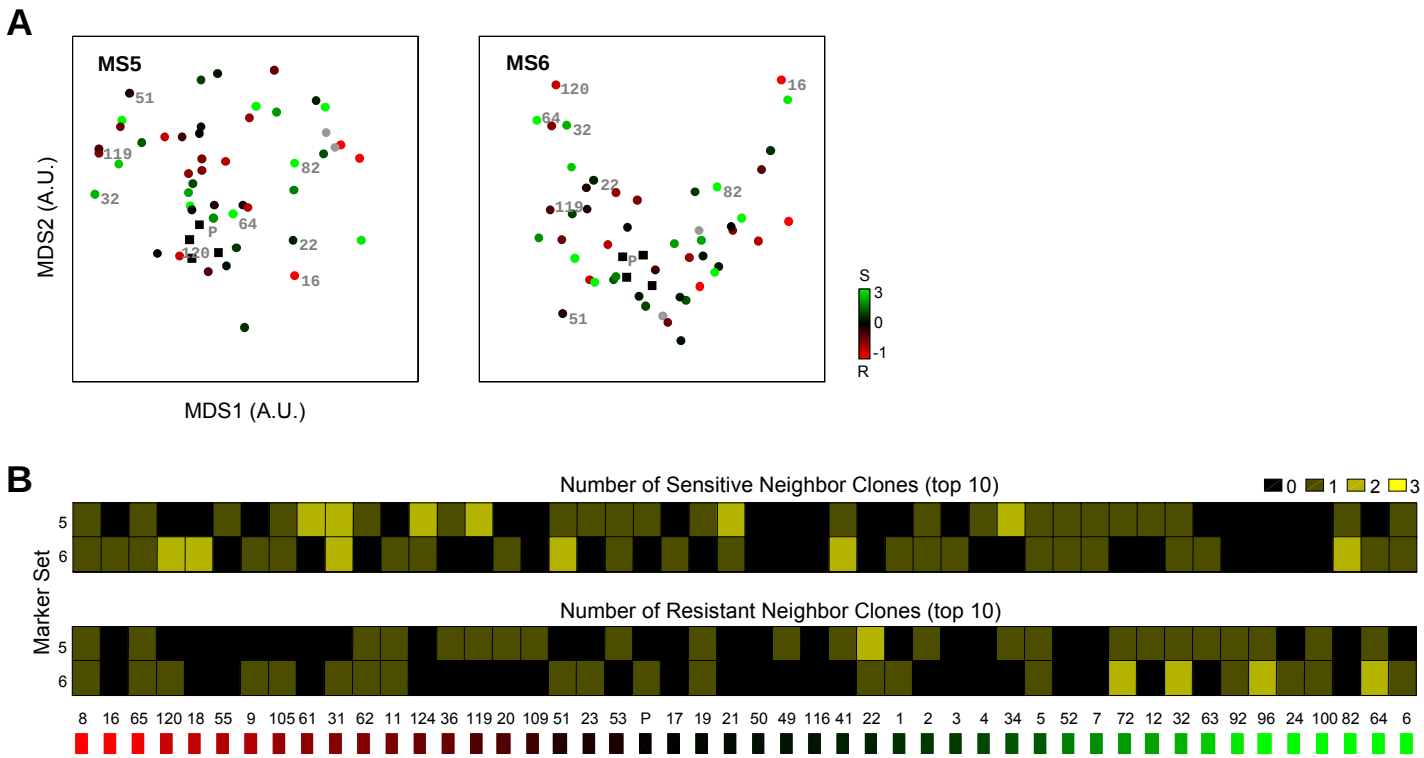


Figure S14. Heterogeneity profiles of H460 clones based on non-signaling markers shows no correlation to paclitaxel drug sensitivity. A. MDS plots for marker sets MS5-6 do not show any separation between paclitaxel-sensitive and resistant clones. **B.** The neighborhood relationship among the extreme paclitaxel-sensitive clones is not preserved in the two non-signaling marker sets.

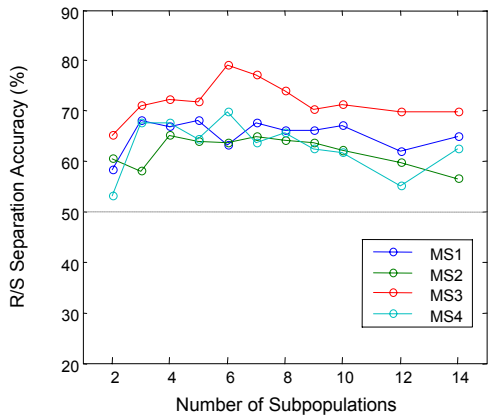


Figure S15. Heterogeneity profiles computed over a range of subpopulations can separate H460 clones by paclitaxel sensitivity. Tenfold cross-validation on drug sensitivity separation consistently shows that a small number of subpopulations gives the best separation performance across all four marker sets.