Supporting Information

Jones *et al*. 10.1073/pnas.0808843106

SI Text

Additional Data. Additional detailed supplemental data is available at www.cellprofiler.org/PNAS2009.html. Features List shows the cytological profile measured for each human cell. Classifier Rules contains rules that were used to score the human phenotypes in Figs. 3 and 4. Combined Scores contains a rank-ordered list of genes tested by RNA interference for the metaphase phenotype in *Drosophila*. AnalysisPIPE.txt contains a description of the image processing pipeline for the human screens. AnalysisPIPE.mat.zip contains the image-processing pipeline for the human screens, which can be loaded into CellProfiler to recreate the analysis. Sample Scores contains rank-ordered lists of human samples, scored for each phenotype.

Statistical Analysis and Validation. Fifty individual regression stumps constituted the final rule for each phenotype, using GentleBoosting. As an example, "IF(Cells_Intensity_Actin_StdIntensity >0.076096, 0.332262, −0.606784)" can be translated as: "If an individual cell has a standard deviation of actin pixel intensities within the whole cell greater than 0.076096, add 0.332262 to its score; otherwise subtract 0.606784 from its score." After all 50 stumps were applied in this manner, the cell was classified as positive for the phenotype if its score was >0 and negative if its score was <0. The rule was applied to all cells in all samples to count the number of phenotype-positive and phenotype-negative cells for each sample. The full set of positive and negative counts was fit with a beta-binomial model (1). To generate an enrichment score for each sample, the negative log (base 10) of the ratio of right and left *P* values was computed, as a fast approximation of the log-odds of enrichment versus suppression. For each human phenotype shown in Figs. 3 and 4, the rules and rank-ordered lists of samples are provided (see *Additional Data* in *SI Text*). For the *Drosophila* metaphase screen with multiple replicates, we computed combined enrichment scores for each gene by: (*i*) combining counts of positive and negative cells for replicate samples within each slide (usually, 3 replicates per gene per slide), (*ii*) computing enrichment scores separately for each slide to minimize any bias in the WT population from slide to slide, and (*iii*) summing the resulting enrichment scores for each gene, generating the rank-ordered list (see *Additional Data* in *SI Text*).

For forced-choice validation, 1 sample (an image showing a cell population) in each pair had been automatically identified as positive (enrichment score roughly ≥3; i.e., better than 1,000:1 odds of enrichment) and the other as neutral (enrichment score of ≈0, to samples with >50 cells). For simplicity, we chose the number of samples to validate in multiples of 10.

1. Gelman A, Stern H, Carlin J, Rubin D (2003) *Bayesian Data Analysis* (Chapman and Hall, London), pp. 118–120.

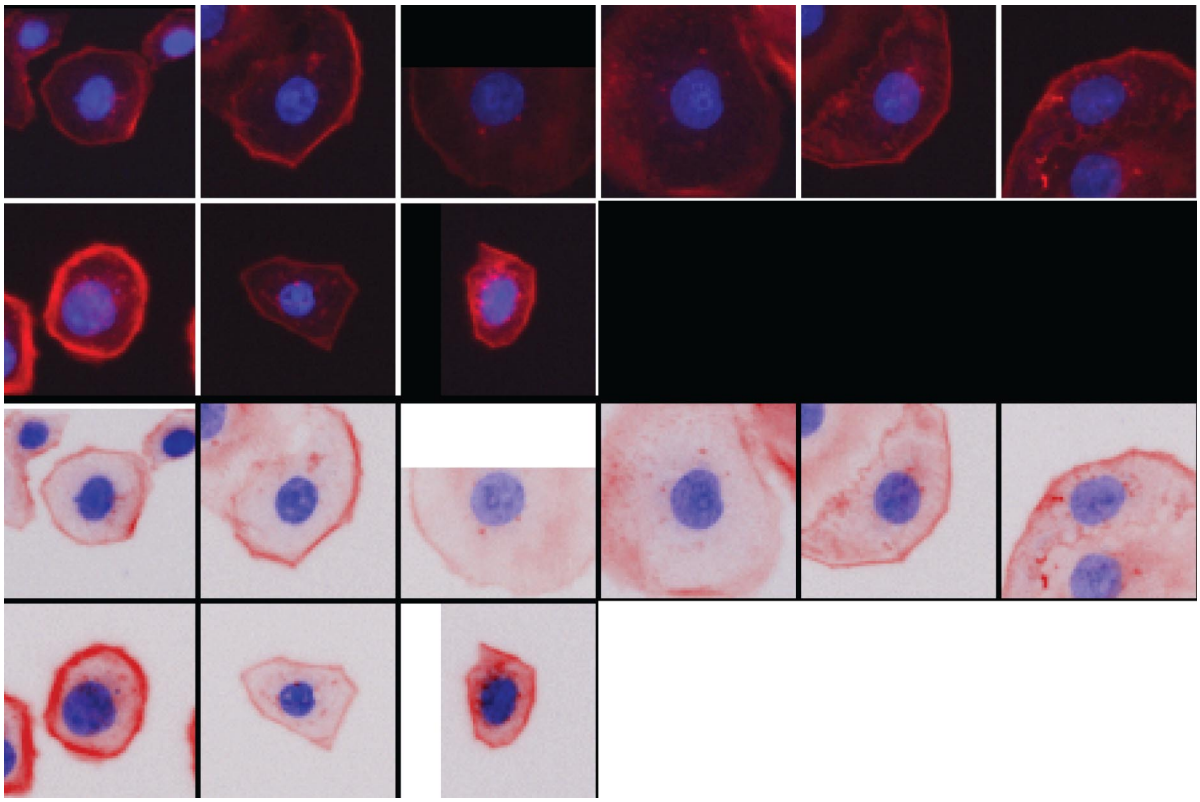

Fig. S1. Example cells showing the ''sparkly actin'' phenotype. Images are as shown in Figs. 3 and 4.

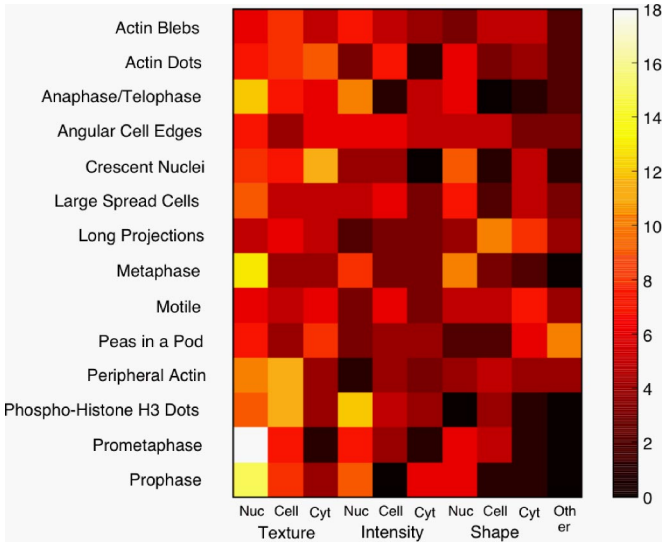

Fig. S2. Categories of features relevant to each human phenotype. For each phenotype's rule, the chart shows the number of regression stumps (out of 50 total for each rule) that reference a particular feature category. ''Texture,'' ''intensity,'' and ''shape'' indicate all features of the particular type, regardless of color channel, and are subdivided by cellular compartment within which the feature was calculated: nucleus, entire cell, or cytoplasm (i.e., cell body excluding nucleus). Correlation between channels is included with the Intensity category. ''Other'' includes neighbor counts of cells and cell location within images.

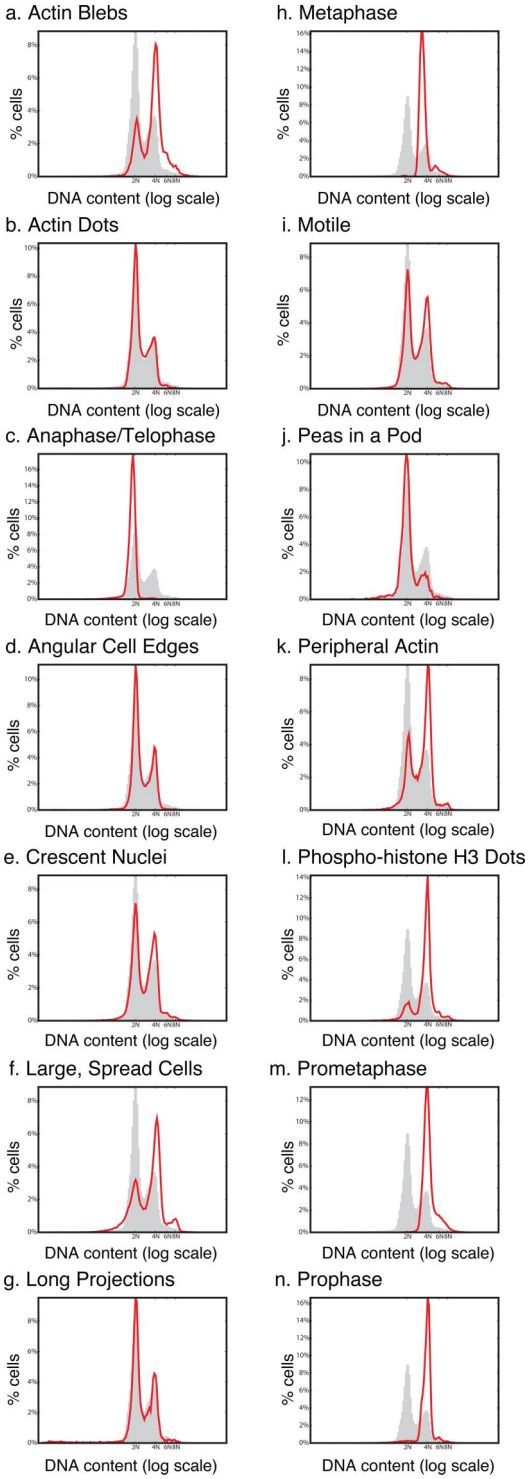

Fig. S3. Cell-cycle distribution of cells displaying each human phenotype. The cell-cycle distribution of cells displaying each phenotype (red) is shown relative to the cell-cycle distribution of all cells in the experiment (gray). DNA content is plotted on a logarithmic scale, with 2N, 4N, 6N, and 8N determined empirically from the base population as multiples of the 2N peak.

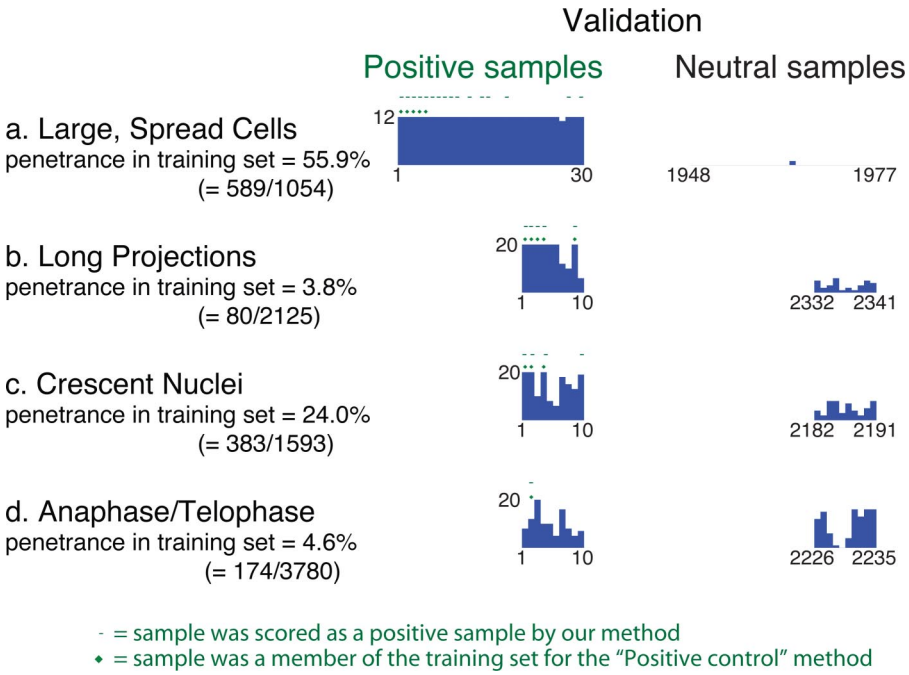

Fig. S4. Results of a previously existing positive control-based method on several phenotypes in human cells. Validation data are shown for samples that were scored as positive and neutral using a positive control-based method. Existing methods based on positive controls (1–3) were not strictly applicable to the phenotypes we selected because we had no a priori positive control samples for these phenotypes. Therefore, to apply a positive control-based method retrospectively, we used the top 5 samples (according to enrichment score) for each phenotype as positive controls (confirmed as positives by eye in Figs. 3 and 4) to create a training set. We chose not to use the entire set of positive samples (as scored by our method) to create the training set because it could have left no positive samples in the test set. We decided not to use only the top 1 or 2 hits to constitute the training set because we did not want the training to overfit to samples with limited variants of the phenotype, causing the method to miss other variants of the phenotype. Furthermore, the number of cells present in 1 or 2 samples was likely too small for adequate training, compared with a typical training set using our method. We therefore selected 5 as a round number that would give some variety and generate a training set size of at least 1,000 cells. The penetrance for the training set samples (noted in the figure as a percentage and as absolute numbers of positive cells/total cells) is as scored by the method presented in the main text. A random set of cells from other samples were used as negative examples, with the same number of negative examples as positives, and a rule was trained and used to score all samples. Negatives were taken from all other samples, rather than a small set of neutral images, to ameliorate potential overfitting to sample-to-sample variation. We selected the top hits from the positive control-based method as those with an enrichment score above roughly 3, and compared them to neutrally scored samples to validate them (according to the method presented in Figs. 2–4 and described in the main text). Note that the 5 strong positive samples that constituted the training set (marked with a green diamond) were included in the scoring process but did not always score as positives by this method. The training set samples, and all other samples that were scored as positives by our method, are marked with a dash.

1. Bakal, C, Aach, J, Church, G, Perrimon, N (2007) *Science* 316:1753–1756.
2. Loo LH, Wu LF, Altschuler SJ (2007) Image-based multivariate profiling of drug responses from single cells. *Nat Methods* 4:445–453.
3. Chen X, Murphy RF (2006) Automated interpretation of protein subcellular location patterns. *Int Rev Cytol* 249:193–227.

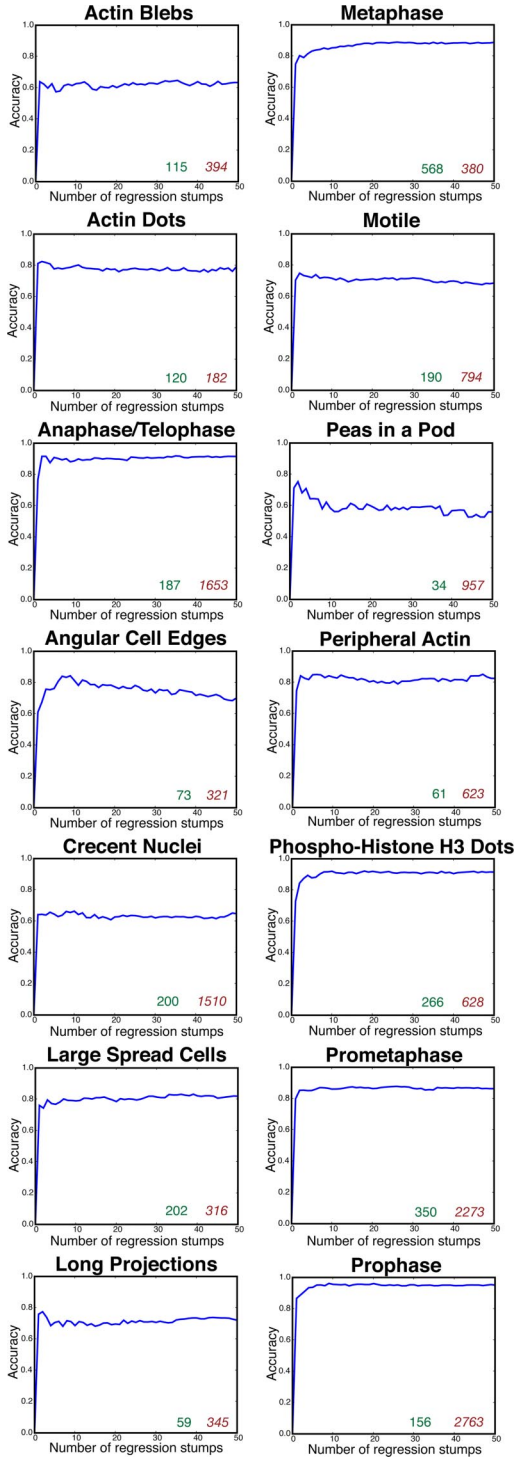

Fig. S5. Cross-validation performance for 14 human cell phenotypes. Using the training sets for each of the phenotypes shown in Figs. 3 and 4, accuracy calculated by 20-fold hold-out cross-validation is shown. Accuracy is plotted as the average of true positive and true negative rates for individual cells (in contrast to per-sample accuracy, as in Figs. 3 and 4). Performance for most phenotypes does not diminish upon increasing the number of individual regression stumps in the rule, up to 50. Exceptions are training sets with few positive examples, which suffer from some overfitting. Numbers in the lower right of each chart show the number of positive, negative example cells in the training set.

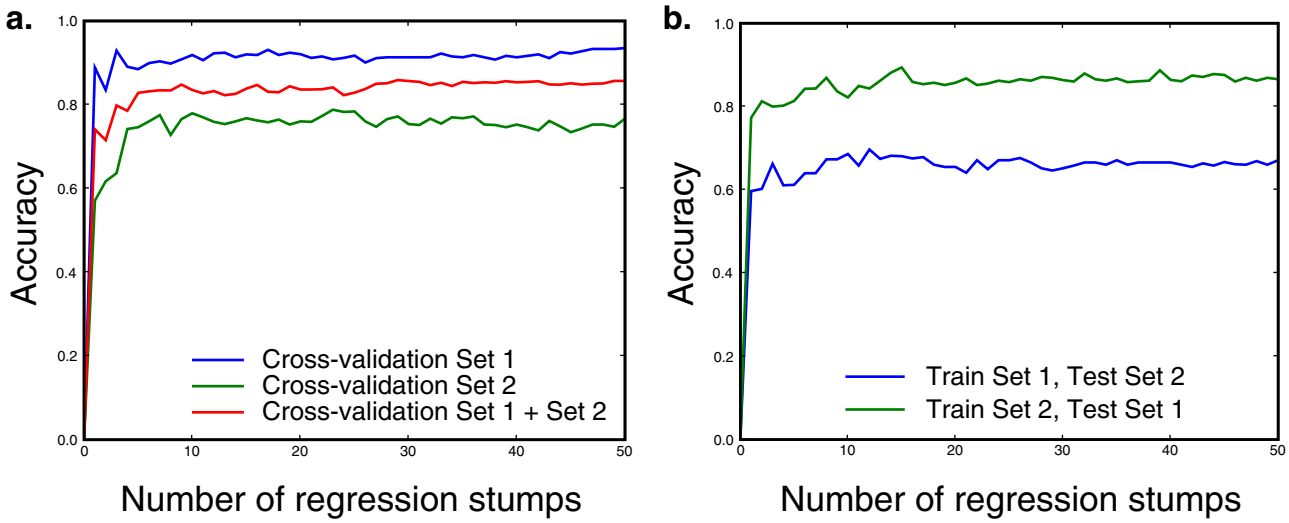

Fig. S6. Cross-experiment performance. A chemical screen probing a binucleate phenotype in *Drosophila melanogaster* was performed in 2 batches. Separate training sets were created for each batch to assess performance within and between experiments. Accuracy is plotted as the average of true positive and true negative rates for individual cells (in contrast to per-sample accuracy, as in Figs. 3 and 4). (*a*) Forty-fold hold-out cross-validation performance of the individual training sets, as well as their combination. (*b*) Cross-experiment performance of a rule trained on the first training set (comprised of 144 positive and 91 negative examples) and applied to the second (comprised of 155 positive and 93 negative examples), and vice versa.

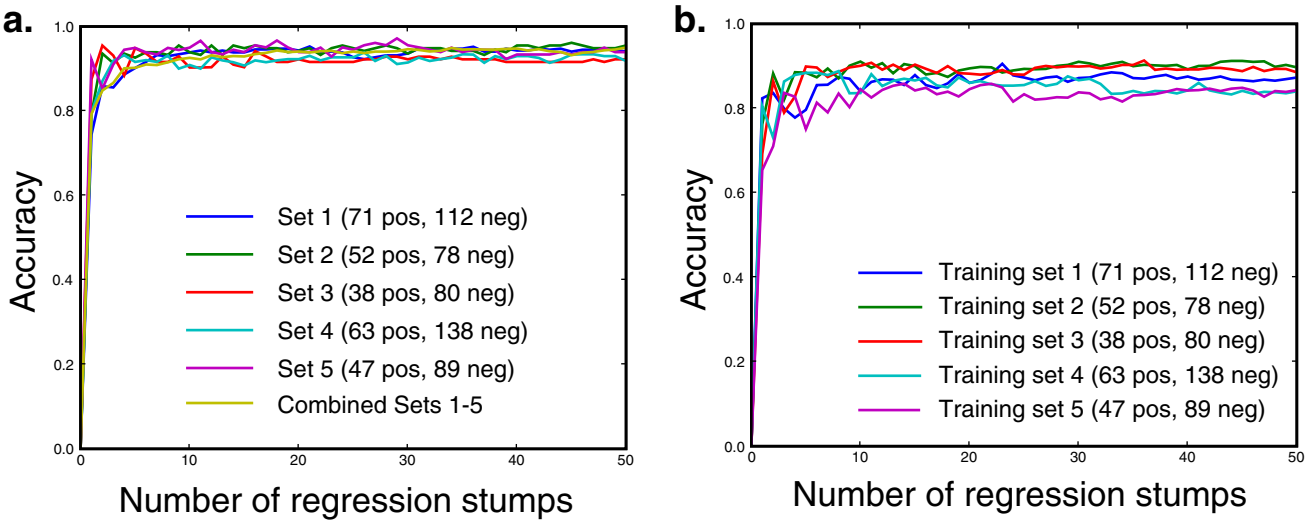

Fig. S7. Cross-validation (cross-experiment) performance on the *Drosophila* metaphase phenotype. An RNA interference screen in *D. melanogaster* was previously performed, as described in the text. The cytological profiles contained measurements based on DNA and actin staining, and a training set was created for cells in metaphase using cells from all 5 replicate slides. The training set was then broken into separate sets by slide for performance evaluation. (*a*) Forty-fold hold-out cross-validation performance of the individual slide's training sets, as well as their combination. (*b*) Cross-slide performance of a rule trained on each individual slide's training set and tested on the other 4 slides' training sets.

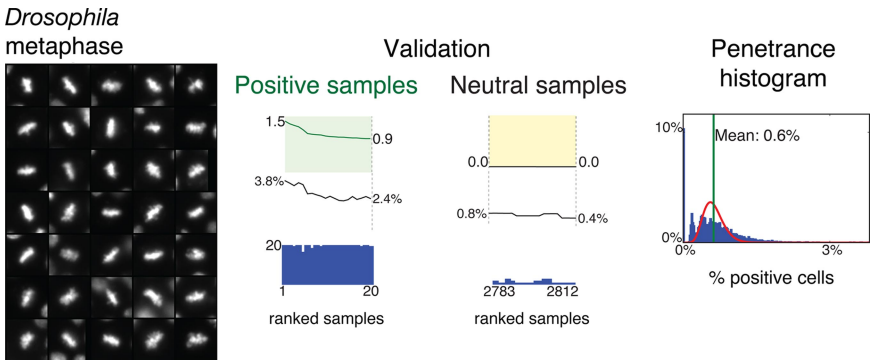

Fig. S8. Results of the phenotype-scoring system, for the metaphase phenotype in *Drosophila* cells. See Fig. 3 for details.

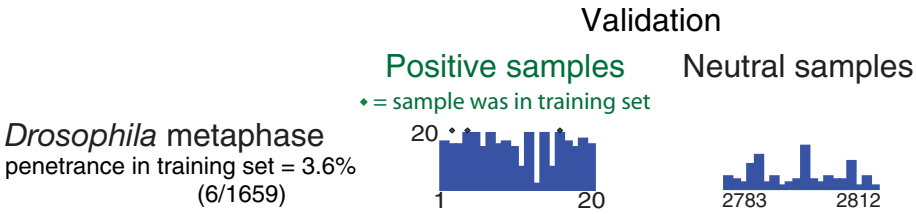

Fig. S9. Results of a previously existing positive control-based method on the *Drosophila* metaphase phenotype. See Fig. S4 for details.

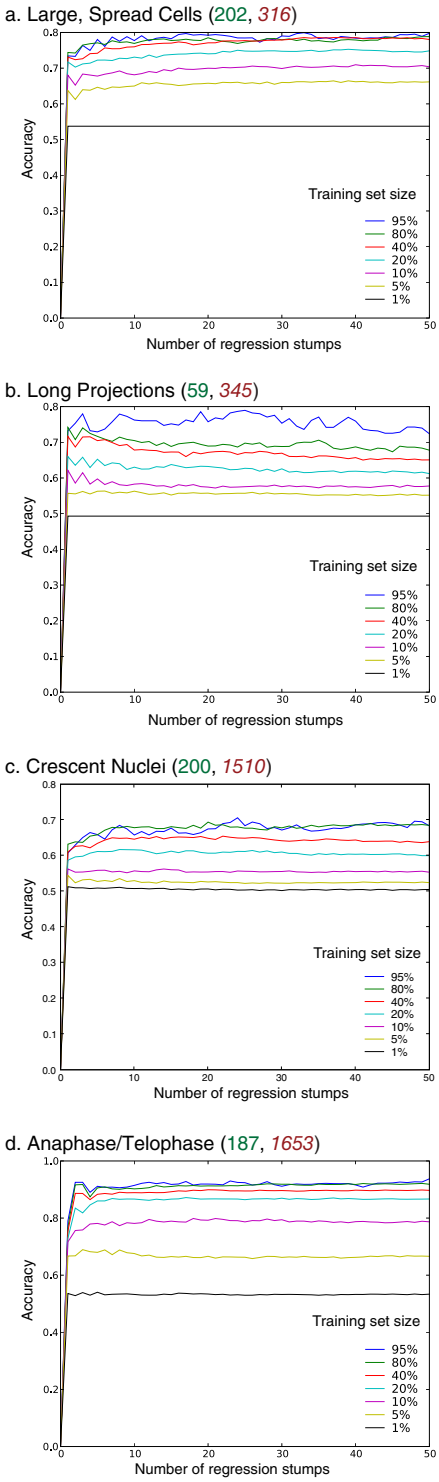

Fig. S10. Cross-validation performance as a training set increases in size. Random subsets of the total available training set of (# positive, # negative) example cells were selected for training, with the remaining cells held out for testing. For example, the 95% plot in *a* used an average of 492 cells for training and 26 for testing. Plots are the average of 40 repetitions of random selection, training, and testing. Accuracy is plotted as the average of true positive and true negative rates for individual cells (in contrast to per-sample accuracy, as in Figs. 3 and 4).

Table S1. Confusion matrices for human and automatic scoring

Phenotype	Average Human	Human 1 Hit	Human 1 Miss	Human 2 Hit	Human 2 Miss	Human 3 Hit	Human 3 Miss	Computer Hit	Computer Miss
Actin Dots	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Peripheral Actin	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Anaphase/Telophase	Hit	91.6	8.4	89.9	10.1			94.5	5.5
	Miss	8.4	91.6	10.1	89.9			5.5	94.5
Angular Cell Edges	Hit	99.5	0.5	98.6	1.4			99.5	0.5
	Miss	0.5	99.5	1.4	98.6			0.5	99.5
Crescents	Hit	93.3	6.7	91	9	92.6	7.4	94.2	5.8
	Miss	6.7	93.3	9	91	7.4	92.6	5.8	94.2
Prophase	Hit	96.6	3.4	94	6			97.5	2.5
	Miss	3.4	96.6	6	94			2.5	97.5
Actin Blebs	Hit	98.1	1.9	96.8	3.2			98.6	1.4
	Miss	1.9	98.1	3.2	96.8			1.4	98.6
Large Spread Cells	Hit	99.5	0.5	98.6	1.4			99.5	0.5
	Miss	0.5	99.5	1.4	98.6			0.5	99.5
Metaphase	Hit	99	1	97.3	2.7			99	1
	Miss	1	99	2.7	97.3			1	99
Motile	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Long Projections	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Peas in a Pod	Hit	98.6	1.4	99.5	0.5			99.5	0.5
	Miss	1.4	98.6	0.5	99.5			0.5	99.5
Prometaphase	Hit	99.3	0.7	99.7	0.3			99.7	0.3
	Miss	0.7	99.3	0.3	99.7			0.3	99.7
Phospho-Histone H3 Dots	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Drosophila metaphase	Hit	96.4	3.6	94.6	5.4			97.5	2.5
	Miss	3.6	96.4	5.4	94.6			2.5	97.5

From the forced-choice visual scoring of samples presented in Figures 2–5, every sample has a probability of being called a ''hit'' versus ''miss'' when presented in a comparison with another sample. This probability can be computed for each human scorer and for our method, as well as an average from all of the humans' scores. We show the probability of agreement/disagreement as percentages when compared to the average of the humans' scores. The confusion matrices are symmetric due to the forced-choice methodology.