SCIENTIFIC REPORTS
nature research

OPEN

Migration rather than proliferation transcriptomic signatures are strongly associated with breast cancer patient survival

Received: 15 February 2019
Accepted: 28 June 2019
Published online: 29 July 2019

Nishanth Ulhas Nair[1,2], Avinash Das[3,4], Vasiliki-Maria Rogkoti[5], Michiel Fokkelman[5], Richard Marcotte[6,7], Chiaro G. de Jong[5], Esmee Koedoot[5], Joo Sang Lee[1,2], Isaac Meilijson[8], Sridhar Hannenhalli[1], Benjamin G. Neel[6,9,10], Bob van de Water[5], Sylvia E. Le Dévédec[5] & Eytan Ruppin[1,2,11]

The efficacy of prospective cancer treatments is routinely estimated by *in vitro* cell-line proliferation screens. However, it is unclear whether tumor aggressiveness and patient survival are influenced more by the proliferative or the migratory properties of cancer cells. To address this question, we experimentally measured proliferation and migration phenotypes across more than 40 breast cancer cell-lines. Based on the latter, we built and validated individual predictors of breast cancer proliferation and migration levels from the cells' transcriptomics. We then apply these predictors to estimate the proliferation and migration levels of more than 1000 TCGA breast cancer tumors. Reassuringly, both estimates increase with tumor's aggressiveness, as qualified by its stage, grade, and subtype. However, predicted tumor migration levels are significantly more strongly associated with patient survival than the proliferation levels. We confirmed these findings by conducting siRNA knock-down experiments on the highly migratory MDA-MB-231 cell lines and deriving gene knock-down based proliferation and migration signatures. We show that cytoskeletal drugs might be more beneficial in patients with high predicted migration levels. Taken together, these results testify to the importance of migration levels in determining patient survival.

Drug development risk is a major contributing factor for spiraling drug prices[1]. Only 1 out of 5000 drugs from pre-clinical studies enter the market after successful clinical testing[2]. Cancer drugs show the highest proportion of failures on the road to clinics[3]. Currently, the prevailing experimental method to initially estimate the pre-clinical efficacy of cancer drug candidates is by measuring their effects on *in vitro* proliferation rates[3–10]. However, even after filtering these findings in animal models, only a fraction of emerging candidates has successfully translated into human trails[11–13]. Many factors contribute to the failure of drugs that are effective in pre-clinical systems. For starters, *in vitro* and *in vivo* systems are obviously only approximate models of patients that do not capture many aspects of human biology. However, another naturally arising possibility is that other cellular phenotypes, such as migration or invasion, may be better indices of tumor response in patients than cellular proliferation. Addressing

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, 20742, USA. [2]Present address: Cancer Data Science Lab, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, USA. [3]Department of Biostatistics and Computational Biology, Harvard School of Public Health, Boston, USA. [4]Massachusetts General Hospital Cancer Center, Harvard Medical School, Boston, USA. [5]Division of Drug Discovery and Safety, LACDR, Leiden University, Leiden, The Netherlands. [6]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, M5G 1L7, Canada. [7]Present address: National Research Council Canada, Montreal, Canada. [8]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, 69978, Israel. [9]Laura and Isaac Perlmutter Cancer Centre, NYU-Langone Medical Center, New York City, NY, 10016, USA. [10]Alexandria Center for Life Science, New York, NY, 10016, USA. [11]The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, 69978, Israel. Nishanth Ulhas Nair and Avinash Das contributed equally. Correspondence and requests for materials should be addressed to E.R. (email: eytan.ruppin@nih.gov)

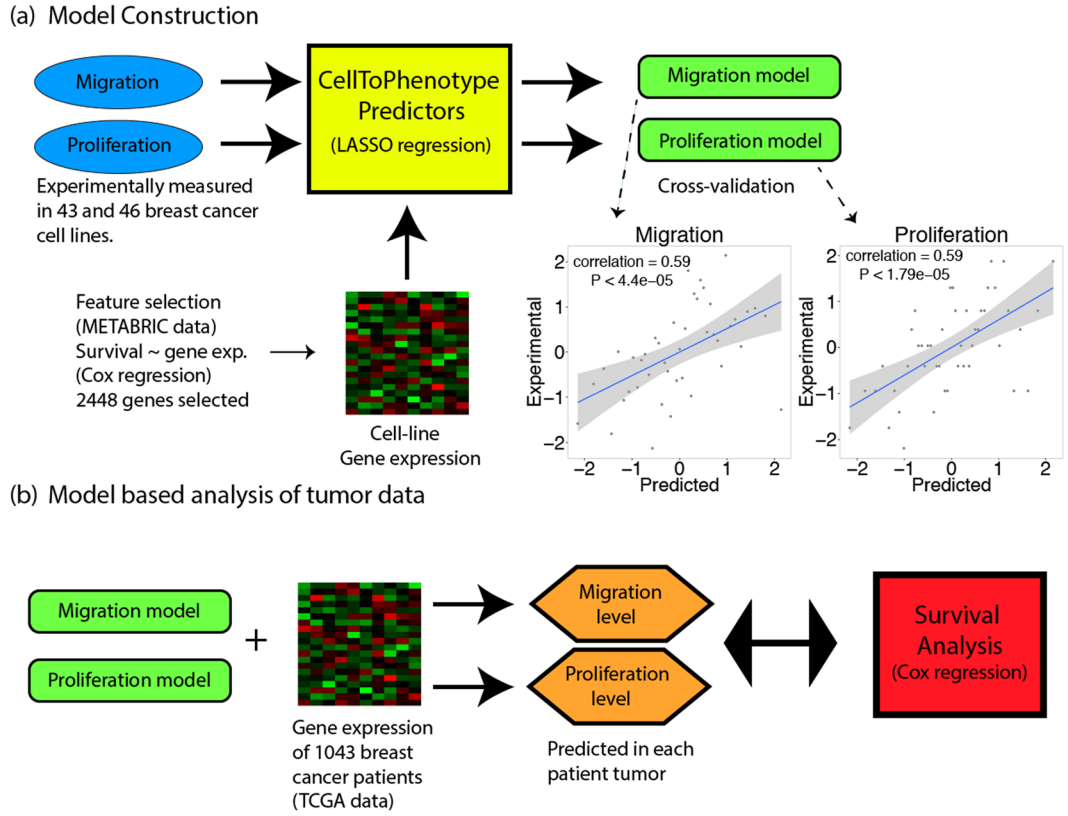

Figure 1. Overview of the method. **(a)** CellToPhenotype predictors of migration and proliferation from gene expression are constructed from experimentally determined migration and proliferation measurements across 43 and 46 breast cancer cell lines respectively. The predictors are built using cross-validation, and the correlations obtained between predicted levels and actual experimentally measured values are depicted as scatter plots. **(b)** The CellToPhenotype predictors are used to analyze the gene expression values of breast cancer patients to predict migration and proliferation levels of 1043 TCGA breast cancer tumors. Subsequently, the association of tumors predicted migration and proliferation levels with different tumor phenotypes and patients' survival is examined.

this question, we aimed here to quantify the relative weight of proliferation versus migration in determining cancer aggressiveness and patient survival.

Ideally, one would have liked to directly measure proliferation and migration levels directly in tumors *in vivo* to study their association with patient survival and treatment response. However, regrettably, such measurements are yet infeasible. We, therefore, set out to build and validate predictors of proliferation and migration levels in breast cancer cell-lines based on gene expression information. With such predictors, we applied them to predict the proliferation and migration levels of TCGA[14] breast cancer patient samples from their gene expression. We tested and verified that the predicted levels of these phenotypes in the tumors are indeed strongly associated with patient survival in the direction expected and that they are associated with cancer aggressiveness as expected. We find that migration is more strongly associated with breast cancer aggressiveness and more importantly, patient survival, than proliferation.

Results

Overview. We built predictors of cell proliferation and migration as follows: First, we experimentally measured migration and proliferation values in 43 and 46 breast cancer cell lines respectively (Table S1a,b, see Data/code availability section). Second, we constructed gene-expression based predictors of migration and proliferation, termed *CellToPhenotype* predictors, using least absolute shrinkage and selection operator (LASSO) based regression[15]. The predictors were tested on the cell-line data using a standard cross-validation procedure (Table S1b). Third, we used the predictors built to estimate the migration and proliferation levels of 1043 breast cancer patients in the TCGA data (Table S1c). Finally, we explored their importance in predicting tumor stage, grade, subtypes, and patients' survival (see Fig. 1 for an overview).

CellToPhenotype predictors (Supplementary Note) consists of two expressions based supervised regression – one for predicting cell migration and other for predicting cell proliferation. Each predictor was trained on *in vitro* cell migration or proliferation as the dependent variable and gene expression of cell lines as the independent variables in the regression. There are two level feature selections to reduce testing error. First, 2448 genes that are significantly associated (FDR < 0.01 using Cox regression) with patient survival (in an independent dataset – METABRIC)[16]. Second, CellToPhenotype uses LASSO regressor to regularize the predictor that enables

a data-driven feature selection using a cross-validation. A five-fold cross validation procedure to compute the minimum λ value for LASSO. Applying the predictor learned above to the gene expression of breast cancer tumor samples, we can predict migratory and proliferation level for each sample/individual. For each test sample, we iterate this procedure 50 times (to obtain a robust estimate) and take the median value of the migration and proliferation levels as the final estimates.

Experimental measurements of migration and proliferation. Doubling times for 46 breast cancer cell lines were estimated by plating a known number of cells and measuring the total number of cells once the culture reached an estimated 80% confluency. Proliferation rates were then calculated using the doubling time measurements. Cell line migration was estimated in 43 breast cancer cell lines using a live cell imaging-based random cell migration assay. The mean speed of cell migration was then quantified as the final migration estimate.

***CellToPhenotype* predictor construction and validation.** We measured migration and proliferation values in a collection of breast cancer cell lines available to us, for which we also had the transcriptomics data of each cell-line (Table S1a). With these data, we constructed predictors of migration and proliferation, termed *CellToPhenotype* predictors, which given the expression of a given cell-line, predict its migration and proliferation levels. These predictors were constructed using least absolute shrinkage and selection operator (LASSO) based regression[15], considering as features the genes whose expression is significantly associated with survival in the METABRIC breast cancer collection (Supplementary Note, Table S1d). The CellToPhenotype predictors accurately estimate cell-line migration (Spearman $\rho = 0.59$, $P < 4.39e\text{-}5$) and proliferation values (Spearman $\rho = 0.59$, $P < 1.79e\text{-}5$) using a standard cross validation procedure (Fig. 1a). The genes selected as the features used by the CellToPhenotype proliferation and migration predictors are shown in Table S3(a–d). These features sets are enriched for RAC1 signaling pathway (RAC1 is associated with cell motility)[17], immune response and cell apoptosis (Table S3(e,f)) in the migration predictor. The gene features of the proliferation predictor are enriched in cell differentiation, promoter transcriptional regulation and tissue development (Table S3(g,h)). A KEGG pathway analysis of these genes shows enrichment in cancer-related pathways known to be involved in migration and proliferation. These include HIF-1 signaling and ECM receptor interaction for migration and ErbB signaling and transcriptional misregulation for proliferation (Table S3i).

Next, we compared CellToPhenotype predictions with the expression of three reported gene markers of migration and proliferation (expression data obtained from CCLE project)[18]. First, the expression of Ki-67, a known marker of cell proliferation and patient survival[19–21]. Its expression is correlated with the experimental cell line proliferation measurements is significant but weaker (Spearman $\rho = 0.32$, $P < 0.03$) than that obtained by the CellToPhenotype predictor. Second, as a control, MIB-1 expression, a marker of tumor cell proliferation and determinant of patient survival in prostate cancer[22], shows no correlation with experimental measured proliferation in breast cancer cell lines (Spearman $\rho = 0.05$, $P < 0.76$). Finally, TPX2 expression, a well-known marker of migration in breast cancer[23], is significantly correlated with the experimental measurements of migration in breast cancer cell lines (Spearman $\rho = 0.4$, $P < 0.008$). Again, this correlation is smaller compared to correlation via the CellToPhenotype migration predictor. Overall, these results show that the latter provides a better estimate of *in vitro* proliferation and migration than known marker genes. We also check that the data used in training the CellToPhenotype predictors do not suffer from multicollinearity (Supplementary Note).

Predicted migration and proliferation levels are significantly higher in tumor samples than in normal samples. We then applied the CellToPhenotype proliferation and migration predictors to analyze breast cancer TCGA tumor data. Given an input tumor sample, each predictor (migration or proliferation) receives as input the levels of expression of its feature genes in that sample and outputs the predicted migration or proliferation levels. We first tested if the CellToPhenotype predicted proliferation and migration levels are higher in the TCGA breast tumors than the matched adjacent normal tissues (analyzing 110 TCGA breast cancer patients for which such matched data exists). Reassuringly, the predicted migration and proliferation levels are significantly higher in the tumors than in the matched non-cancerous tissues (paired Wilcoxon rank-sum test[24], $P < 5.5e\text{-}20$ and $P < 4.4e\text{-}20$ respectively, Fig. 2a). Random linear combinations of survival-significant genes used as control predictors do not show any such differences either for migration or proliferation (paired Wilcoxon rank-sum test, $P < 0.6$ and $P < 0.47$ respectively, Supplementary Note).

Advanced stages of breast cancer have higher predicted migration and proliferation levels than early stages. We next tested if CellToPhenotype predicted migration and proliferation levels would be higher in advance-stage patients, as expected. 937 breast cancer patients have cancer stage information available in TCGA (Table S2a). Indeed, the predicted migration levels increased significantly from stage I to stage II (Wilcoxon rank sum test, $P < 9.3e\text{-}4$); and from stage II to stage III-IV ($P < 8.7e\text{-}3$). Predicted proliferation levels also increase from stage I to stage II ($P < 8.3e\text{-}4$) and from stage I to stage III-IV ($P < 5.8e\text{-}4$) (Fig. 2b). As a control, random linear combinations of survival-significant genes do not exhibit any significant increase in their levels with higher tumor stages (Supplementary Note).

Predicted migration and proliferation levels increase with cancer grade. We next tested if the predicted migration and proliferation increase with cancer grade. Tumor grade information was available for 1706 breast cancer patients in METABRIC (Grade information is absent in TCGA breast cancer patients; Table S2a). Predicted migration levels were indeed significantly higher in grade 3 patients when compared to grade 2 (Wilcoxon rank-sum $P < 1.8e\text{-}13$) and grade 1 patients ($P < 8.7e\text{-}06$). Similarly, the predicted proliferation levels increase significantly from grade 1 to grade 2 ($P < 1.7e\text{-}4$), and from grade 2 to grade 3 patients ($P < 9.9e\text{-}48$) (Fig. 2c). Random linear combinations of survival-significant genes do not exhibit any such association with tumor grade (Supplementary Note).

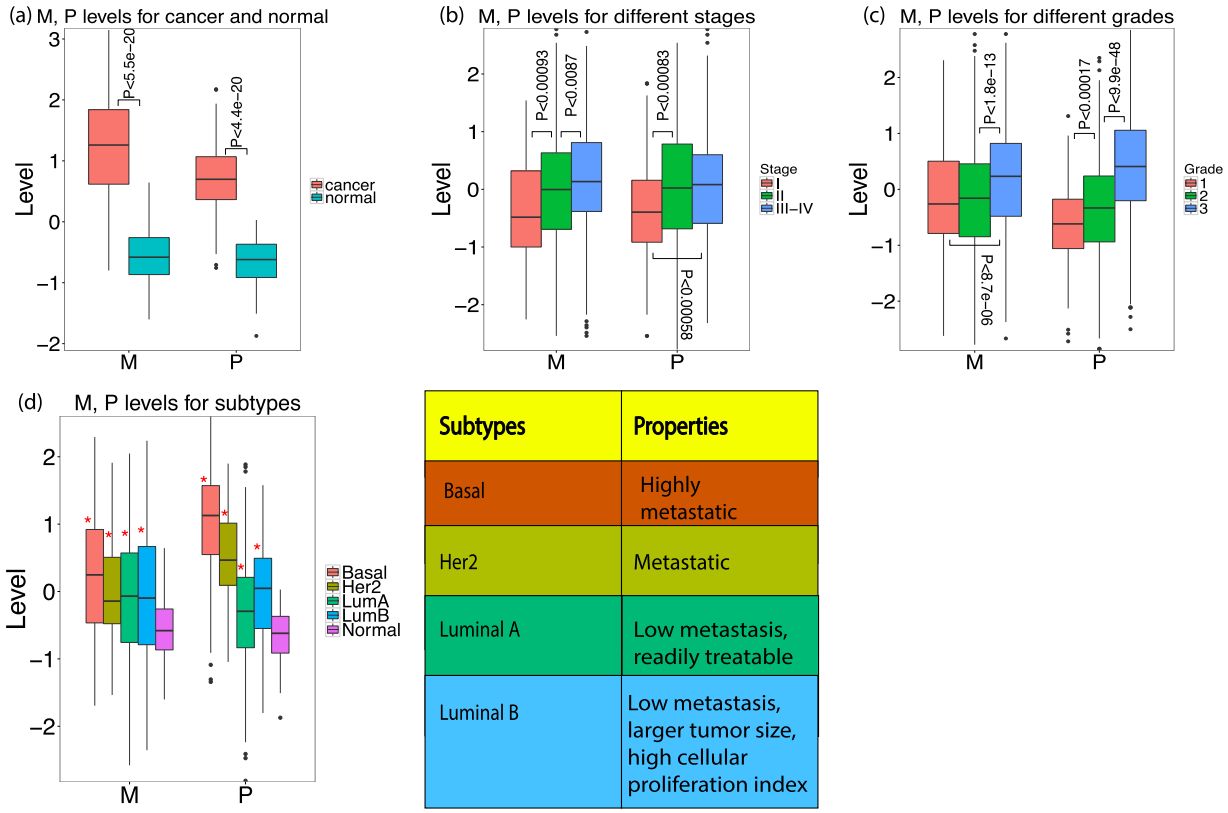

Figure 2. (**a**) Predicted migration (M) and proliferation (P) levels of breast cancer tumors and their association with various clinical phenotypes. (a) M, P levels of 110 breast cancer patients for their tumor (cancer) and matched non-cancerous breast samples (normal). **(b)** Predicted M, P levels for 937 breast cancer TCGA tumors for which cancer stage information is available. **(c)** Predicted M, P levels for 1706 METABRIC breast cancer patients for which cancer grade information is available. **(d)** Predicted M, P levels for 497 breast TCGA tumors dataset having subtype information: Basal or Triple-Negative (91 patients), Her2 (55 patients), Luminal A (LumA, 224 patients), Luminal B (LumB, 127 patients), and noncancerous samples (110 patients). The properties of these subtypes are shown in a table. Significant differences (when comparing tumors to non-cancerous samples) are marked via '*'.

Predicted migration and proliferation levels match known attributes of breast cancer subtypes. Different breast cancer subtypes have been associated with different migration and proliferation phenotypes. We therefore asked whether the predicted levels using CellToPhenotype predictors recapitulate these attributes. We find that all four subtypes of breast cancer (Table S2a; basal or triple negative breast cancer (TNBC), Her2, Luminal A and B) have significantly higher migration and proliferation levels than that of non-cancerous samples (n = 110) (Fig. 2d). Consistent with the observation that TNBC tumors are highly metastatic[25], we find that TNBC patients (n = 91) exhibit the highest predicted migration and proliferation levels amongst subtypes. Luminal A patients (n = 224) exhibit the lowest predicted proliferation amongst subtypes and significantly lower migration levels than TNBC (Supplementary Note), consistent with the observation that they have low metastasis levels and respond relatively well to treatment. Luminal B (n = 127) patients have higher predicted proliferation levels than Luminal A patients, consistent with the observation that Luminal B has larger tumor size and higher cellular proliferation index[25], and higher rates of lymph node involvement than Luminal A[26] (Fig. 2d). We also find similar results when we looked at both the experimentally-measured/predicted migration and proliferation values in breast cancer cell lines (Supplementary Fig. S3).

Predicted migration levels are more strongly associated with patient overall survival than predicted proliferation levels. We studied the association of predicted migration and proliferation levels with patients' overall survival in the TCGA breast cancer dataset (1043 patients). For this analysis, we first built new predictors of migration and proliferation by analyzing the 40 breast cancer cell lines for which we have both migration and proliferation measurements to enable a head-to-head comparison of the effects of migration and proliferation on survival, when built from exactly the same cell-lines (the construction itself followed the same predictor generation procedure described above). Given these CellToPhenotype predictors, we estimated the migration and proliferation levels of the 1043 TCGA breast tumors from their expression data, as before. We then employed Cox regression to examine the association between the predicted values and the patients' survival, controlling for various covariates including age, race, and genomic instability. We find a stronger association of predicted migration levels with patient survival (risk factor = 0.45, P < 2.06e-5) than the association between

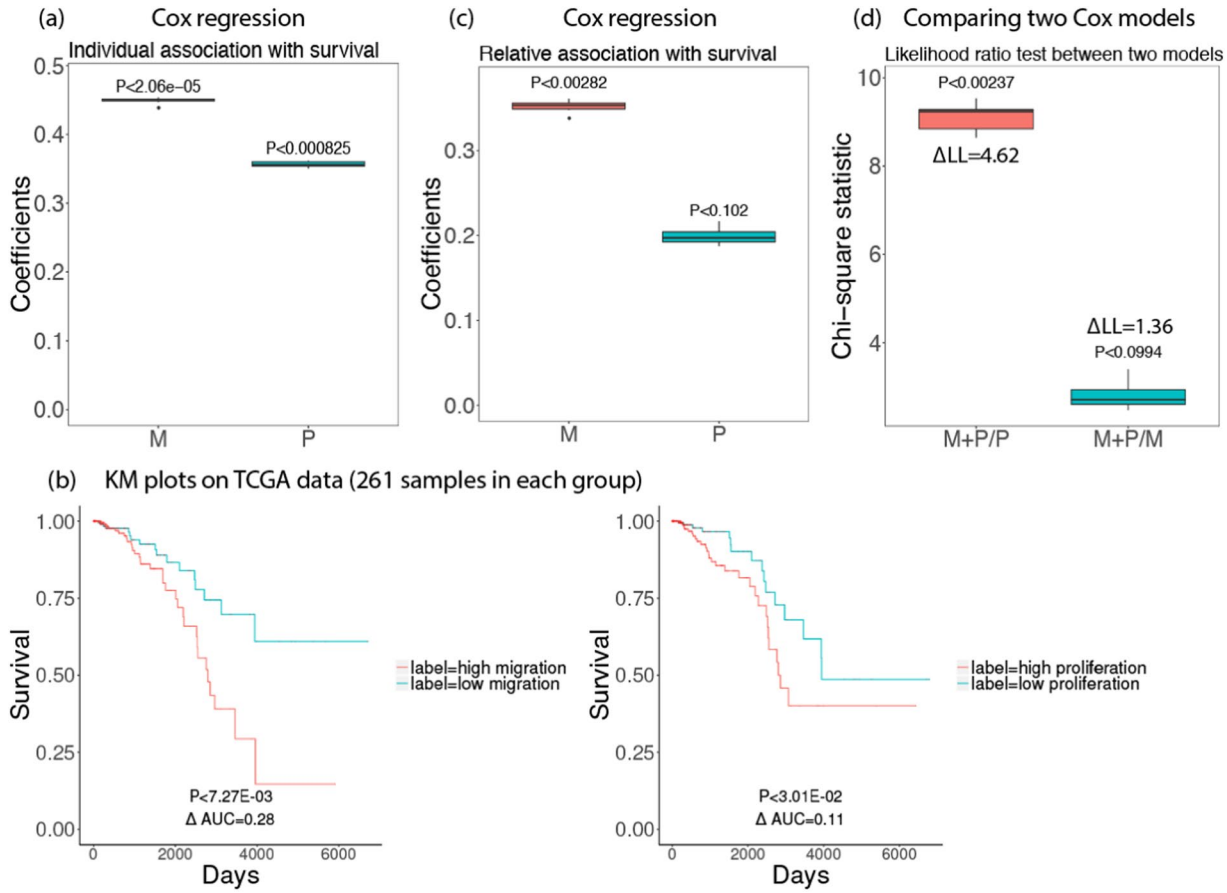

Figure 3. Survival analysis for 1043 breast cancer patients in TCGA data using predicted migration (M) and proliferation (P). Box plots of 10 iterations are shown and median p-value of each coefficient is given above the box plots. A positive coefficient (risk factor) for M (or P) indicates that the higher value of M (or P), the lower the patient survival. **(a)** Coefficients of M, P when used to predict survival individually using Cox regression after controlling for age, race, and genomic instability. **(b)** A Kaplan Meyer (KM) survival analysis of tumors' predicted migration and proliferation levels. Top 25 and bottom 25 percentile of the predicted migration/proliferation samples in each group were considered (261 samples in each group). **(c)** Relative coefficients of M, P when used to predict survival when they are controlled by each other (multivariate Cox-regression). **(d)** Likelihood ratio test comparing how significantly different are two Cox regression models with each other (Chi-square test statistics with p-values are provided): (i) Migration and Proliferation vs Proliferation only (M + P/P); (ii) Migration and Proliferation vs Migration only (M + P/M). The difference between log-likelihood (ΔLL) between the two models is also shown.

proliferation and survival (risk factor = 0.36, P < 8.25e-4) (Fig. 3a). In both cases, the higher the predicted migration/proliferation levels are, the worse is the patients' survival. A similar trend is revealed using a Kaplan Meier (KM) analysis (Kaplan and Meier, 1958) comparing tumors with high vs low predicted migration and proliferation levels (Fig. 3b). Analysis with multivariate Cox-regression (Fig. 3c) again shows that migration is more strongly associated with patient survival (relative risk factor = 0.35, P < 2.82e-3) than proliferation (relative risk factor = 0.20, P < 0.102). Random linear combinations of survival-significant genes do not exhibit any significant associations with patient survival (risk factor = 0.067, P < 0.24 for migration; risk factor = −0.0021, P < 0.24, for proliferation; Supplementary Note).

To estimate the significance of the difference of predictability, we additionally performed a likelihood ratio test comparing the survival predictive power of combined migration and proliferation compared to only proliferation or only migration. We see that the migration + proliferation model is significantly better than a proliferation only model (Chi-square statistic = 9.24, P < 2.37e-3) but not significantly better from a migration only model (Chi-square statistic = 2.72, P < 0.099). Thus, adding migration to a proliferation-only model improves the survival prediction, however adding proliferation does not add significant predictive power to a migration-only model., A migration only model is significantly better than a proliferation only model (Chi-square statistic = 6.40, P < 2.2e-16) in predicting patient survival (Fig. 3d). Finally, we observed that survival prediction accuracy was considerably reduced if we use a smaller number of cell lines for building the CellToPhenotype predictors (Supplementary Note), showing the importance of studying a large number of cell lines.

A siRNA-based analysis further supports that migration is more strongly associated with survival than proliferation. We next turned to study our basic research question by building an additional set of predictors of migration and proliferation levels. These predictors are based on siRNA knockdown (KD) experiments that we have conducted in a highly migratory breast cancer cell line, MDA-MB-231. Keeping translational goals in mind, we considered around 4600 druggable proteins. Most of these genes come from protein kinases and GPCRs (G-protein-coupled receptors). (Kinases which are often over expressed and activated in cancer are easily targetable through their ATP binding sites.) Among these, we knocked down 248 proteins whose gene expression levels are significantly negatively correlated with experimentally determined migration values in the 40 breast cancer cell lines that we studied above. We experimentally measured the effect of each knockdown on cell migration using a 2D migration assay. We termed the genes whose knockdown significantly enhances cellular migration *migration-suppressive genes*. Similarly, we also carried out siRNA KD experiments on 227 proteins whose gene expression levels are significantly positively correlated with experimentally determined migration values in these 40 breast cancer cell lines, and determined the effect of each knockdown on cell migration. Among these, we identified all genes whose knockdown significantly decreased cellular migration in MDA-MB-231 cell line and termed them as *migration-enhancer* genes. Migration suppressive genes (using the siRNA-based analysis, Table S3j) show enrichment in gene sets involved in metastasis and breast cancer and cell migration (Table S3k). Migration-enhancer genes (Table S3j) show enrichment in the actin cytoskeleton, focal adhesion and cell-cell junctions (Table S3(i,m)). Gene sets like actin cytoskeleton, focal adhesions and cell-cell junctions are known to be involved in cell migration[27–29].

The number of migration-suppressive genes that had low expression were downregulated in a given breast cancer cell-line (S-count) was highly correlated with its CellToPhenotype predicted migration levels (Spearman $\rho = 0.81$, $P < 3.43e\text{-}10$) and also with its experimentally measured migration values (Spearman $\rho = 0.79$, $P < 1.83e\text{-}9$). This suggests that the S-count could be considered as an approximation of cellular migration. Similarly, the number of migration-enhancer genes that had high expression in a given breast cancer cell-line denotes their migration-enhancer scores (E-count). The E-score is also highly correlated with the predicted migration levels (Spearman $\rho = 0.827$, $P < 8.62e\text{-}11$) and the experimentally measured migration values (Spearman $\rho = 0.77$, $P < 1.03e\text{-}8$), suggesting that E-count also provide approximate estimates of cellular migration. The sum of S-count and E-counts, termed the KD-migration-score, has a slightly higher correlation with the predicted migration levels (Spearman $\rho = 0.83$, $P < 4.36e\text{-}11$, Supplementary Fig. S1a) and experimentally measured ones (Spearman $\rho = 0.79$, $P < 1.9e\text{-}9$, Supplementary Fig. S1b) across the cell-lines.

Similarly, in an analogous manner, we computed a KD-proliferation-score, using published shRNA/siRNA knockdown data done in MDA-MB-231 cell line[30]. The KD-proliferation-scores are highly correlated with both the predicted proliferation levels (Spearman $\rho = 0.75$, $P < 3.07e\text{-}8$, Supplementary Fig. S1c) and the experimentally measured proliferation values (Spearman $\rho = 0.82$, $P < 2.57e\text{-}10$, Supplementary Fig. S1d). Reassuringly, we find that the cross-correlations between KD-migration-score and experimentally-measured proliferation levels (and vice-versa) are much lower (Supplementary Note). Gene Set Enrichment Analysis (GSEA) analysis on proliferation-enhancer/suppressive genes, showed enrichment in relevant gene sets, including, cell proliferation, regulation of developmental processes, genes associated with breast cancer (Table S3(n,o)) – gene sets which are known to play a role in proliferation and breast cancer.

Having these scores in hand, we next computed *KD-migration-scores* and *KD-proliferation-scores* for every TCGA breast cancer tumor. Reassuringly, these scores are significantly correlated with the CellToPhenotype predictions of migration (Spearman $\rho = 0.4$, $P < 5.95–42$) and proliferation levels of these tumors (Spearman $\rho = 0.69$, $P < 2.77–150$). We then examined the association of the *KD-migration* and *KD-proliferation scores* of the TCGA breast cancer tumors and patient survival, after controlling for age, race, and genomic instability via Cox regression. The results reinforce the trend observed previously with the CellToPhenotype analysis, as we find a significant association of *KD-migration-scores* with patient survival (risk factor $= 0.3$, $P < 2.98e\text{-}3$) but a lower association between *KD-proliferation-scores* and survival (risk factor $= 0.24$, $P < 0.0396$) (Fig. 4a). Analysis with multivariate cox-regression using both the *KD-migration-scores* and *KD-proliferation-scores* as covariates while controlling for age, race, and genomic instability show a similar trend (relative risk factor $= 0.29$, $P < 0.037$ for migration and relative risk factor $= 0.044$, $P < 0.76$ for proliferation, Fig. 4b). Thus, ruling out model-based biases of CellToPhenotype predictors, the analysis further corroborates our findings that migration is better predictor patient survival than proliferation. As it is knock-down based, it suggests that the stronger association of migration with survival may have a causal basis.

CellToPhenotype estimates of migration and proliferation levels are associated with patient drug response. We next asked if predicted migration or proliferation levels determine patient response to cytoskeletal vs cytotoxic drugs. Although migration and proliferation occur via partially overlapping cellular processes, as a first approximation cytotoxic drugs mainly target proliferation by inhibiting nucleotide synthesis or inducing DNA break while cytoskeletal drugs are known to affect cell migration by targeting microtubules (though they also target proliferation)[31,32]. Accordingly, our working hypothesis has been that tumors with high migratory estimates may respond better to cytoskeletal drugs and conversely, tumors with high proliferation estimates will respond better to cytotoxic drugs.

To test this hypothesis, we analyzed data of TCGA breast cancer patients, out of which 389 patients were given at least one of the 7 cytoskeletal drugs and 331 patients were given at least one of 31 cytotoxic drugs (Table S2b). In the patients treated with cytoskeletal drugs, migration levels (estimated from CellToPhenotype predictors) are significantly higher than the levels in the rest of TCGA breast cancer patients ($P < 6.73e\text{-}5$, Fig. 4c), and higher than the levels of the patients treated only with cytotoxic drugs ($P < 4.9e\text{-}5$, Supplementary Fig. S2). Importantly, among the patients with high predicted migration levels, the patients treated with cytoskeletal drugs have better survival than those that were not treated with these drugs (KM Δ AUC $= -0.47$, log rank $P < 4.38e\text{-}4$, Fig. 4d).

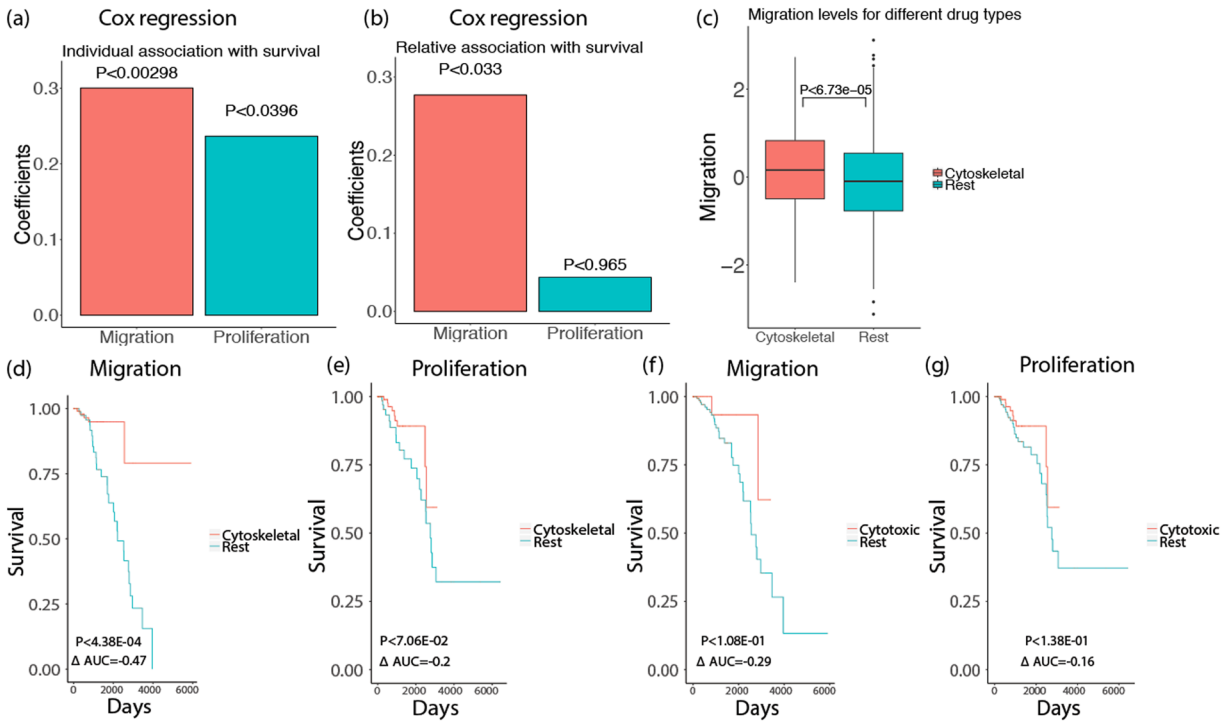

Figure 4. (**a**) Cox regression of KD-migration-scores and KD-proliferation-scores with patients' survival, after controlling for age, race, and genomic instability for 1043 breast cancer patients in TCGA data. **(b)** Relative association of KD-migration-scores and KD-proliferation-scores with survival. **(c)** Migration levels (estimated from CellToPhenotype predictors) for breast cancer patients who have taken cytoskeletal drugs versus the rest. **(d)** Among the breast cancer patients who have high migration levels (greater than 75 percentile), a KM analysis was done between those who have taken cytoskeletal drugs versus the rest of the patients. **(e)** Similarly, among the patients that have high proliferation levels, a KM analysis was done between those who have taken cytoskeletal drugs versus the rest of the patients. **(f)** A KM analysis of patients who have high migration levels and who have taken only cytotoxic drugs versus the rest. **(g)** A KM analysis of patients who have high proliferation levels and have taken only cytotoxic drugs versus the rest.

Thus, cytoskeletal drugs are more likely administered to patients with high (predicted) migration, and such administration seems to preferentially benefit such patients. This suggests that the predicted migration levels may serve as a biomarker for better response to cytoskeletal drugs. As controls, patients with high predicted proliferation levels do not have a significant survival benefit from taking cytoskeletal drugs (KM Δ AUC $= -0.2$, log rank P $<$ 7.06e-2, Fig. 4e). Finally, and interestingly, neither migration nor proliferation is a good predictor of cytotoxic efficacy, as predicted migratory and proliferation levels are not associated with the survival of patients who have taken only cytotoxic drugs: (KM Δ AUC $= -0.29$, log rank P $<$ 1.08e-1, Fig. 4f) for patients with high predicted migratory levels and (KM Δ AUC $= -0.16$, log rank P $<$ 1.38e-1, Fig. 4g) for patients with high proliferation levels.

Discussion

The central focus of our work is that among proliferation and migration assays–routinely used in the experimental laboratories – we found that the migration assays capture components of clinical parameters, including patient survival, which are not captured as well by proliferation assays. This suggests that conducting migration assays in addition to proliferation assays could aid the success of laboratory findings in clinics. To show this, we first experimentally measured migration and proliferation values in over 40 breast cancer cell lines, and built and cross-validated CellToPhenotype predictors of these phenotypes. Applying these predictors to the gene expression data of breast cancer patients, we predicted the migration and proliferation levels of every tumor and studied their association with cancer stage, grade, subtype, and most importantly, with patient survival. We also find some evidence (based on a small number of samples) that the predicted migration levels are higher in circulating tumor cells than in normal samples or TCGA cancer samples (Supplementary Fig. S6). We find a stronger association of predicted migration levels with the patients' survival compared to the predicted proliferation levels. Controlling for tumor purity[33] in Cox regression (along with age, race, genomic instability) gives similar results as before, i.e. migration is a stronger predictor of breast cancer patient survival than proliferation (risk factor $= 0.44$, P $<$ 4.55e-05 for migration; risk factor $= 0.39$, P < 0.00041 for proliferation). We also find that patients with high predicted migration levels respond better to cytoskeletal drugs than patients with low predicted levels. siRNA-based predictors of migration and proliferation that we additionally built further testify that migration is indeed a better predictor of survival and the association may even have a causal basis. The choice of MDA-MB-231 cell-line (for

the siRNA-based predictor) is unlikely to bias this conclusion because MDA-MB-231 is both highly proliferative and migratory cell line (MDA-MB-231 has the highest proliferation rate and the third highest migration rate among all the breast cancer cell lines used), and since the expression levels of the selected genes (features) are significantly correlated with the experimentally measured migration values in the 40 breast cancer cell lines. Our conclusion about the relative importance of migration over proliferation is robust to a few different migration like wound-healing assays (explained in detail in the Supplementary Note, Supplementary Figs S3, S4). To the best of our knowledge, this is the first study which aims to quantify migration and proliferation levels in cancer patients by collecting and analyzing pertaining *in vitro* data. Such an investigation is particularly relevant since the majority of cancer drugs are developed by measuring their effect on *in vitro* proliferation rates[10]. However, further studies are required to validate our findings in other *in vitro* and *in vivo* model systems and to elucidate the underlying mechanisms.

Many of the genes identified as migratory signatures by CellToPhenotype predictors are known to play a role in cell migration. For instance, the LAMA2 gene produces an extracellular protein, Laminin, is thought to play a role in migration and organization of cells in tissues during embryonic development[34]. CX_3CR1 is known to play a role in adhesion and migration of leukocytes[35]. CX_3CR1 is also amongst the top hits in our siRNA-based. RNT4 is another gene that is associated with cell migration[36,37]. Our migration and proliferation signatures (either CellToPhenotype or siRNA-based analysis) identified many additional genes (Table S3p) that may be important in migration/proliferation, and their investigation may provide future leads for enhancing our understanding of these cellular phenotypes.

There is no clear distinction between cytoskeletal and cytotoxic drugs, and their effects on migration and proliferation. Cytoskeletal drugs are also cytotoxic in nature, that is, they also target cell proliferation. Also, most patients who have taken a cytoskeletal drug have also taken another cytotoxic drug, and many patients have taken more than one cytotoxic drug but no cytoskeletal drug (Table S3q). Our drug-response analysis broadly suggests that drugs targeting both migration and proliferation seem to give an additional treatment benefit over drugs that only target proliferation.

In summary, our analysis highlights the importance of tumor migration in determining its aggressiveness and patients' survival. It puts forward the need to put more effort on *in vitro* assays of cell migration (and possibly, invasion) in the early stages of cancer drug development, which currently focusses on proliferation screens.

Methods

Quantification of cell migration and proliferation of a panel of breast cancer cell lines. *Doubling times.* Breast cancer cell line population doublings were estimated by plating a known number of cells at day 0 and measuring the total number of cells once the culture reached an estimated 80% confluency (usually 4–5 days). Cell numbers were calculated using a hemocytometer and population doublings (PDL) were determined using the following formula: PDL = 3.32 (log(total cells at harvest/total cells plated at day 0)). The doubling time was calculated by dividing the number of days or hours between harvest and seeding by the PDL. Example: Number of cells plated at day 0 = 5×10^6, Number of cells at harvest = 20×10^6, days in culture = 4 days, then the doubling time would be 2 days or 48 hours (4 days/2 PDL). The proliferation rate for each cell lines was computed using the equation 70/(doubling time). We did this for 46 breast cancer cell lines.

Live Cell imaging-based random cell migration assay. Cells were seeded on 96-well glass bottom plates (Greiner Bio-one, Monroe, NC, USA) coated with 10 µg/ml collagen type I (isolated from rat tails) in PBS for 1 hour at 37 °C[38]. Before imaging, the cells were pre-exposed for 30–45 min to 0.1 µg/ml Hoechst 33342 (Fisher Scientific, Hampton, NH, USA) to visualize the nuclei. The plates were placed on a Nikon Eclipse TE2000-E microscope fitted with a 37 °C incubation chamber and 5% CO2 supplier, a 20x objective (0.75 NA, 1.00 WD), an automated stage and perfect focus system. Up to four positions per well were automatically defined and nuclei (stained with live Hoechst) were imaged overnight every 10 to 20 minutes using NIS controlling software (Nikon) and a CCD camera (Pixel size: 0.78 or 0.32 µm). The.nd2 files acquired from NIS were exported to.tiff files as mono image for each channel and then converted to.avi files and analyzed using custom made ImagePro Plus macros as previously described[39]. The mean speed of cell migration was quantified per time-lapse by tracking each nucleus separately over time. This was done for 43 breast cancer cell lines.

siRNA-image based migration assay using the MDA-MB-231 cell line. *Transient siRNA-mediated gene knockdown.* Human siRNA of 475 proteins (whose gene expressions are significantly correlated with experimentally determined migration values in 40 breast cancer cell lines studied above, Table S2c) were purchased in siGENOME format from Dharmacon (Dharmacon, Lafayette, CO, USA)[40]. 'Transient siRNA knockdown was achieved by reverse transfection of 50 nM single or SMARTpool siRNA in 2,500-5,000 cells/well in a 96-well plate format (PKT assay) using the transfection reagent INTERFERin (Polyplus, Illkirch, France) according to the manufacturer's guidelines. The medium was refreshed after 20 h and transfected cells were used for various assays between 65 to 72 h after transfection'[40]. (Some of the above text in this paragraph has been similarly described in Fokkelman *et al.*[40]).

Phagokinetic track (PKT) assay. 'PKT assays were performed as described before[41]. Briefly, black 96-well µClear plates (Greiner Bio-One, Frickenhausen, Germany) were coated with 10 µg/ml fibronectin (Sigma-Aldrich, Zwijndrecht, The Netherlands) for 1 h at 37 °C. Plates were washed twice with PBS, using a HydroFlex plate-washer (Tecan, Männedorf, Switzerland). Subsequently, the plates were coated with white carboxylate modified latex beads (400 nm, $3.25 \cdot 10^9$ particles per well; Life Technologies, Carlsbad, CA, USA) for 1 h at 37 °C, after which the plate was washed 7 times with PBS. 65 h after siRNA transfection, transfected cells were washed twice with PBS-EDTA and trypsinized. Cells were resuspended into single cell suspensions, then diluted, and finally seeded

at low density (~100 cells/well) in the beads-coated plate. Cells were allowed to migrate for 7 h, after which the cells were fixed for 10 min with 4% formaldehyde and washed twice with PBS. For each transfection, duplicate bead plates were generated (technical replicates); transfection of each siRNA library was also performed in duplicate (independent biological replicate). Procedures for transfection, medium refreshment and PKT assay were optimized for laboratory automation by a liquid-handling robot (BioMek FX, Beckman Coulter)'[40]. (The above text in this paragraph has been similarly described in Fokkelman *et al.*[40]).

PKT imaging and analysis. 'Migratory tracks were visualized by acquiring whole well montages (6 × 6 images) on a BD Pathway 855 BioImager (BD Biosciences, Franklin Lakes, NJ, USA) using transmitted light and a 10x objective (0.40 NA). A Twister II robotic microplate handler (Caliper Life Sciences, Hopkinton, MA, USA) was used for automated imaging of multiple plates. Montages were analyzed using WIS PhagoTracker. Migratory tracks without cells or with more than 1 cell were excluded during image analysis. The quantitative output of PhagoTracker was further analyzed using KNIME. Wells with <10 accepted tracks were excluded. Next, data was normalized to mock to obtain a robust Z-score for each treatment and each parameter. After normalization, an average Z-score of the 4 replicates was calculated. Knockdowns with <3 images were removed, as well as knockdowns with <150 accepted tracks'[40]. (The above text in this paragraph has been similarly described in Fokkelman *et al.*[40]). Major Axis score (Z-score) as a measure of cell speed was further used and computed in the modeling.

CellToPhenotype predictors. CellToPhenotype predictors consists of two expressions based supervised predictors – one for predicting cell proliferation and other for cell migration. The gene expression data was obtained from the Cancer Cell Line Encyclopedia project[18]. Each predictor was trained on *in vitro* cell migration or proliferation as the dependent variable and gene expression of cell lines as the independent variables in the regression. CellToPhenotype adopts two level feature selection to reduce testing error. First genes that are significantly associated with patient survival (in an independent dataset – METABRIC) were selected to be included in the subsequent regression model. Secondly, CellToPhenotype uses LASSO shrinkage to regularize the predictor that enables a data-driven feature selection using a cross-validation. Both above feature selection was conducted in dataset independent of the testing set on which performance of CellToPhenotye was evaluated. This ensures an unbiased evaluation of the predictive power of CellToPhenotype.

To achieve the robust final estimate of migration and proliferation, CellToPhenotype uses bootstrapping. CellToPhenotype predictors were conducted on the training data and then phenotypes were predicted for test samples. The process is repeated for 50 bootstraps. Median of each bootstrap are taken as the final estimates of migration and proliferation levels (details are provided in the Supplementary Note).

CellToPhenotype predictive performance using cross-validation. Leave-one-out cross validation was conducted to assess the CellToPhenotype predictive power as follows. The migration and proliferation models were trained on all *in vitro* data leaving one sample. Migration and proliferation was estimated for the left-out sample. Spearman ρ between the predicted and actual phenotypes was computed.

Association of cellular phenotypes with cancer stages. Using CellToPhenotype predictors, migration and proliferation levels were predicted for 937 breast cancer patients in TCGA (Table S2a). Out of this, 83 and 9 individuals are stage IA and stage IB respectively (grouped as stage I); 348 and 238 individuals are stages IIA and IIB respectively (grouped as stage II); 148, 29, 62, and 19 individuals in stages IIIA, IIIB, IIIC, and IV respectively (grouped as stage III-IV). Patients with stage III and IV were grouped because there is only 19 sample from stage IV. These predicted levels were used to check how they vary with stages.

Association of cellular phenotypes with cancer grade. Using CellToPhenotype predictors, migration and proliferation levels were predicted for 1706 breast cancer patients whose cancer grade information was available in the METABRIC dataset (146 grade 1 patients, 673 grade 2 patients, 887 grade 3 patients, Table S2a). These predicted levels were then used to check how they vary with grade.

Association of cellular phenotypes with breast cancer subtypes. Migration and Proliferation levels are predicted for the 497 breast cancer patients (Table S2a) in TCGA dataset for whom we have the four different breast cancer subtypes information available, and of the 110 normal non-cancerous breast samples. A one-sided Wilcoxon rank-sum test was used to compare migration and proliferation levels of each of the subtypes with that of the normal samples.

Association of cellular phenotypes with patient survival. Migration and proliferation models were built by training on 40 breast cancer cell lines that have experimentally measured migration and proliferation. Migration and proliferation levels of 1043 TCGA breast cancer patients were estimated using CellToPhenotype predictors. To check the association of the predicted migration with patients' survival we fit following Cox regression:

$$\text{Survival} \sim \text{migration} + \text{strata(race)} + \text{age} + \text{GII}. \quad (1)$$

Patient survival is known to be confounded by age, race, and genomic instability (GII). Accordingly, the above model systematically controls for these confounders. Strata (race) in the above model implies Cox regression was conducted in each patient stratification based on race separately and likelihood were combined. We repeated the procedure for 10 iterations, and median coefficients (risk factor) of migration were computed. The association of survival with proliferation was estimated similarly. Each Kaplan Myer (KM) analysis was done by comparing the migration/proliferation levels of on the top 25 percentile of patients with bottom 25 percentile patients.

To estimate the relative contribution of migration and proliferation to predict patient survival we fit following Cox regression, which also controls for age, race, and genomic instability:

$$\text{Survival} \sim \text{migration} + \text{proliferation} + \text{strata(race)} + \text{age} + \text{GII} \quad (2)$$

siRNA-based KD-migration-score of a sample. Out of around ~4600 druggable genes that we considered, we selected 475 genes whose expression was significantly correlated with migration across 40 cell lines (Spearman ρ, $P < 0.01$). Out of this, 248 genes are negatively correlated and 227 genes are positively correlated. We conducted siRNA knockdown of 475 genes described above in MDA-MB-231 breast cancer cell line. MDA-MB-231 was chosen because it is highly migratory breast cancer cell line[42]. Following the siRNA, we experimental measured change cell migration by measuring factors including Major Axis score (Z-score). Migration-suppressive genes ($n = 26$) were identified by selecting genes whose knockdown led to high migration (above 90 percentile) in MDA-MB-231 cell line and whose gene expression was significantly negatively correlated with experimentally determined migration values in 40 cell lines. Migration-enhancer genes ($n = 24$) are those whose knockdown significantly decreased cellular migration (below 10 percentile) and whose gene expression was significantly positively correlated with experimentally determined migration values in 40 cell lines. We count the number of lowly expressed (below 50 percentile of the entire expression dataset – for all cell lines and genes) migration-suppressive genes and highly expressed (above 50 percentile) upregulated migration-enhancer genes; and assign the count as KD-migration-score of each sample (cell line or patient). For this analysis, we use the median expression of all genes as the threshold for upregulation or downregulation.

siRNA/shRNA-based KD-proliferation-score of a sample. siRNA/shRNA-based KD-proliferation-score of a sample is determined in an analogous manner described for migration. Briefly, we used proliferation measurement post ~15400 genes shRNA knockout in MDA-MB-231 cell lines from Marcotte *et al.*[30]. Among these genes, we select 1248 genes whose expressions are significantly correlated with experimentally determined proliferation values in 40 BC cell lines (606 genes positively correlated and 642 genes negatively correlated). Proliferation-suppressive genes were identified by selecting genes whose knockdown led to high proliferation (above 90 percentile) in MDA-MB-231 cell line and whose gene expression was significantly negatively correlated with experimentally determined proliferation values in 40 cell lines. Proliferation-enhancer genes are those whose knockdown significantly decreased cellular proliferation (below 10 percentile) and whose gene expression was significantly positively correlated with experimentally determined proliferation values in 40 cell lines. The count of lowly expressed proliferation suppressive and highly expressed proliferation-enhancer genes in a sample was assigned as KD-proliferation-score of the sample.

Drug response analysis. Drug response information is available for 720 TCGA breast cancer patients in TCGA: 389 patients administering at least one of the 7 cytoskeletal drugs, and 331 patients administering at least one of the 31 drugs targeting only proliferation (i.e., cytotoxic drugs, Table S2b). Among the 1043 breast cancer patients who have high migration levels (top 25 percentile), we do a KM analysis among patients administering cytoskeletal drugs and the rest (Fig. 4d). compares top 25 percentile patient with high migration levels with the rest of the 1043 patients. Similar KM analysis was conducted on patient administering cytotoxic drugs.

Pathway and GSEA enrichment analysis. We also carried out GSEA analysis[43,44] on the genes selected by the LASSO regression in the CellToPhenotype predictors. The annotated gene sets from the Molecular Signature Database was used for this analysis[45]. GSEA analysis was done separately on these sets of genes. Enriched gene sets with FDR q-value < 0.05 is shown (Table S3). Similarly, we carried out GSEA analysis on *migration-enhancer/suppressive* and *proliferation-enhancer/suppressive* genes.

For each iteration of the LASSO regression in the CellToPhenotype predictors, we did KEGG pathway analysis. We also did KEGG pathway analysis on genes selected from the siRNA-based analysis.

Data/Code Availability

The experimentally measured proliferation values are added in Table S1b. The experimentally measured migration values are provided in https://drive.google.com/file/d/1AY3M0Nadt1B0InYZQ5XvrSwDTV5Sr2rl/view?usp=sharing (from Rogkoti *et al.*[38]). The siRNA-based analysis data based on MDA-MB-231 cell line is from Fokkelman *et al.*[40]. The R code for developing CellToPhenotype predictors is available as free software with the GNU General Public License in the GitHub repository (https://github.com/nishanth83/CellToPhenotype).

References

1. Kaitlin, K. I. The Landscape for Pharmaceutical Innovation: Drivers of Cost-Effective Clinical Research. *Pharm Outsourcing* 1–6 (2010).
2. Torjesen, I. Drug development: the journey of a medicine from lab to shelf. *Pharmaceutical Journal* (2015).
3. Arrowsmith, J. Trial watch: Phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.* **10**, 87–87 (2011).
4. Weinstein, J. N. *et al.* An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science (80-.).* **275**, 343–349 (1997).
5. Staunton, J. E. *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci.* **98**, 10787–10792 (2001).
6. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
7. Solit, D. B. *et al.* BRAF mutation predicts sensitivity to MEK inhibition. *Nature* **439**, 358–362 (2006).
8. Dry, J. R. *et al.* Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). *Cancer Res.* **70**, 2264–2273 (2010).
9. Greshock, J. *et al.* Molecular target class is predictive of *in vitro* response profile. *Cancer Res.* **70**, 3677–3686 (2010).
10. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in. *Cancer. Cell* **166**, 740–754 (2016).

11. Hackam, D. G. & Redelmeier, D. A. Translation of Research Evidence From Animals to Humans. *Jama* **296**, 1727–1732 (2006).
12. FDA & Food and Drug *Administration. Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. Review Literature And Arts Of The Americas* (2004).
13. Mak, I. W., Evaniew, N. & Ghert, M. Lost in translation: animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.* **6**, 114–8 (2014).
14. Weinstein, J. N. *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–20 (2013).
15. Tibshirani, R. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B* (1996).
16. Curtis, C. *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–52 (2012).
17. Parri, M. & Chiarugi, P. Rac and Rho GTPases in cancer cell motility control. *Cell Communication and Signaling* **8** (2010).
18. Barretina, J. *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
19. Li, L. T., Jiang, G., Chen, Q. & Zheng, J. N. Ki67 is a promising molecular target in the diagnosis of cancer (Review). *Molecular Medicine Reports* **11**, 1566–1572 (2015).
20. Hoos, A. *et al*. High Ki-67 proliferative index predicts disease specific survival in patients with high-risk soft tissue sarcomas. *Cancer* **92**, 869–874 (2001).
21. Scholzen, T. & Gerdes, J. The Ki-67 protein: From the known and the unknown. *Journal of Cellular Physiology* **182**, 311–322 (2000).
22. Borre, M., Bentzen, S. M., Nerstrom, B. & Overgaard, J. Tumor cell proliferation and survival in patients with prostate cancer followed expectantly. *J. Urol.* **159**, 1609–1614 (1998).
23. Yang, Y. *et al*. TPX2 promotes migration and invasion of human breast cancer cells. *Asian Pac. J. Trop. Med.* **8**, 1064–70 (2015).
24. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **18**, 50–60 (1947).
25. Chikarmane, S. A., Tirumani, S. H., Howard, S. A., Jagannathan, J. P. & Dipiro, P. J. Metastatic patterns of breast cancer subtypes: What radiologists should know in the era of personalized cancer medicine. *Clinical Radiology* **70**, 1–10 (2015).
26. Inic, Z. *et al*. Difference between Luminal A and Luminal B subtypes according to Ki-67, tumor size, and progesterone receptor negativity providing prognostic information. *Clin. Med. Insights Oncol.* **8**, 107–111 (2014).
27. Yamaguchi, H. & Condeelis, J. Regulation of the actin cytoskeleton in cancer cell migration and invasion. *Biochimica et Biophysica Acta - Molecular Cell Research* (2007).
28. Friedl, P. & Mayor, R. Tuning collective cell migration by cell-cell junction regulation. *Cold Spring Harb. Perspect. Biol.* (2017).
29. Nagano, M., Hoshino, D., Koshikawa, N., Akizawa, T. & Seiki, M. Turnover of focal adhesions and cancer cell migration. *International Journal of Cell Biology* (2012).
30. Marcotte, R. *et al*. Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance. *Cell* **164**, 293–309 (2016).
31. Small, J. V., Geiger, B., Kaverina, I. & Bershadsky, A. How do microtubules guide migrating cells? *Nat. Rev. Mol. Cell Biol.* **3**, 957–64 (2002).
32. Yang, H., Ganguly, A. & Cabral, F. Inhibition of cell migration and cell division correlates with distinct effects of microtubule inhibiting drugs. *J. Biol. Chem.* **285**, 32242–32250 (2010).
33. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* (2015).
34. Moran, T., Gat, Y. & Fass, D. Laminin L4 domain structure resembles adhesion modules in ephrin receptor and other transmembrane glycoproteins. *FEBS J.* **282**, 2746–2757 (2015).
35. Imai, T. *et al*. Identification and molecular characterization of fractalkine receptor CX3CR1, which mediates both leukocyte migration and adhesion. *Cell* **91**, 521–530 (1997).
36. Acevedo, L. *et al*. A new role for Nogo as a regulator of vascular remodeling. *Nat. Med.* **10**, 382–388 (2004).
37. Hui, L. *et al*. RNT4 3′-UTR insertion/deletion polymorphisms are not associated with atrial septal defect in Chinese Han population: a brief communication. *DNA and cell biology* **31**, 1121–1124 (2012).
38. Rogkoti, V. M. *et al*. An integrated systems microscopy and transcriptomics analysis identifies gene signatures of breast cancer cell migratory and invasive behavior. *Submitted*.
39. Van Roosmalen, W., Le Dévédec, S. E., Zovko, S., De Bont, H. & Van De Water, B. Functional screening with a live cell imaging-based random cell migration assay. *Methods Mol. Biol.* **769**, 435–448 (2011).
40. Fokkelman, M. *et al*. Uncovering the signaling landscape controlling breast cancer cell migration identifies splicing factor PRPF4B as a metastasis driver. *bioRxiv* 479568 (2018).
41. Van Roosmalen, W. *et al*. Tumor cell migration screen identifies SRPK1 as breast cancer metastasis determinant. *J. Clin. Invest.* **125**, 1648–1664 (2015).
42. Struckhoff, A. P. *et al*. PDZ-RhoGEF is essential for CXCR4-driven breast tumor cell motility through spatial regulation of RhoA. *J. Cell Sci.* **126**, 4514–26 (2013).
43. Mootha, V. K. *et al*. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
44. Subramanian, A. *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
45. Liberzon, A. *et al*. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Cancer Institute and the Center for Cancer Research. S.H. was partly funded by NSF award 1564785. We would like to thank Kuoyuan Cheng and Dr. Asha Krishna for their valuable comments and inputs into this work.

Author Contributions

N.U.N., A.D., E.R. formulated the research question and study design. V.R., M.F., C.G.D.J., E.K., B.V.D.W., SELD carried out the experimental measurements for migration assays. R.M., B.G.N. carried out the experimental measurements for proliferation assays. N.U.N., A.D., E.R., I.M. designed the methods. N.U.N. and A.D. carried out the bioinformatic analysis. N.U.N., A.D., E.R., J.S.L., S.H. and S.E.L.D. analyzed the results. N.U.N., A.D., E.R. wrote the manuscript with help from all the other authors. The manuscript has been read and approved by all the authors.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-47440-w.

Competing Interests: The authors declare no competing interests.