Published in final edited form as:

Mol Biosyst. 2017 November 21; 13(12): 2615–2624. doi:10.1039/c7mb00416h.

Hedgehog-mesenchyme gene signature identifies bi-modal prognosis in Luminal and Basal breast cancer sub-types

Wandaliz Torres-García¹ and Maribella Domenech²

¹Department of Industrial Engineering, University of Puerto Rico Mayagüez, Mayagüez, PR

²Department of Chemical Engineering, University of Puerto Rico Mayagüez, Mayagüez, PR

Abstract

Hedgehog signaling (Hh) has been shown to be hyper-activated in several cancers. However, active Hh signaling can promote or inhibit tumor growth, thus identification of markers beyond main canonical Hh target genes are needed to improve patient selection and clinical outcome in response to Hh inhibitors. Cancer-associated fibroblast (CAFs) have been linked with tumor progression and beneficial response to Hh inhibitors. Thus, we hypothesized that genes associated with Hh-activated CAFs can be used for stratification of tumors that will benefit from Hh inhibitors. In this work, we evaluated a 15-gene fingerprint that combines Hh and mesenchymal genes associated to CAFs phenotype to profile breast cancer sub-types based on gene expression patterns among clustered groups. About 3,800 cancer samples were evaluated using random forest models and linear discriminant analysis to sort breast cancer by subtypes and therapeutic approach. Results showed that the Hh-mesenchyme gene fingerprint has a highly sensitive and differential expression pattern among basal and luminal A sub-groups. Basal samples with high levels of Hh target genes had better prognosis than Luminal A samples. Luminal A samples with tendency towards Hh signaling suppression had higher overall and disease-free survival rates particularly if deprived of hormone therapy. Hh transcriptional repressor GLI3 and signaling activator SMO were the top 2 genes to discriminate among samples with active Hh signaling in human breast cancer subtypes and Hh-inhibitor resistant tumors. Caveolin-1 (CAV1), a gene associated to low expression in CAFs, shows strong correlation with active Hh signaling and discrimination among survival curves in Luminal A patients with active or inactive Hh signaling. Our data suggests that CAV1 is an important gene for monitoring Hh inhibition in tumors and supports further stratification by hormone therapy status prior to use of Hh inhibitors.

Introduction

Hedgehog (Hh) signaling is a developmental pathway that becomes activated during tissue patterning in embryogenesis(1, 2), and in expansion of rapid turnover cells like stem cells localized to normal adult tissues(3, 4). Canonical Hh signaling is driven by binding of Hh ligand Desert, Indian, or Sonic hedgehog (SHH) ligands to the transmembrane protein receptor Patched1 (PTCH1). PTCH1-ligand interaction relieves inhibition on Smoothened (SMO) promoting translocation of the glioma-associated oncogene homolog (GLI)

transcription factors Gli1 and Gli2 to the nucleus. Gli1 and Gli2 transcription factors then promote upregulation of genes associated with cell cycle (e.g. Cyclin D, MYC), vascularization (e.g. VEGF) and a Hh feedback control loop including GLI1, GLI2, GLI3 (transcriptional repressor) and PTCH1.

Overexpression of Hh ligands and target genes have been shown to be enriched in several cancers such as basal cell carcinoma(5), pancreas(6), prostate(7, 8) and breast(9–11). Hh inhibitors have been successful in treating basal cell carcinoma which is enriched for mutations in PTCH1 that lead to uncontrolled activation of Smoothened(5). Unfortunately, clinical outcomes in patients with non-mutation Hh-driven tumors have been unsuccessful. Hh inhibitors have failed in most clinical trials in which overexpression of Hh target genes has been observed(12, 13). For example, combinatorial use of gemcitabine with the Smo inhibitor vismodegib (Genentech) didn't improve overall survival rates in patients with active Hh signaling in advanced pancreatic tumors when compared to standard gemcitabine treatment alone(13). In several clinical trials, even though the overall group of patients didn't improve with combinatorial treatment, some patients did benefit from Smo inhibitors (14) suggesting that other factors besides overexpression of Hh targets are relevant for clinical response.

One of the main factors that can influence clinical response is the bimodal growth and metastasis observed in tumors in response to Hh inhibitors. Independent studies in colon and pancreatic cancers show that inhibition of Hedgehog signaling can either promote or inhibit tumor growth. Studies done by Beachy et al. show that inhibition of Hh signaling in the adjacent stroma promotes tumor growth and metastasis in 3 genetically different mouse models of pancreatic cancer (15). In contrast, studies done by Hwang *et al*, show that inhibition of Hh signaling in adjacent stroma inhibit tumor growth in both colon and pancreatic cancer models (16, 17). However, the clinical benefit of Hh inhibitors was limited to targeting Hh-activated cancer-associated fibroblasts and embryonic fibroblasts but not normal adult adjacent stroma. Thus, the source of Hh-responsive mesenchymal cells can modulate tumor suppression and promoting mechanisms during Hh signaling, and thereby new strategies that incorporate mesenchymal markers can lead to improve methods for identification of patients that will benefit from Hh inhibitors.

Currently, there are no tools to discriminate between tumor promoting and inhibiting Hh signaling. Candidates for inhibitors are typically identify by activation of main Hh target genes (SMO, GLI1, PTCH1, GLI2 and GLI3) which will be similar regardless of the adjacent mesenchymal sub-type. Thus, potential clinical benefit of Hh inhibitors may be underscored in clinical trials due to lack of screening tools to sample candidates and monitor overall response to Hh inhibitors. In this study, we selected 21 genes associated to Hh signaling, mesenchymal cell sub-types and tumor-promoting response in Hh-activated mesenchymal cells to sort samples and identify discrete gene patterns among breast cancer sub-types that could lead to differences in clinical outcome. We identified a Hh-mesenchyme 15-gene fingerprint that can recognized 2 sub-groups in basal and luminal A breast cancer sub-types with opposite survival prognosis curves in response to Hh signaling. Evaluation of this gene fingerprint in basal cell carcinoma and medulloblastoma mouse models treated with SMO inhibitors reinforces its potential to also monitor response to Hh inhibition.

Inclusion of mesenchymal markers into monitoring of Hh-activity in tumors will enhance our understanding of the bi-modal growth regulation mechanism and improve the identification of patient candidates that will benefit from Hh inhibitors.

Results and Discussion

We present the results from the sorting of breast cancer sub-types based on the similarities of the expression pattern of selected 15 genes associated to Hh signaling and mesenchymal phenotype. Both random forests (RFs) and linear discriminant analysis (LDA) models were implemented measuring error rates and area under the curve (AUC) in a cross validated approach to assess its predictive power. Initially, we measured the predicting performance of this 15-gene signature towards sorting breast cancer sub-types (i.e. Basal, Luminal A, Luminal B, HER2, Normal) among 3,800 gene expression profile samples. In Table 1, the overall error rates for breast cancer subtyping ranged between 30–50% with similar results regardless of the model (Supp.B). Thus, error rates from RFs are used to highlight the selectivity of gene expression patterns found across sorted subgroups. Although, the error rates were not as low as those reported for other models(18), it is important to mention that key breast-associated oncogenes (i.e. ESR1, BRCA1/BRCA2) were not included in the study since the objective here is to identify subgroups with differential prognosis based on the gene pattern prevalence of the selected Hh-mesenchymal genes, and thereby identify sub-groups that may benefit of Hh inhibitors.

Sorting of samples per gene expression highlighted a discrete gene expression pattern for basal and luminal A sub-types. This gene signature has non-overlapping expression patterns among Basal and Luminal A samples compared to other subtypes with error rates as low as 12% (Table 1). The majority of basal samples (70–80%) that display high expression levels of SMO and low levels of GLI3 whereas the majority of Luminal A samples display the opposite pattern across all data sets indicating active and repressed state for Hh signaling respectively (See Figure 1). HER2, Luminal B and Normal samples were consistently difficult to sort using this 15-gene signature. However, in most cases HER2 and normal subtypes were the minority of the total samples which negatively affects error rates due to low statistical power. A permutation test was performed to assess the predictive importance of this signature in comparison with any random 15-gene-set. The Hh-mesenchyme signature ranked among the 85th percentile using the data from The Cancer Genome Atlas (TCGA) (Supp.C). To further highlight the predictive performance of this signature for sub-types, AUC values were computed across all models and obtained high values fluctuating between 65% and 80% (Table 1). High AUC values indicate good sensitivity and specificity metrics.

To identify which genes were more relevant for sorting of basal and luminal A sub-groups, we looked at the higher scores from the mean decrease gini (MDG) obtained in RFs. The top 5 genes from seven databases identify GLI3 as the main gene resulting in all seven (7/7) followed by CAV1 (6/7) and SMO (5/7) (Figure 2A). To determine whether Hh signaling was active in both basal and luminal sub-groups, we studied their marginal effect through the computation of partial dependency plots (PDPs). These plots offer a graphical representation of the log-odds for a subtype class depending on an input value of the

predictor gene which is useful to interpret the gene expression impact on the subtype. Figure 2B shows the PDPs for the three most important genes in our seven datasets across Basal versus Luminal A. We observe that samples with tendency towards suppression of Hh signaling (high GLI3 and low SMO levels) have higher likelihood to be classified as Luminal A subtype (Figure 2B). As expected, opposite trend is observed for classified Basal samples which correlates with activation of Hh signaling. Clearly, >80% of basal breast tumors show high levels of Hh target genes which further supports previous small sample number studies that correlate overexpression of Hh target genes in triple negative breast cancer (TNBC) (10). A specific trend was not clear for CAV1 as its relationship between Basal and Luminal A sub-types is not dualistic as in the case of GLI3 and SMO. Loss of CAV1 has been linked to cancer-associated fibroblasts and could indicate the abundance of this cell phenotype in tumors with active Hh signaling.

Survival Analysis

The construction of univariate and multivariate Cox proportional hazard regression models showed clinical potential for the 15-gene signature and its gene expression profiles. Univariate hazard models were built using all samples from the METABRIC and they revealed IGFBP6, HER2 status, PR status, CAV1, and whether the classifier was able to predict subtype correctly (CorrectlyClassified) among the most relevant individual predictors of survival risk with log-rank test p-values less than 0.0001 (See Supp.D. Table D1). IGFBP6, CAV1 and PR status covariates were found statistically significant to describe overall survival in Luminal A samples as well as shown in Table 2. These models showcased the importance of receptors (i.e. PR_STATUS) and expression levels of IGFBP6 and CAV1 towards patient survival status. The increase in expression levels for both of these genes were highly associated with better prognosis (HR: 0.8 and 0.88). Moreover, the expression values of CAV1 and IGFBP6 are somewhat correlated ($r=0.66$), hence the multivariate models found no benefit in including both of them. The hazard models for Basal samples (Supp.D. Table D2) did not produced significant expression patterns to discriminate survival. However, for Luminal A samples, further subgrouping of these samples based on expression levels of genes such as IGFBP6 could improve clinical outcome and better therapeutic options assessments (see Table 2 and Table D4).

Furthermore, we assessed the overall survival prognosis of patients based on receptors' status (i.e. ER, PR, HER2) through the construction of Kaplan Meier curves and log-rank tests. We found that there are statistical differences in overall survival for the METABRIC samples grouped based on the inactivation or activation of these three receptors as shown in Supp.D. For example, overall survival between ER+ versus ER- patients showed a p-value of 0.033 for the log-rank test using the G-rho family of survival tests. These results were similarly observed for HER2 and PR groupings with p-values less than 0.0001. This result is not surprising as the status of these receptors are commonly used to categorize clinical interventions. However, the gene expression signature presented in this work can further subdivide these patients into different subgroups with significant survival differences. For example, ER+ patients had better survival outcome (hazard ratio (HR) = 0.813) when the model could classify them correctly than those patients that did not match the gene expression patterns found in the subtype characterization model. The differences between

these subgroups were found statistically significant through the log-rank test for all possible groupings of receptors status and model prediction but one (See Supp.D. log-rank test p-values ER+: 0.00225, ER– 0.0215, HER2+ 0.123, HER2– 0.00556, PR+ 0.0254, PR – 0.00278). This suggests that differences in the gene expression values of this signature could stratify patients with better survival probability even when they all share the same receptor status.

Bi-modal prognosis of Hh signaling in Basal and Luminal A samples

Overall survival and disease-free survival curves were compared among basal and luminal sub-groups with high and low Hh signaling. Correctly and incorrectly classified subgroups represent the majority and minority of the sample population for each breast cancer sub-type respectively. Our results highlight an association of the Hh pathway with the Luminal A subtype in patients which has never been reported before suggesting that Hh activity has a role in the progression of these tumors. In fact, we observed statistical differences between survival curves for all samples sub-types (p-val<0.0001), and for those Luminal A (p-val=0.0343) correctly classified vs those incorrectly classified by our RFs model using the METABRIC dataset (Figure 2C). Correctly classified Luminal A and Basal samples had better prognosis rates than incorrectly classified. Considering that correctly classified Basal and Luminal A subgroups show active and repressed Hh signaling signatures, our results suggest that Hh signaling is tumor inhibiting for basal subgroup and tumor promoting for the Luminal A sub-group.

Overall survival and DFS rates were further stratified by therapeutic approach (Supp.D: Figures D4–D6), Luminal A samples with a Hh-signaling suppressive profile (Luminal A correctly classified sub-group) had significantly lower survival rates if deprived of hormone therapy suggesting a regulatory interplay between hormone and Hh signaling (Figure 2A/2C). Basal samples with Hh-suppressive gene pattern (incorrectly classified sub-group) had significantly lower overall survival and DFS rates if treated with hormone therapy (Figure 2B/2D). Thus, hormone therapy status is another important factor for potential discrimination of tumor promoting and inhibiting mechanisms of Hh signaling and selection of patients for pharmacological Hh-inhibitors.

CAV1/IGFBP6 a potential gene for identification of tumor-promoting Hh signaling

As CAV1 expression pattern was a top classifier gene and its down regulation is associated with the phenotype of Hh-responsive stroma, its potential to distinguish among survival prognosis was evaluated. CAV1 was an important gene to discriminate across breast cancer sub-groups including Basal and Luminal A (See Figure 1). IGFBP6 was a relevant gene to discriminate among survival curves within the Luminal A sub-group (Table 2). Both, CAV1 and IGFBP6 were selected for their association to the mesenchymal phenotype of Hh-responsive stroma. Downregulation of CAV1 has been linked with cancer-associated fibroblasts (19) and its overexpression in tumor adjacent stroma is a good predictor of improved outcome in breast cancer(20). IGFBP-6 is a target of Hedgehog signaling in the tumor adjacent stroma (21, 22). Interestingly, the gene expression pattern for IGFBP-6 and CAV1 were very similar across sub-groups. Downregulation of both CAV1 and IGFBP6 was observed in luminal A subgroup with lowest survival. A similar expression pattern between

these two genes was reported to distinguish among thyroid cancer sub-groups(23). Thus, CAV1 and IGFBP6 seem to have an important role in endocrine tumors, and further evaluation at the molecular level may provide new insights into the regulatory mechanisms and potential crosstalk at the cellular level.

Detection of pharmacological sensitivity

To determine which genes could discriminate among normal or resistant tumors to Hh inhibitors, we evaluated the performance of the signature among pharmacological responsive and resistant tumors. As clinical datasets from patients were not available, we used data from mice models of medulloblastoma and basal cell carcinoma treated with Hh inhibitors. GLI1, GLI2 and SMO were found as most relevant genes to discriminate among these groups using MDG score from the RFs. All three genes tend to be less expressed in most of the drug sensitive instances and highly expressed in the drug resistant ones as shown in the heatmaps from Supplemental Data D. When comparing, the top genes used for sample classification, SMO is highlighted as top gene for identification of pharmacologically responsive tumors and bi-modal prognosis sub-groups among Luminal A and Basal samples. Low levels of SMO resulted in better clinical outcome in mice models, and for has higher log-odds for Luminal A which in general tends to be a less aggressive subtype than basal tumors (Figure 4 and Figure 2C). High levels of SMO are abundant in TNBC which strongly indicates activation of Hh signaling but correlated with a better prognosis for Basal samples suggesting a tumor inhibitory rather than tumor promoting effect. Overexpression of SMO hasn't been shown to be a result of canonical Hh signaling and no mutations were found to be associated with this gene in the tumor samples examined. Thus, the cause of SMO upregulation in breast and other tumors is still unknown but critical for the improvement of Hh-targeted therapies.

Conclusion

Overall, this work characterized the expression pattern of a 15-gene Hh-mesenchyme signature among 3,800 samples across breast cancer subtypes and drug sensitivity models using data mining techniques. In this work, we show that this gene signature identifies bi-modal tumor promoting and inhibiting behavior among Basal and Luminal A subgroups. Further stratification of sorted sub-groups highlights hormone therapy as a key parameter in the clinical prognosis of samples with Hh suppressive signals. Even though, GLI3, the most important gene for subtype classification, was not found as one of the most relevant genes in the drug sensitivity study its counterparts GLI1 and GLI2 were. These genes have intrinsic dependencies and their interactions in breast cancer should be further explored. Our results support the use of this signature and therapeutic regimen for identification of candidate patients for Hh inhibitors.

Methods

General analysis framework

This work presents the results of a computational data-driven approach aiming to discover potential expression patterns in a gene expression signature consisting of 15 genes from

Hedgehog signaling and mesenchymal markers. The multi-step analysis processed a large amount of gene expression samples, constructed linear and non-linear models to distinguish breast cancer subtypes and drug resistance samples, and evaluated the survival impact of revealed patterns (See Figure 5).

Selection of Hedgehog-mesenchymal signature

To identify the Hh-mesenchyme 15-gene signature, we first evaluated a total of 21 genes associated to active Hh signaling in a 47 samples data set (GSE3744) composed by TNBC, Non-TNBC and normal groups. The 21 genes can be grouped as follows: 6 genes associated to canonical Hh signaling (GLI1, GLI2, GLI3, SMO, PTCH1, HHIP), 6 mesenchymal markers observed in cancer-associated fibroblasts in breast (CAV1, FAP, VIM, ACTA1, CDH1 and CDH2) (24–27) and 9 genes identified in highly Shh-responsive embryonic myofibroblasts from a mouse xenograft model of paracrine-driven Hh signaling in prostate tumors (CXCL14, FBN2, ANGPT4, TIMP3, IGFBP6, ADAM12, FGF5, HES1 and HSD11B1) (21). Genes expressed in the adjacent stroma of paracrine-driven Hh tumors in breast cancer hasn't been described. The rationale in selecting markers of paracrine-driven Hh tumors in the prostate for sorting of breast samples is that both gland share similarities with regard to physiology and pathology (28) such as tissue markers, hormonal regulation and cancer metastatic sites (28, 29). We used linear discriminant regression analysis (LDA) sort tumor samples based on mRNA expression levels of the panel of 21 genes in gene microarray profiles. LDA was performed to exclude genes that are not differentially expressed across normal, non-TNBC and TNBC (ER⁽⁻⁾PR⁽⁻⁾Her2⁽⁻⁾) groups(30, 31). This type of analysis builds a prediction model by progressively adding the variables (genes) with the most significant (individual p-value = 0.05) at each step. Based on this analysis, a 15 of 21 genes were found to discriminate among breast cancer patients and normal tissues with >90% accuracy. To further validate the significance of the 15 gene signature, we performed a gene permutation test to compare our results with other sets of 15 genes. Our results placed the Hh-mesenchyme gene signatures in the 85th percentile. (Supp. A & B).

Datasets

The gene expression datasets used in this work includes data generated by TCGA in RNA sequencing format and METABRIC(32) microarrays as well as several other publicly available data from the NCBI GEO data repository(33–40). METABRIC collected close to 2000 primary fresh-frozen breast cancer cases from UK and Canada for gene expression studies (32). METABRIC dataset was used for all the analyses involving survival significance since it was the largest repositories publicly available. Intrinsic information on breast cancer subtypes based on PAM50 classifier(41) was available for all the datasets. We evaluated ~3800 breast cancer samples using gene expression platforms Agilent and Affymetrix. Datasets were preprocessed using median-based or logarithmic Lowess normalization for two-color arrays. In the case of multiple probes for the same gene symbol we used the maximum normalized score as representation of its gene expression. For the drug resistance, NCI GEO GSE58375 and GSE77042 expression datasets for basal cell carcinoma (BCC) and medulloblastoma were used. The BCC data consists of 21 samples (9-resistant/12-sensitive) with RNA-sequencing FPKM values((42)). For medulloblastoma, this

set contained 75 samples (50-resistant/25-sensitive to drug therapy) from mice with gene expression intensities calculated using RMA and MAS-5(43).

Classification Models

Linear discriminant analysis (LDA) and Random Forest models (RFs) were implemented to predict breast cancer subtypes from gene expression profiles. Both methods, LDA and RFs, aim to separate several classes using linear combinations or subsets of features respectively. The features are represented by the expression of genes from the Hedgehog-mesenchymal markers and the classes are the breast cancer subtypes.

LDA attempts to find linear combinations among features (i.e. 15-gene signature) to describe the response variable (i.e. breast cancer subtype). Given a set of gene expression profiles, $x_1, \dots, x_n$ and their subtype classification, $y_1, \dots, y_n$ where n represents the number of samples, LDA aims to maximize the variability between classes over the within-class variability through a set of linear combinations. Hence, it computes the within-class variability (S_w) and the between-classes variability (S_b) as stated in Equation 1 where c represents the total number of classes. To maximize the ratio of S_b to the S_w , a set of linear combinations gathered through w weight coefficients are found solving a generalized eigenvalue problem.

$$S_w = \sum_{i=1}^n (x_i - \mu_{y_i})(x_i - \mu_{y_i})' \text{ and } S_b = \sum_{k=1}^c n_k (\mu_k - \mu)(\mu_k - \mu)' \quad \text{Equation 1}$$

RFs are ensemble classifiers based on decision trees that uses the benefits of bootstrapping through its modeling. They are also simple to train, can handle a large amount mixed-type data and provide ways to evaluate variable importance scores. Ensemble methods such as random forest are known to have strong accuracy performance and reliable identification of important variables and interactions between them. A forest is composed of a set of individual decision trees where each tree is constructed from a random sample of instances and at each tree node a random sample of features is evaluated. This bootstrapping sampling improves predictive performance and computational efforts as well as reducing overfitting. Two key parameters to tune in RFs models are the number of individual trees (ntree) and the number of randomly sampled variables at each tree split node (mtry). The models presented here shown mtry values equal to 3 and ntree equal to 10000 (See Supp.B. for more details).

LDA and RFs models were implemented in R statistical program (44) and packages MASS and randomForest(45) respectively. The metrics used to evaluate the performance of the classifiers were error rates and AUC which are commonly used in the field to assess model behavior. For RF models, other metrics such as Mean Decrease Gini measure (MDG) were used to evaluate variable importance. The larger the MDG score the more important is a predictor, meaning that the variable makes purer splits in the tree on average across the forest. A low Gini (i.e. more homogeneous) represents that a predictor variable has a key role in separating the data into the defined groups (i.e. subtypes). Each time a predictor is used as a split node the difference in Gini metric (MDG) between the split node and previous node is computed. If this difference is large, it presents an improvement by that

predictor at the split node. Furthermore, we used partial dependency plots (PDPs) to extract the marginal effect of an important variable in RFs. PDPs are graphical representations of average predictor's behavior for a given class. This marginal effect of a particular predictor variable can be computed as an averaged log-odds metrics across all trees in the forest for a specific class. In each partial dependency plot, the x axis has the range of possible gene expression values for that gene predictor and the y axis provides the log-odds values of belonging to the specific class (either basal/luminal or resistant/sensitive). The scale of the x axis across all datasets varied because the gene expression values are different for each independent study (See Methods Section for more details). The log-odds are plotted on the y

axis as follows $y = \log p_k(x) - \frac{1}{K} \sum_{j=1}^K \log p_j(x)$ where K is the number of classes (i.e. subtypes or resistant/sensitive) and the possible expression values are explored and plotted in the x axis for the genes studied here.

Survival Models

Survival curves and Cox-Hazard proportion models were implement using R survival(46) package. Univariate and multivariate analyses were performed and analyzed. Specifically, Kaplan-Meier curves were constructed for different groups (receptor status, correctly classified by the signature versus not correctly classified; hormone therapy, radiotherapy, chemotherapy). These curves were compared using Fleming-Harrington G-rho family of test (survdif).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the NIH-NCI K01-CA188167 and small research grant program from the Puerto Rico Science and Technology Thrust.

References

1. Apelqvist A, Ahlgren U, Edlund H. Sonic hedgehog directs specialised mesoderm differentiation in the intestine and pancreas. *Curr Biol*. 1997 Oct 01; 7(10):801–4. [PubMed: 9368764]
2. Chiang C, Litingtung Y, Lee E, Young KE, Corden JL, Westphal H, et al. Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. *Nature*. 1996 Oct 03; 383(6599): 407–13. [PubMed: 8837770]
3. Solanas G, Benitah SA. Regenerating the skin: a task for the heterogeneous stem cell pool and surrounding niche. *Nature Reviews Molecular Cell Biology*. 2013; 14(11):737–48. [PubMed: 24064540]
4. Katoh Y, Katoh M. Hedgehog signaling pathway and gastrointestinal stem cell signaling network (review). *Int J Mol Med*. 2006 Dec; 18(6):1019–23. [PubMed: 17089004]
5. Fecher LA, Sharfman WH. Advanced basal cell carcinoma, the hedgehog pathway, and treatment options - role of smoothened inhibitors. *Biologics*. 2015; 9:129–40. [PubMed: 26604681]
6. Thayer SP, Pasca di Magliano M, Heiser PW, Nielsen CM, Roberts DJ, Lauwers GY, et al. Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis. *Nature*. 2003; 425(6960):851–6. [PubMed: 14520413]
7. Gonnissen A, Isebaert S, Haustermans K. Hedgehog signaling in prostate cancer and its therapeutic implication. *Int J Mol Sci*. 2013 Jul 04; 14(7):13979–4007. [PubMed: 23880852]

8. Bushman, W. Hedgehog Signaling in Development and Cancer. In: Chung, LWK, Isaacs, WB., Simons, JW., editors. Prostate Cancer: Biology, Genetics, and the New Therapeutics. Totowa, NJ: Humana Press; 2007. p. 107-18.
9. Habib JG, O'Shaughnessy JA. The hedgehog pathway in triple-negative breast cancer. *Cancer Med*. 2016 Oct; 5(10):2989–3006. [PubMed: 27539549]
10. Noman AS, Uddin M, Rahman MZ, Nayeem MJ, Alam SS, Khatun Z, et al. Overexpression of sonic hedgehog in the triple negative breast cancer: clinicopathological characteristics of high burden breast cancer patients from Bangladesh. *Sci Rep*. 2016 Jan 05;6:18830. [PubMed: 26727947]
11. Hui M, Cazet A, Nair R, Watkins DN, O'Toole SA, Swarbrick A. The Hedgehog signalling pathway in breast development, carcinogenesis and cancer therapy. *Breast Cancer Research*. [journal article]. 2013; 15(2):203.
12. Berlin J, Bendell JC, Hart LL, Firdaus I, Gore I, Hermann RC, et al. A randomized phase II trial of vismodegib versus placebo with FOLFOX or FOLFIRI and bevacizumab in patients with previously untreated metastatic colorectal cancer. *Clinical Cancer Research*. 2013; 19(1):258–67. [PubMed: 23082002]
13. Kim EJ, Sahai V, Abel EV, Griffith KA, Greenson JK, Takebe N, et al. Pilot clinical trial of hedgehog pathway inhibitor GDC-0449 (vismodegib) in combination with gemcitabine in patients with metastatic pancreatic adenocarcinoma. *Clin Cancer Res*. 2014 Dec 01; 20(23):5937–45. [PubMed: 25278454]
14. Sandhiya S, Melvin G, Kumar SS, Dkhar SA. The dawn of hedgehog inhibitors: Vismodegib. *Journal of Pharmacology & Pharmacotherapeutics*. 2013 Jan-Mar;4(1):4–7. [PubMed: 23662017]
15. Lee JJ, Perera RM, Wang H, Wu DC, Liu XS, Han S, et al. Stromal response to Hedgehog signaling restrains pancreatic cancer progression. *Proc Natl Acad Sci U S A*. 2014 Jul 29; 111(30):E3091–100. [PubMed: 25024225]
16. Hwang RF, Moore T, Arumugam T, Ramachandran V, Amos KD, Rivera A, et al. Cancer-associated stromal fibroblasts promote pancreatic tumor progression. *Cancer Res*. 2008 Feb 01; 68(3):918–26. [PubMed: 18245495]
17. Hwang RF, Moore TT, Hattersley MM, Scarpitti M, Yang B, Devereaux E, et al. Inhibition of the hedgehog pathway targets the tumor-associated stroma in pancreatic cancer. *Mol Cancer Res*. 2012 Sep; 10(9):1147–57. [PubMed: 22859707]
18. Narvaez, IT-G., Wandaliz. Data-driven Approach to Extract Molecular Patterns in Breast Cancer using Transcriptomic and Clinical Data. *IIE Annual Conference Proceedings ISERC*; 2015; 2015.
19. Mercier I, Casimiro MC, Wang C, Rosenberg AL, Quong J, Minkeu A, et al. Human breast cancer-associated fibroblasts (CAFs) show caveolin-1 downregulation and RB tumor suppressor functional inactivation: Implications for the response to hormonal therapy. *Cancer Biol Ther*. 2008 Aug; 7(8):1212–25. [PubMed: 18458534]
20. Shan-Wei W, Kan-Lun X, Shu-Qin R, Li-Li Z, Li-Rong C. Overexpression of Caveolin-1 in Cancer-Associated Fibroblasts Predicts Good Outcome in Breast Cancer. *Breast Care*. 2012; 7(6): 477–83. [PubMed: 24715830]
21. Shaw A, Gipp J, Bushman W. The Sonic Hedgehog Pathway Stimulates Prostate Tumor Growth by Paracrine Signaling and Recaptures Embryonic Gene Expression in Tumor Myofibroblasts. *Oncogene*. 2009; 28(50):4480–90. [PubMed: 19784071]
22. Lipinski RJ, Cook CH, Barnett DH, Gipp JJ, Peterson RE, Bushman W. Sonic hedgehog signaling regulates the expression of insulin-like growth factor binding protein-6 during fetal prostate development. *Developmental Dynamics*. 2005; 233(3):829–36. [PubMed: 15906375]
23. Aldred MA, Huang Y, Liyanarachchi S, Pellegata NS, Gimm O, Jhiang S, et al. Papillary and Follicular Thyroid Carcinomas Show Distinctly Different Microarray Expression Profiles and Can Be Distinguished by a Minimum of Five Genes. *Journal of Clinical Oncology*. 2004; 22(17):3531–9. [PubMed: 15337802]
24. Shiga K, Hara M, Nagasaki T, Sato T, Takahashi H, Takeyama H. Cancer-Associated Fibroblasts: Their Characteristics and Their Roles in Tumor Growth. *Cancers*. 2015; 7(4):2443–58. [PubMed: 26690480]

25. Shan-Wei W, Kan-Lun X, Shu-Qin R, Li-Li Z, Li-Rong C. Overexpression of caveolin-1 in cancer-associated fibroblasts predicts good outcome in breast cancer. *Breast Care (Basel)*. 2012 Dec; 7(6): 477–83. [PubMed: 24715830]
26. Sotgia F, Del Galdo F, Casimiro MC, Bonuccelli G, Mercier I, Whitaker-Menezes D, et al. Caveolin-1(–/–) Null Mammary Stromal Fibroblasts Share Characteristics with Human Breast Cancer-Associated Fibroblasts. *The American Journal of Pathology*. 2009; 174(3):746–61. [PubMed: 19234134]
27. Buchsbaum RJ, Oh SY. Breast Cancer-Associated Fibroblasts: Where We Are and Where We Need to Go. *Cancers*. 2016; 8(2):19.
28. Lopez-Otin C, Diamandis EP. Breast and prostate cancer: an analysis of common epidemiological, genetic, and biochemical features. *Endocr Rev*. 1998 Aug; 19(4):365–96. [PubMed: 9715372]
29. Coffey DS. Similarities of prostate and breast cancer: Evolution, diet, and estrogens. *Urology*. 2001 Apr; 57(4 Suppl 1):31–8. [PubMed: 11295592]
30. Brabender J, Marjoram P, Salonga D, Metzger R, Schneider PM, Park JM, et al. A multigene expression panel for the molecular diagnosis of Barrett's esophagus and Barrett's adenocarcinoma of the esophagus. *Oncogene*. 2004 Jun 10; 23(27):4780–8. [PubMed: 15107828]
31. Nebozhyn M, Loboda A, Kari L, Rook AH, Vonderheid EC, Lessin S, et al. Quantitative PCR on 5 genes reliably identifies CTCL patients with 5% to 99% circulating tumor cells with 90% accuracy. *Blood*. 2006 Apr 15; 107(8):3189–96. [PubMed: 16403914]
32. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012; 486(7403):346–52. [PubMed: 22522925]
33. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*. 2002 Jan 1; 30(1):207–10. [PubMed: 11752295]
34. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*. 2013 Jan 1; 41(D1):D991–D5. [PubMed: 23193258]
35. Esserman LJ, Berry DA, Cheang MCU, Yau C, Perou CM, Carey L, et al. Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Research and Treatment*. 2012; 132(3):1049–62. 12/25 11/18/received 11/21/accepted. [PubMed: 22198468]
36. Kao K-J, Chang K-M, Hsu H-C, Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC cancer*. 2011; 11(1):143. [PubMed: 21501481]
37. Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, et al. DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO molecular medicine*. 2011; 3(12):726–41. [PubMed: 21910250]
38. Sabatier R, Finetti P, Adelaide J, Guille A, Borg J-P, Chaffanet M, et al. Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. *PloS one*. 2011; 6(11):e27656. [PubMed: 22110708]
39. Ciriello G, Gatz Michael L, Beck Andrew H, Wilkerson Matthew D, Rhie Suhan K, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*. 163(2):506–19.
40. Alimonti A, Carracedo A, Clohessy JG, Trotman LC, Nardella C, Egia A, et al. Subtle variations in Pten dose determine cancer susceptibility. *Nat Genet*. 2010 May; 42(5):454–8. //print. [PubMed: 20400965]
41. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*. 2009; 27(8):1160–7. [PubMed: 19204204]
42. Atwood SX, Sarin KY, Whitson RJ, Li JR, Kim G, Rezaee M, et al. Smoothed variants explain the majority of drug resistance in basal cell carcinoma. *Cancer cell*. 2015; 27(3):342–53. [PubMed: 25759020]

43. Filocamo G, Brunetti M, Colaceci F, Sasso R, Tanori M, Pasquali E, et al. MK-4101-a potent inhibitor of the hedgehog pathway-is highly active against medulloblastoma and basal cell carcinoma. *Molecular cancer therapeutics*. 2016 molcanther. 0371.2015.
44. RTeam. Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2005.
45. Liaw A, Wiener M. Classification and regression by randomForest. *R news*. 2002; 2(3):18–22.
46. Therneau, T. R package version 2.37-4. R-project org/package= survival. 2014. A package for survival analysis in S.

Figure 1. Hh target genes (GLI3 and SMO) and mesenchymal marker (CAV1) expression levels for Basal and Luminal samples with their respective survival outcome. (A) Z-scores expression heatmap (B) Z-scores expression boxplots.

Figure 2.

RFs model performance for subtype classification and its top relevant genes. **(A)** The top 5 variable important scores using MDG score for all seven datasets in study. MDG is a metric that measures how well a variable classifies the data, the greater the decrease in MDG the gene contributes more to better classification. **(B)** PDPs for top genes GLI3, CAV1, and SMO presenting the marginal effect of a gene per subtype for the models performed through the calculation of log-odds. The log-odds are plotted on the y-axis and the possible expression values are plotted in the x-axis. SMO gene was not found in (GSE22226). **(C)** Survival curves for overall survival (months) for METABRIC samples that were correctly (--) vs incorrectly (-) classified for all (p=0.000138), Luminal A (p=0.0343) and Basal (p=0.355) samples.

Figure 3.

Overall and DFS curves by prediction/hormone therapy. (A–B) shows overall survival for Luminal A and Basal samples from the METABRIC datasets respectively; inspected for differences on whether the 15-gene signature correctly classify them (TRUE/FALSE) and if hormone therapy was received(Yes/NO). (C–D) shows DFS curves for Luminal A and basal respectively. Time is in months.

Figure 4. PDPs for top drug resistance genes GLI1, GLI2, SMO and HHIP from RFs models. PDPs quantify the marginal effect of a gene in a particular class for the models performed through the calculation of log-odds. The log-odds are plotted on the y axis and expression values plotted in the x axis.

Figure 5.
General analysis framework: Overview of processed data, methods implementation, performance evaluation and signature interpretability.

Table 1
Hh-mesenchymal gene signature performance across subtypes using random forest algorithm

Classification Model	Dataset	Number of Samples*	AUCs	Error Rates (%)										Resistance	Sensitive
				Overall	Basal	HER2	LumA		LumB	Normal					
							IDC	ILC							
Subtype Signature Characterization	TCGA_Cell	587	0.80	32.03	14.95	94.11	22.39	18.87	48.36	--	--				
	METABRIC	1974	0.68	42.65	31.31	75.00	24.23		49.18	72.86					
	GSE 20685	327	0.76	41.59	24.32	77.33	34.33		28.40	--					
	GSE 20711	90	0.66	50.00	31.82	47.52	43.47		77.27	--					
	GSE 21653	266	0.65	42.11	17.33	95.83	25.84		69.39	65.51					
	GSE 22226	129	0.65	44.19	16.28	90.48	34.38		48.00	100.00					
	GSE 31448	294	0.72	37.07	12.24	80.77	25.56		67.35	64.52					
Clinical Impact: Resistant vs Sensitive	GSE 58375	21	0.79	19.05	---							0.33	0.08		
	GSE 77042	75	0.77	17.33								0.06	0.40		

* Resulting samples after preprocessing steps including the removal of samples with missing values or subtypes.
-- Normal samples were not collected. NA not applicable. IDC: Invasive Ductal Carcinoma, ILC: Invasive Lobular Carcinoma.

Table 2

Univariate Cox proportional hazard regression for overall survival for Luminal A samples using METABRIC dataset.

Covariates	beta	HR	(95% CI of HR)	wald.test	p.value	
IGFBP6	-0.39	0.67	(0.58–0.78)	27	2.60E-07	***
PR_STATUS	-0.42	0.65	(0.52–0.82)	14	0.00021	***
CAV1	-0.21	0.81	(0.72–0.92)	11	0.00091	***
FAP	-0.16	0.85	(0.76–0.95)	8.5	0.0036	**
CorrectlyClassified	-0.24	0.79	(0.63–0.98)	4.5	0.035	*
GLI2	-0.42	0.66	(0.43–1)	3.8	0.05	
VIM	-0.11	0.9	(0.79–1)	2.9	0.087	
GLI1	-0.51	0.6	(0.34–1.1)	2.9	0.088	
FBN2	-0.12	0.89	(0.77–1)	2.7	0.1	
HER2_STATUS	0.44	1.6	(0.89–2.7)	2.4	0.12	
ANGPT4	-0.4	0.67	(0.33–1.4)	1.3	0.26	
CDH2	-0.069	0.93	(0.81–1.1)	0.93	0.33	
TIMP3	-0.038	0.96	(0.86–1.1)	0.49	0.48	
SMO	-0.11	0.9	(0.66–1.2)	0.48	0.49	
ER_STATUS	-0.43	0.65	(0.16–2.6)	0.36	0.55	
FGF5	0.15	1.2	(0.65–2.1)	0.26	0.61	
HHIP	-0.14	0.87	(0.39–1.9)	0.12	0.73	
CDH1	0.011	1	(0.93–1.1)	0.07	0.79	
GLI3	-0.017	0.98	(0.85–1.1)	0.05	0.82	