



***ABI1*-based expression signature predicts breast cancer metastasis and survival**

Angelina Regua^{1,2,*}, Csaba Papp^{1,2}, Andre Grageda^{1,2}, Baylee A. Porter^{1,2}, Tiffany Caza³, Isabelle Bichindaritz⁴, Mira Krendel⁵, Abirami Sivapiragasam⁶, Gennady Bratslavsky^{1,2}, Vladimir A. Kuznetsov^{1,2}  and Leszek Kotula^{1,2} 

1 Department of Urology, SUNY Upstate Medical University, Syracuse, NY, USA

2 Department of Biochemistry and Molecular Biology, SUNY Upstate Medical University, Syracuse, NY, USA

3 Department of Pathology, SUNY Upstate Medical University, Syracuse, NY, USA

4 Department of Computer Science, SUNY Oswego, NY, USA

5 Department of Cell and Developmental Biology, SUNY Upstate Medical University, Syracuse, NY, USA

6 Department of Medicine, SUNY Upstate Medical University, Syracuse, NY, USA

Keywords

Abi1; breast cancer; metastasis; preclinical mouse; survival signature; WAVE

Correspondence

L. Kotula and V. A. Kuznetsov, Department of Urology, SUNY Upstate Medical University, 750 East Adams Street, Syracuse, NY 13210, USA
Tels: +1 315-464-1690; +1 315-464-7664
E-mails: kotulal@upstate.edu; kuznetsov@upstate.edu

***Present address**

Department of Cancer Biology, Wake Forest University School of Medicine, Winston-Salem, NC, 27101, USA

Angelina Regua and Csaba Papp contributed equally to this article.

(Received 1 November 2021, accepted 29 December 2021, available online 26 January 2022)

doi:10.1002/1878-0261.13175

Despite the current standard of care, breast cancer remains one of the leading causes of mortality in women worldwide, thus emphasizing the need for better predictive and therapeutic targets. *ABI1* is associated with poor survival and an aggressive breast cancer phenotype, although its role in tumorigenesis, metastasis, and the disease outcome remains to be elucidated. Here, we define the *ABI1*-based seven-gene prognostic signature that predicts survival of metastatic breast cancer patients; *ABI1* is an essential component of the signature. Genetic disruption of *Abi1* in primary breast cancer tumors of PyMT mice led to significant reduction of the number and size of lung metastases in a gene dose-dependent manner. The disruption of *Abi1* resulted in deregulation of the WAVE complex at the mRNA and protein levels in mouse tumors. In conclusion, *ABI1* is a prognostic metastatic biomarker in breast cancer. We demonstrate, for the first time, that lung metastasis is associated with an *Abi1* gene dose and specific gene expression aberrations in primary breast cancer tumors. These results indicate that targeting *ABI1* may provide a therapeutic advantage in breast cancer patients.

Abbreviations

ABI1/2/3, abelson interactor protein-1, -2, -3; ABL, tyrosine-protein kinase ABL1; ARP2/3, actin-related proteins ARP2 and ARP3; BRK1, BRICK1 subunit of SCAR/WAVE actin nucleating complex; CK14, cytokeratin 14; CK8, cytokeratin 8; CNA, copy number alteration; CYFIP, cytoplasmic FMR1-interacting protein; DDg, data-driven grouping; DFS, disease-free survival; DMFS, distant metastasis-free survival; ECL, enhanced chemiluminescence; HSPC300, hematopoietic stem/progenitor cell protein 300; KO, knockout; MAPK, mitogen-activated protein kinase; MEF, mouse embryonic fibroblast; MMTV, mouse/murine mammary tumor virus; NCKAP1, NCK-associated protein 1; OS, overall survival; PCA, principal component analysis; PI3K/AKT, phosphoinositide 3-kinase/protein kinase B; PMF, primary myelofibrosis; PyMT, polyoma middle T antigen; RT, room temperature (25 °C); SRA, steroid receptor RNA activator; SWVg, statistically weighted voting grouping; TEB, terminal end buds; WAVE, Wiskott-Aldrich Verprolin homologous protein.

1. Introduction

Breast cancer is the most commonly diagnosed noncutaneous cancer in American women, causing an estimated 200 000 deaths and over 40 000 new diagnoses each year [1]. Despite current treatment modalities that combine surgical intervention, radiation, and adjuvant chemotherapy, many patients relapse after years of treatment and present with metastatic and often incurable diseases. Metastasis of breast tumors accounts for the majority of breast cancer-related deaths [2]. Thus, there is an urgent need to identify novel molecular targets for the development of new treatments against breast cancer.

The critical role of actin polymerization in breast tumor progression and invasion is well established, but the underlying mechanisms remain to be elucidated. Candidate mechanisms of tumor progression involving actin include cell–matrix interactions, invadopodia formation, and increased cell motility, which can all be attributed to increased actin polymerization in invading cells [3,4]. The WAVE complex is a heteropentameric nucleation-promoting factor of F-actin polymerization and comprises WAVE proteins (1/2/3), Abelson interactor (1/2/3), SRA1/CYFIP1, NAP1, and BRK1/HSPC300 [5–7]. These proteins are encoded by genes *WASF(1,2,3)*, *ABI* (1/2/3), *CYFIP(1,2)*, *NCKAP1*, and *BRK1*, respectively [7]. The WAVE regulatory complex in response to RAC1 activation has been proposed to act as a regulator of cell motility by promoting ARP2/3-dependent actin polymerization at the leading cell edge [8,9]. Importantly, increased levels of ARP2/3 and WAVE2 are correlated with an increased risk of invasive breast cancer [10].

The integrity and activity of the WAVE complex are reliant on the presence of all complex members; the loss of any single constituent can lead to altered cell phenotypes [11]. Upstream pathway signaling partners of WAVE complex such as RAC1 [6,12,13] and NUDEL modify its activity [14]. Abelson interactor 1 (ABI1) is crucial for WAVE complex stability and regulation of specific actin-dependent processes such as cell motility and adhesion, macropinocytosis, and embryonic development [15–17]. Our previous studies demonstrated that constitutive *Abi1* loss results in murine embryonic lethality [16]. ABI1 is an adaptor protein that promotes phosphorylation of substrates, such as WAVE2, by ABL kinase and has also been shown to be important for capping of F-actin filaments, thus highlighting its regulatory role in cellular homeostasis and actin turnover [18]. WAVE1 and WAVE2 have differential roles in actin polymerization

output resulting in distinct effect on actin meshwork at the plasma membrane [19,20].

In cancers, WAVE complex's molecular composition is dynamic and can be represented by distinct molecular subcomplexes due to deregulation of component levels [7,11,14]. Furthermore, several cell context-dependent WAVE/ABI1 subcomplexes can form and exhibit distinct functions activated and maintained through different mechanisms [7,11,14,20]. For instance, enhanced levels of *WASF3* gene expression could promote cancer cell invasiveness and are associated with the highly aggressive breast cancer subtypes [21,22]. However, recent studies also demonstrate potential tumor suppressor function of *Wasf3* upon overexpression in PyMT breast cancer cells [22] thus indicating heterogeneity of WAVE3-based complex signaling through differential effect on actin cytoskeleton and cell proliferation [23,24].

WAVE complex dysregulation in cancer provides input into cell cycle progression and warrants the study of its role in breast cancer [25]. Although the specific molecular mechanism has yet to be uncovered, WAVE2 was linked to regulation of cell cycle progression through RAC1, Arp2/3, and ARPIN. Upregulation of the Arp2/3 subunit, ARPC1B, is associated with very poor metastasis-free survival of breast cancer patients, but inhibition of ARP2/3 prevents cycle progression through RAC1 transformation [7,25].

Alterations in *ABI1* expression have been associated with tumor initiation and progression in human cancers, thus indicating that ABI1 protein levels must be tightly regulated in cells. *ABI1* dysregulation has been implicated in several cancers, such as breast, brain, colon, stomach, ovarian, and prostate cancers [26–29]. Notably, the role of *ABI1* in cancer is not always the same; in some cancers, such as PMF, glioblastoma, and prostate cancer, *ABI1* expression is downregulated [26,27,30,31], whereas in breast cancer, *ABI1* expression is enhanced [32], thus suggesting the tissue and disease-involving pathway specificity of the role of *ABI1* in oncogenic transformation and indicating the importance of mechanistic studies. The important role of ABI1 in breast cancer has been established in clinical samples. Previously, immunohistochemical studies of over 900 human breast tumor samples showed that *ABI1* overexpression is positively correlated with poor survival and a shorter relapse time in human breast cancer patients [32]. Indeed, the analysis revealed that invasive breast tumors have higher ABI1 protein expression than poorly invasive tumor samples and that increased ABI1 protein levels are significantly correlated with earlier recurrence and shortened survival.

These findings have been supported by xenograft models of highly aggressive breast cancer cells (MDA-MB-231) lacking *ABI1*, which were unable to grow into large tumors in immunocompromised mice [33]. Taken together, previous data suggest that *ABI1* plays a driving role in the progression of metastatic breast cancers [32–34].

Several *in vitro* studies have shown the impact of ABI1 in driving breast cancer cell motility, division, and invasiveness; however, its exact role during *in vivo* tumor initiation, progression, and metastasis remains to be elucidated. Thus, we aimed to study the impact of *Abil* loss on mammary tumor initiation and progression using the polyoma middle T (PyMT) breast cancer mouse model. The PyMT antigen is a transmembrane scaffolding protein with key tyrosine residues that, upon phosphorylation, can activate signaling pathways involved in cell proliferation and survival (e.g., PI3K/AKT and MAPK), making it a reliable model for aggressive breast tumor formation [35]. The PyMT breast cancer model has been well characterized and recapitulates human breast cancer pathology, especially that of the triple-negative subtype [36].

High levels of ABI1 have been associated with the risks of metastasis of primary tumors and breast cancer mortality, as well as associated with the metastatic phenotype of human breast cancer cell lines *in vitro* [3,32,34,37,38]. The deficiency of ABI1 has been shown to reduce cell migration and invasiveness of aggressive breast cancer cells and is associated with activity in pathways such as PI3 kinase/AKT and SRC [32,33].

Here, we define the ABI1-associated gene expression signature, which predicts the disease metastasis-free survival (DMFS) of patients with primary breast cancer. The signature includes a subset of WAVE complex genes (*ABI1*, *BRK1*, *WASF3*, *CYFIP1*, *CYFIP2*), and the direct interactors of WAVE complex (*RAC1* and *NDEL1*). *ABI1* is an essential component of the signature. To model the role of *Abil* in breast cancer tumor progression and metastasis, we conditionally depleted *Abil* gene expression in the mammary epithelium of PyMT breast cancer mice using the mammary-specific Cre recombinase mouse. Our analysis shows that *Abil* knockout (KO) mice, both with homozygous and heterozygous deletion had more diverse tumor growth kinetics compared to the controls. In KO animals, a significant proportion (between 54% and 64%) of the primary tumors grew slower or not at all. However, the number of identified metastatic foci in lung and their size were significantly reduced in both homozygous and heterozygous KO mice, with the more significant metastasis suppression effect observed in the

former. These results indicate that *Abil* gene dosage in primary tumors is critical for the progression of metastasis in breast cancer. Western blotting analysis of primary tumors supports our previous findings that ABI2 protein expression is increased in animals with homozygous deletion of *Abil*. Collectively, both our analyses utilizing both human breast cancer gene expression data and genetically engineered *Abil* knock-out breast cancer mouse models support the critical role of ABI1 and *ABI1*-based gene prognostic signature as novel biomarkers of breast cancer metastases.

2. Methods

2.1. Reannotation and legacy comparison microarray datasets

The updated and reannotated Rosetta microarray dataset [39] and the Metadata dataset [40,41] were used for the statistical testing and survival prediction analyses. The Metadata dataset is comprised of Uppsala and Stockholm data cohorts, which totals 249 samples (Affymetrix U133A, U133B) [40,41]. Rosetta expression microarray dataset of 295 primary breast cancer samples has been downloaded [39] and reprocessed. Probe sequences (60 bp) obtained from the Rosetta dataset were aligned using NCBI's command line *blastn* program with the following arguments: `-reward 2 -penalty -3 -word_size 11 -gapopen 5 -gapextend 2`. Coordinates with the most significant e-value were used for each sequence. Ensembl GRCh38.p13 was used to annotate each probe's given genome coordinates. RefSeq gene symbols were used to annotate probes that were not annotated by Ensembl and contained RefSeq IDs. A total of 32439 expression data points are present with 24479 unique probes (GSE159956).

Our newly updated Rosetta probe set annotation was compared to the original probe set annotation. Our newly updated Rosetta probe set annotation was compared to the original probe set annotation. A total of 11847/24479 (48.4%) of our probe's gene symbols exactly matched the original gene symbol. A total of 804/24479 (3.2%) probes that have identical gene symbols are on the opposite strand of the given gene. Of the 12632/24479 (51.6%) unique probes that were not an exact match, there were instances where the original set either had a false negative, a false positive, or an alternative gene symbol was used (Fig. S1). For an example of a false negative, see probe 'Contig44690_RC'. In the original probe set annotation, there is no gene symbol for this probe. However, we

found that the gene symbol for this probe was *PTEN*, which was confirmed with the UCSC genome browser.

A total of 7375/24479 (30.1%) probes matched this characteristic. For an example of a false positive, see probe ‘Contig52193_RC’. In the original probe set annotation, there is a gene symbol present for this probe. However, we found that the location of the probe is neighboring the gene body rather than overlapping the gene body. A total of 92/24479 (0.003%) probes matched this characteristic. The remaining percentage of probes that do not exactly match are either instances where the official gene symbol has been updated or instances where a single probe maps to a locus containing multiple genes. For an example of an updated gene symbol, see probe ‘NM_017546’. The original states this probe maps to gene *C40*; however, our method maps this probe to *CNOT11*. Using GeneCards, we found that *CNOT11* and *C40* are the same genes. For an example of a probe that maps to multiple genes, see probe ‘NM_006340’. This probe maps to both *BAIAP2* and *AATK*. The original probe annotation only links this probe to *BAIAP2*. Overall, we improved upon the original probe set annotation by providing gene symbols for a significant portion of false negatives (Fig. S1). Additionally, we show increased consistency of ABI1 expression values between groups following KS-weighted means batch effect correction (Fig. S2).

2.2. Characterization of ABI1 expression, copy number alterations, and associations with breast cancer clinical data

To analyze ABI1 expression, copy number alterations and associations of these characteristics with breast cancer clinical data, we used The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [42] observed in cBioPortal for Cancer Genomics. https://www.cbioportal.org/study/summary?id=brca_metabric. ABI1 profiles from 1904 breast cancer patients including microarray expression, copy number variation, and clinical and cancer samples were downloaded and analyzed.

2.3. Survival prediction analysis and multigene prognostic signature identification

We used our data-driven grouping (DDg) methods (one-dimensional (univariate), 1D-DDg, two-dimensional (bivariate) 2D-DDg), and statistically weighted voting grouping (SVWg) algorithms for patient’s risk stratification onto two and three survival groups representing Kaplan–Meier survival functions (K–M functions) [37,43–45]. These are well-established

statistically based computational methods to identify optimized cutoff values of high-dimensional variable domains that transform large-scale variables to low-dimensional (discrete) scale-independent statistically weighted variables allowing for the selection of the most informative, robust, and reproducible categorical variables with the ability to stratify patient survival risk. In this study, we used an advanced version of the previously published software and algorithm [37,43]. The following is a general description of how patient stratification in the risk groups can be utilized as a measure of survival prediction and as a method of selecting survival predictors (genes) for multivariate prognostic model.

2.3.1. 1D-Data Driven grouping (1D-DDg)

Assume a gene expression data set with $i = 1, 2, \dots, N$ genes whose intensities are measured for $k = 1, 2, \dots, K$ patients. The log-transformed intensities of gene i and patient k are denoted as $y_{i,k}$. Associated with each patient are a clinical outcome continuous data (e.g., survival time) and a nominal (yes/no) clinical event (e.g., tumor recurrence). Assuming that K clinical outcomes are negatively correlated with the vector of expression signal intensity y_i of gene i , patient k can be assigned to the high-risk or the low-risk group according to

$$x_k^i = \begin{cases} 1(\text{high risk}), & \text{if } y_{i,k} > c^i, \\ 2(\text{low risk}), & \text{if } y_{i,k} < c^i, \end{cases} \quad (1)$$

where c^i denotes the predefined cutoff of the i th gene’s intensity level. The clinical outcomes or events are subsequently fitted to the patients’ groups by the Cox proportional hazard regression model [46]:

$$\log h_k^i(t_k | x_k^i, \beta_i) = \alpha_i(t_k) + \beta_i \cdot x_k^i, \quad (2)$$

where h_k^i is the hazard function and $\alpha_i(t_k) = \log h_0^i(t_k)$ represents the unspecified log-baseline hazard function; β is the $1 \times N$ regression parameters vector; and t_k is the patients’ survival time. To assess the ability of each gene to discriminate the patients into two distinct genetic classes [defined by Eqn (1)], the Wald statistic (W) [46] of the β_i coefficient of the model Eqn (2) is estimated by using the univariate Cox partial likelihood function [47], estimated for each gene i as

$$L(\beta_i) = \prod_{k=1}^K \left\{ \frac{\exp(\beta_i^T x_k^i)}{\sum_{j \in R(t_k)} \exp(\beta_i^T x_j^i)} \right\}^{e_k}, \quad (3)$$

where $R(t_k) = \{j : t_j \geq t_k\}$ is the risk set at the time t_k and e_k is the clinical event at the time t_k . The actual fitting of the model Eqns (2,3) is conducted by the survival package in R (<https://cran.r-project.org/web/packages/survival/index.html>). The genes with the largest β_j Wald statistics are assumed to have better group discrimination ability and thus called survival significant genes. These genes are selected for further confirmatory analysis or inclusion in a prospective gene signature set. Note that log-rank statistics were also included in our algorithm and shown similar or in some cases slightly better P -value.

Note, how the stratification of patients in Eqn (1) depends on predefined cut-off values (c^i). In most real-world scenarios, such values are not known in advance. Our 1D-DDg method builds on the described workflow, by identifying the ideal cutoff without needing any prior information. First, for each gene i , we compute the tenth quantile (q_{10}^i) and the 90th quantile (q_{90}^i) of the distribution of K^* signal intensity values. For every value, the algorithm performs the splitting of patients (1), fits the clinical event to the patient groups Eqn (2), and finally calculates the Wald statistic of β_i Eqn (3). In other words, within (q_{10}^i, q_{90}^i), we search for the value* which corresponds to the minimum β_i^z P -value (here $z = 1, \dots, Q$) and that most successfully discriminates the two unknown risk groups.

We note that at the time of patient stratification we cannot tell which group is associated with higher or lower risk. The 1D-DDg method predicts risk by analyzing the survival times of the groups. The group with lower mean survival times will be classified as 'higher risk', while the group with higher mean survival will be labeled as 'lower risk'. According to this classification, two possible relationships exist between patient risk (lower risk, higher risk) and the expression pattern of a given gene (higher expressed, lower expressed). In the case of a parallel pattern, 'higher risk – higher expression' or 'low risk – low expression', the relatively higher prognostic gene expression level is associated with the poorer prognosis (a gene exhibits pro-oncogenic behavior). In the case of antiparallel pattern 'higher risk – low expression' or 'low risk – high expression', the relatively higher prognostic gene expression level is associated with better prognosis (a gene exhibits tumor suppressor-like behavior).

In our current work, the Rosetta and MetaData cohort datasets were used that contain both expression microarray data and the corresponding clinical information. Our survival prediction analysis was focused on the identification of the shortlist of survival significant genes of the *WAVE* complex, *RAC1* and *NDELI*, all of which encode proteins constituting or interacting with

the *WAVE* complex. Input list of the genes includes genes *WASF(1,2,3)*, *ABI (1/2/3)*, *CYFIP(1,2)*, *NCKAPI*, *BRK1*, *RAC1*, and *NDELI*, represented by the probe and probe sets localized in the 3'UTR of the selected genes on both microarray platforms.

The mRNA expression profiles of the selected genes are considered as putative predictors of the disease outcome. The 1D-DDg analyzed the survival prediction property of the Rosetta and Metadata expression microarray signals corresponding to *WAVE* complex members and also *RAC1* and *NDELI* as independent variables. An expression signal, called prognostic variable, is selected for further analysis if in both cohorts the DDg cutoff value(s) provide discrimination of the patients onto survival risk groups at $P \leq 0.05$. To keep a reasonable compromise between sample size, the imbalance of distinct risk groups in a patient cohort, reproducibility across the different cohort and prognostic significance of the putative prognostic variables selection step, we were also allowed to include in the prognostic variables set up to two variables, if for a given variable in one dataset (e.g., DFS, Metadata) $P \leq 0.15$. Thus, the output of 1D-DDg analysis for Metadata or Rosetta cohorts includes the same list of reproducible prognostic variables (gene IDs) defined by the same gene lists, a similar survival prediction pattern of the identical variable (gene ID), cohort-specific gene expression cutoff values dichotomizing the patients on to relatively low-risk (code 1) and high-risk (code 2) groups [37,43–45].

Next, using the results of 1D-DDg, our two-dimensional grouping (2D-DDg) method, and statistically weighted voting grouping (SWVg) we constructed the robust and synergistic multigene prognostic signature. The ability of individual prognostic variables (voting weight) to stratify patients in risk groups is represented by the P -values associated with log-rank statistics.

2.3.2. Statistically Weighted Voting grouping (SWVg)

SWVg is an automatic method of prognostic feature selection and disease risk prediction that allows the construction of an optimized, multivariable, prognostic classifier [37]. The input data were provided by the 1D-DDg method. The ability of individual prognostic variables to stratify patients is represented by the P -values associated with the Wald statistic (calculated in the 1D-DDg). These P -values are used to calculate the relative weight of individual variables in the multivariable classifier. This information is used to construct a decision rule and to assign a patient to one of the risk subgroups.

In practice, the list of genes is ordered in ascending order according to the P -values generated from 1-DDg. The weight w_j is calculated by the formula

$$w_j = \frac{-\log(P_j)}{\sum_{m=1}^N (-\log(P_m))}, \quad (4)$$

where P_j is the P -value of gene j in the 1D-DDg procedure. Then, the new numeric grouping value for sample i could be calculated by the formula

$$G_i^N = \sum_{j=1}^N w_j G_{ij}, \quad (5)$$

where N is the number of genes and G_{ij} is the group allocation for sample i assigned by gene j in the 1D-DDg. In the case that samples are divided into two groups, patient i could be separated into two groups (2 = 'high-risk', 1 = 'low-risk') at a predefined cutoff value (G_c) of G_i^N with the following:

$$y_i^N = \begin{cases} 1(\text{high-risk}), & \text{if } G_i^N > G_c \\ 0(\text{low-risk}), & \text{if } G_i^N \leq G_c \end{cases} \quad (6)$$

A Cox proportional hazard regression model is estimated by using a univariate Cox partial likelihood function with the method described in the 1D-DDg procedure. Wald statistic of $\hat{\beta}^j$ is estimated and serves as an indicator to evaluate the ability of group discrimination for gene j at cutoff G_c . The searching space of G_c is from 0.2 to 0.8, with an increment of 0.01 for each step. The G_c that provides the minimum log-rank P -values in the searching space is the optimized G_c . The above-described procedure is repeated for different N , which varies from 3 to the number of genes assigned. The number (N_{opt}) and combination of genes are optimized for minimum log-rank P -values.

A similar procedure is applied when the samples are divided into three groups. Two cutoff values (G_{c1} , G_{c2} , $G_{c1} < G_{c2}$) of F_i^N selected and then used to calculate the grouping variable according to the following formula**:

$$y_i^N = \begin{cases} 1 \text{ (low risk)} & \text{if } G_i^N > G_{c2} \\ 2 \text{ (intermediate risk)} & \text{if } G_{c1} < G_i^N \leq G_{c2} \\ 3 \text{ (high risk)} & \text{if } G_i^N \leq G_{c1} \end{cases} \quad (7)$$

A Cox proportional hazard regression model and log-rank statistic estimates are computed. G_{c1} in Eqn (7) is searched in the range of 0.2 and 0.44, with an increment of 0.01 for each step, while G_{c2} is searched in the range 0.56 and 0.8, with an increment of 0.01 for each step. G_{c1} , G_{c2} are optimized for the minimum value of summation of pair-wise log-rank P -values of three survival curves.

The most significant and robust cut-off value does not always result in balanced groups (i.e., one group may only contain a few patients). In our work, we aim to define risk groups that the smallest group contains at least 10% of the total patients. In cases where 'the best' cut-off value resulted in unbalanced groups, and other values could stratify patients with statistically significant Wald statistics, we opted to use these alternatives.

Note that before the execution of Eqn (7), we recode the group values of the high-risk group from 2 to 0. Using the modified values to calculate G_i^N in Eqn (5) will result in G_i^N closer to 0 for patients with higher risk. Conversely, patients with lower risk will have G_i^N closer to 1. As a result, patients for whom $G_i^N > G_{c2}$ is true will constitute the lower risk group. Patients with G_i^N that is below G_{c1} will be in the high-risk group. Patients whose G_i^N falls between G_{c1} and G_{c2} are classified as moderate risk.

To construct the multivariate prognostic signature, the SVWg starts with paired gene expression data using the two-dimensional grouping (2D-DDg) method [37,44,45]. For the given two variables domain and the 1D-DDg determined cutoff values of these variables, the 2D-DDg identifies two mutually excluded subdomains in the 2D domain that maximize discrimination of all subjects (patients) onto low- and high-risk groups. The possible distinct subdomain combinations in the 2D domain are called 'designs/models' of the patient's grouping.

SWVg adds the next prognostic variable that increases differentiation between risks of the groups and allows a selection of a synergistic multivariable signature based on the summation of the statistically weighting variables in a stepwise multivariable fashion. Less stringent statistical criteria (weights) are used by SWVg when the next most significant prognostic variable is added to the survival prediction model. The sample size and data quality constraints are included in the algorithm allowing the SWVg to minimize the number of prognostic variables (predictors) and reduce the signature identification overfitting risk keeping high confidence and reproducible prognostic multivariate model.

The multivariate method starts with the most significant prognostic variable (1-st rank predictor) paired with the next most significant predictor. These features in combination provide a synergistic effect, robust prognostic signature, and provide consistency between the signatures derived in the cohorts.

2.3.3. Optimization of the 2D-DDg method for correlated covariates (gene expression value pairs)

In many datasets, the gene pairs (A and B) expressions may be correlated positively or negatively due to some

context regulatory mechanisms (interaction due to common medical condition(s) and similar treatment). The paired correlation analysis could be used to improve the significance and robustness of patient's risk group stratification. In this section, we describe an extension of our 2D-DDg method.

Let N denote the number of nonduplicated samples of the population (patient cohort). Let $\{X, Y\}$ denote the N random variable (r.v.) pairs (e.g., gene expression levels in the N samples that associated with N patient survival data (event and time after disease diagnostics or last follow up)), where the expression levels X and Y of the genes A and B , respectively. If the correlation measure between r.v. X and Y significant, the DDg defined risk group separation cutoff value (gene expression value determined a patient to the given risk groups) of bivariate r.v., could be optimized due to the variable's dependence. In such cases, we can define the 'interaction effect' (synergy) between A and B data into 2D-DDg prediction analysis as follows.

The method calculates the Kendal tau (or Spearman) correlation coefficient between all possible paired of r.v., specifies significantly correlated pairs, and then parameterizes the linear regression model quantifying the stochastic association between two r.v.

$$Y = \alpha + \beta X + \varepsilon, \tag{8}$$

where x is the vector of gene A expression values, $x = \{x_1, x_2, \dots, x_N\}$; y is the vector of gene B expression values, $y = \{y_1, y_2, \dots, y_N\}$; ε represents an additive error term that may stand un-modeled determinants or random statistical noise: $\varepsilon = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N\}$ N is the number of samples; α and β are parameters of the linear regression model. α is a y-intersect of the line and β is a slope of the line. We estimate the parameters using the least squares method. The estimated parameter values denote as $\hat{\alpha}$ and $\hat{\beta}$.

Using parameterized Eqn (8) for the vector component pair $(X, Y - \hat{\alpha})$ defined in the form

$$(y_i - \hat{\alpha}) = \hat{\beta}x_i, \quad i = 1, 2, \dots, N, \tag{9}$$

we calculated the shortest distance of a particular point Q (x,y) from the regression line. To do this, we use a rotation of orthogonal coordinate system formula of point Q $\{x, y - \hat{\alpha}\}$ as the following

$$\bar{x}_i = x_i \cos\gamma + (y_i - \hat{\alpha}) \sin\gamma, \tag{10}$$

$$\bar{y}_i = -x_i \sin\gamma + (y_i - \hat{\alpha}) \cos\gamma, \tag{11}$$

Where $\{\bar{x}_i, \bar{y}_i\}$ are the coordinates of point Q in the new orthogonal coordinate system rotated on the angle γ . Using trigonometric formula, $\hat{\beta} = \tan\gamma$, and we obtain.

$$\bar{x}_i = (x_i + \hat{\beta}(y_i - \hat{\alpha})) / (1 + \hat{\beta}^2)^{1/2}, \tag{12}$$

$$\bar{y}_i = (-\hat{\beta}x_i + (y_i - \hat{\alpha})) / (1 + \hat{\beta}^2)^{1/2}, \tag{13}$$

Equations (12,13) are used in our study for the calculation of new coordinates of the objects and corrected cutoff values $C\{\bar{x}^*, \bar{y}^*\}$ defined by DDg for prediction of the low- and high-risk groups in the patient cohort.

We included and used our rotation of orthogonal coordinate system approach in the 2D-DDg method to improve the significance of the patient's separation on the relatively low- and high-risk groups. Our analysis showed that in the high-correlated genes, this method improves the statistical significance of results obtained in DDg methods, but also could lead to more robust grouping and reproducibility of the risk model across distinct patient cohorts. For instance, in the case of *AB11- BRK1* pair of the Rosetta cohort, our standard 2D-DDg survival prediction analysis of DMFS provided the borderline statistical significance of patients grouping ($P < 0.05$). However, a strong positive correlation between expressions of these two genes was found ($P < 0.0001$), suggesting common coregulatory mechanisms.

2.3.4. Prognostic models, correlations, and reproducibility of the AB11-based prognostic signature genes

According to our selection criteria of prognostic variables (Methods), 1D-DDg selected 5 genes of WAVE complex (*AB11, BRK1, CYFIP1, CYFIP2, and WAVE3*) and 2 genes (*RAC1* and *NDEL1*) encoding the proteins RAC1 and NADEL exhibit 'interaction' with WAVE complex components. Our 7 genes were representative be unique probe sets on Affymetrix U133 A&B and Rosetta microarray platforms (Table S1). Figs. S3,S4 and Table S2 show that across different microarray platforms DFS and DMFS survival patterns *AB11, BRK1, CYFIP1, and RAC1* are commonly reproducible and classified as pro-oncogenic, while *CYFIP2, NDEL1, and WAVE3* are mostly classified as tumor suppressor-like genes. However, because the system of interactive molecules is open, stochastic, and nonlinear for some genes (e.g., *WAV3*), the variations of these prognostic properties (as a

component of the system) could be unstable and expressed alternative functions.

Note that over data sets and event types (e.g., DFS, DMFS), the prognostic pattern of expression changes for some individual genes (e.g., *WASF3*) was in some cases not the same (classified as proto-oncogene or tumor suppressor like). However, the pro-oncogenic pattern (upregulated expression–poor prognosis) of *AB11*, *BRK1*, and *RAC1* or the ‘tumor suppressor’ pattern (upregulated expression–good prognosis) of *NDELI* and *CYFIP2* expression was highly reproducible between our datasets.

In the context of co-expression, the METABRIC, Rosetta, and Metadata data, *AB11* expression is positively correlated with the expression of *BRK1*, *CYFIP1*, and *NDELI*. It is also not correlated with *RAC1* expression and is negatively correlated with *CYFIP2* expression.

2.4. Mouse primary tumors RNA-seq

Gene expression profiles from primary breast tumors of PyVT heterozygous and homozygous mice with and with *Abil* disruption were detected with the Illumina NextSeq platform (GSE162815). Two tumors from a single mouse from each of the four groups were sequenced. Two runs were performed on consecutive days for increased depth. Illumina’s breast cancer12-fastq program was used for the conversion of base calls to FASTQ files. This resulted in two read files due to the paired-end sequencing protocol. STAR was used to align sequences to the GRCm38/mm10 mouse genome. STAR was also used for the quantification of reads per gene. Raw counts between tumors and days for each genotyped were summed for maximum depth. Fold change was calculated between *Abil* wild-type and *Abil* knockout mice for each genotype.

2.5. Animals

Transgenic PyMT mice (JAX no. 022974; C57BL6) and mammary-specific *Cre* mice (JAX no. 003553, Line D; mixed strain) were purchased from Jackson Laboratory. *Abil*-floxed mice were generated by the Kotula Laboratory [16] (MGI : 4950557; *Abil*^{tm1.1Lko}; C57BL6). Female PyMT mice with conditional *Abil* knockout were generated by crossing PyMT transgenic males to homozygous *Abil* females to produce PyMT transgenic males heterozygous for *Abil* floxed allele (PyMT; *Abil* fl/wt). PyMT; *Abil*(fl/wt) males were backcrossed to homozygous *Abil* females to produce PyMT; *Abil*(fl/fl) males. In parallel, transgenic *Cre* animals were crossed with homozygous *Abil* animals to generate transgenic *Cre* animals heterozygous for

Abil [MMTV-*Cre*; *Abil*(fl/wt)]. To generate experimental animals, male PyMT; *Abil*(fl/fl) were crossed to female MMTV-*Cre*; *Abil*(fl/wt). All breeders used were at least 8 weeks of age. Genotyping was performed using ear snips (Transnetyx, Cordova, TN). As mammary glands were the tissue of interest, only female experimental animals were analyzed. Female animals were sacrificed at designated time points (5, 7, and 12 weeks, for developmental studies, $n = 5$ animals per genotype; or seven weekly time points starting with the tumor detection (at week 0, 1, 2, 3, 4, 5, and 6, $n \geq 6$ mice), when tumors reached 2.0 cm, or when animals displayed signs of distress as per the guidelines of the National Research Council Committee on Recognition and Alleviation of Distress in Laboratory Animals. For primary and lung metastasis tumor studies, animals ($n \geq 6$ mice) were sacrificed age between 17 and 26 weeks. All animals used in the studies described herein were housed in ventilated microisolator caging under HEPA-controlled environmental conditions and maintained under the supervision of the SUNY UMU Institutional Animal Care and Use Committee (IACUC no. 393).

2.6. Tumor palpation and measurements

Starting at weaning age, all female PyMT animals were palpated and measured for tumors biweekly. Tumor measurements and volume calculations were performed as previously described [48]. Total tumor burden over time was calculated for each animal ($n = 6$ /genotype) and was plotted against the time since primary breast tumor was detected by palpation.

2.7. Mathematical Models and Estimated parameters in analyses of primary tumor kinetics and pulmonary metastatic node’s size frequency distribution

We estimated parameters of the tumor volume kinetics using the exponential function

$$f(t; a, b) = (a - b) * \exp(ax),$$

where t is time, parameter a is the rate of cell volume growth, and $(a - b)$ is the initial tumor volume. SIGMAPLOT-13 software was used to perform nonlinear regression analysis and the results visualization.

2.8. Proliferative activity and histological parameters of the mouse primary tumors

Mouse mammary parenchyma less than (<20 weeks) or greater than 20 weeks (>20 weeks) of age were

examined by a blinded pathologist. Histologic sections of healthy control, homozygous control, heterozygous control, homozygous *Abil*(KO, -/-), and heterozygous *Abil*(KO, +/-) breast parenchyma were compared. The murine grades were determined according to published histologic criteria [35]. The Ki-67 index is expressed as percent positivity from 500 nuclei counted in areas of highest positivity. The comparative analysis was performed for each group of mice vs normal and corresponding negative contours (breast parenchyma) of heterozygous *Abil* (KO, +/-) and homozygous *Abil* (KO, -/-) samples. Unpaired nonparametric Mann–Whitney U-test was performed at $P < 0.05$.

2.9. Lung metastasis quantification

To quantify the metastatic area throughout the lung tissue, three 5 μ m sections from formalin-fixed paraffin-embedded mouse lungs (sectioned every 50 μ m) were collected from each group ($n \geq 6$ animals per genotype), stained with hematoxylin and eosin and imaged using an Omnyx digital pathology scanner (GE Healthcare, Boston, MA, USA) [49]. Images were quantified for the total number of metastatic foci using IMAGEJ software (NIH) and subjected to statistical analyses.

2.10. Western blot analyses

Western blots were performed as previously described [16]. Blots were probed with the following primary antibodies: ABI1 (Rockland, Pottstown, PA, USA; 1 : 1000), ABI2 (P-20, Santa Cruz Biotechnology, Dallas, TX, USA; 1 : 500), ABI3 (GeneTex, Irvine CA, USA; 1 : 1000), WAVE1 (K91/36, MilliporeSigma, Burlington, MA, USA; 1 : 1000), WAVE2 (H-110, Santa Cruz Biotechnology, Dallas, TX, USA; 1 : 1000), WAVE3 (W4642, Sigma-Aldrich, St. Louis, MO, USA; 1 : 1000), or β -actin (AC-15, Sigma-Aldrich; 1 : 10 000). Blots were incubated with SuperSignal West Pico or Femto ECL reagents (Thermo Fisher, Waltham, MA, USA) and imaged using a PxiTouch imaging system (SynGene, Bengaluru, Karnataka, India).

2.11. Immunohistochemistry, histology, and whole mount analysis

Immunohistochemical staining was performed with antigen retrieval following standard protocols. Tissue sections of normal mammary tissue were stained with anti-CK8 (TROMA-I, DSHB, Iowa, 1.1000) and anti-CK14 (PRB-155P, Covance, 1.250). Tumor sections (≥ 3 animals/genotype) were stained with the following antibodies : ABI2 (P-20, Santa Cruz Biotechnology, 1.250),

WAVE1 (K91/36, Millipore, 1.250), WAVE2 (H-110, Santa Cruz Biotechnology, 1.250), and WAVE3 (Abreast canceram ab110739, 1.100). Stained sections were mounted on coverslips using Cytoseal XYL (Fisher) and imaged using a Nikon Eclipse Ci-L upright microscope. Formalin-fixed tumor specimens were stained with hematoxylin and eosin for histopathologic review. Grading of murine tumors was performed according to Fluck and Schaffhausen's review of the model pathology [35]. Briefly, tumors were assigned a score of 0 (normal breast parenchyma), 1 (mammary hyperplasia consisting of dense lobules), 2 (mammary intraepithelial neoplasia; the murine correlate of ductal carcinoma *in situ*), 3 (early carcinoma characterized by early stromal invasion), or 4 (late carcinoma). The mitotic rate was determined by counting the number of mitotic cells in 10 high-power fields (hpf). The mitotic rate was calculated for the areas of the tumor with the highest grade. Tumor sections were also stained with Ki-67, with nuclei in cells in the highest grade areas counted to determine expression, which was reported as the percentage of positivity.

For whole mount staining, mammary glands were processed as previously described [50]. Stained whole-mounted tissues were imaged using a Nikon D610 camera, and images were subjected to morphometry using ImageJ software (NIH). Terminal end buds, ductal length, and ductal branching were quantified as previously described [51,52].

2.12. Statistics

Each cellular or biochemical experiment had technical ($n \geq 3$) and biological ($n \geq 3$) repeats. To determine statistically significant differences involving more than 2 biological groups, we used 1-way and 2-way ANOVA followed by *t*-test, nonparametric tests, generalized univariate and multivariate linear models, correlations other analyses as stated elsewhere in the manuscript using Statistica 13, StatSoft); *P*-value less than 0.05 was considered significant. Categorical data analyses were carried out using Sytel Studio-9 software (SyteL Inc. Pune). Kinetic analysis and nonlinear models parameterization were done using SigmaPlot-13 (SYSTAT Software takes) software.

2.13. Ethics approval and consent to participate

All animal studies were performed according to guidelines approved by the Institutional Animal Care and Use Committee of SUNY Upstate Medical University (Protocol no. 393). Publicly available datasets were used for all patient-associated bioinformatics analyses in this manuscript.

3. Results

3.1. Upregulation of ABI1 gene expression in primary breast cancers correlates with aggressive, basal-like phenotype and metastatic predisposition

ABI1 is an essential part of the WAVE regulatory complex, a major promoter of actin filament nucleation is often exploited by invasive tumor cells [53]. To elucidate the significance of the ABI1 in the pathobiology of human breast cancer, we carried out a retrospective analysis of METABRIC data of 1904 breast cancer patients (Fig. 1). We found that expression of ABI1 in primary tumors is strongly associated with copy number alterations (CNA) (Fig. 1A), overexpression with histologic grade 3 (Fig. 1B). There was a significant negative correlation ABI1 mRNA as well as ABI1 CNA with ER (+) status (Fig. 1C) and no correlation with the lymph node (LN) status of the patients (Fig. S5).

Moreover, ABI1 overexpression is associated with highly aggressive (grade 3) basal-like and claudin-low breast cancer subtypes (Fig. 1D). Additionally, using cBioPortal for Cancer Genomics tools (<https://www.cbioportal.org/>), we observed that high-expressed and gained and amplified CNA *ABI1* are significantly enriched in the high genome instability integrative cluster 10 [42]. The cluster 10 molecular subtype is enriched by basal-like cancer subtype tumors and clinically defined as triple-negative, highly aggressive, drug resistance, and high-risk metastasis tumor genes that includes numerous signaling molecules, transcription factors, mitotic, and other cell division genes associated in trans with this deletion event in the basal

cancers, including alterations in *AURKB*, *BCL2*, *BUB1*, *FOXM1*, *KIF2C*, *KIF1C*, *RAD51AP1*, *TTK*, and *UBE2C*. Notably, many of these molecules are included genetic grade and poor survival outcome signatures [40,42,54,55]. For instance, TTK (MPS1), a dual-specificity kinase that assists AURKB in chromosome alignment during mitosis and promotes aneuploidy in breast cancer [42].

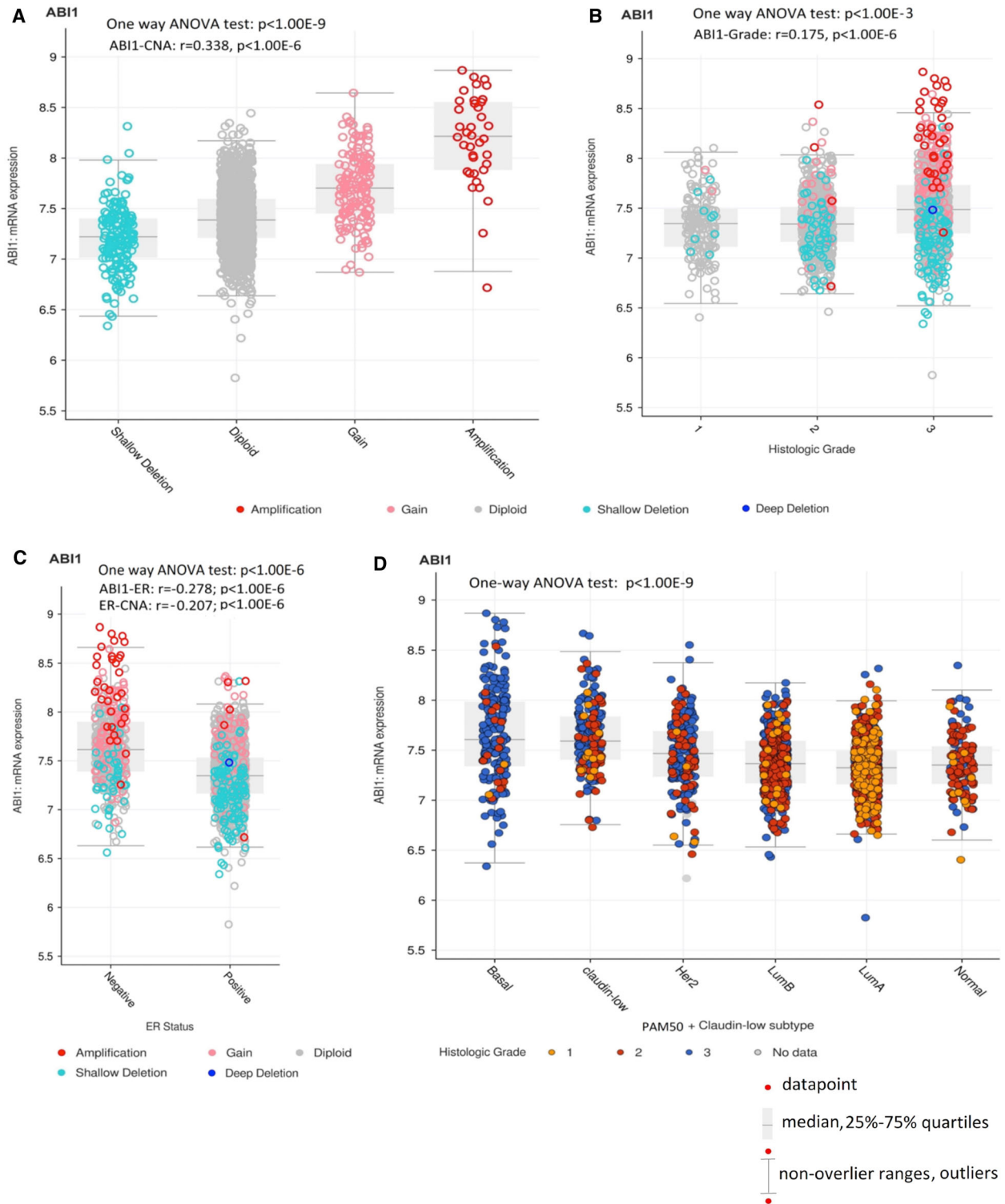
Thus, ABI1 expression shows strong positive correlates with histologic grading, negative correlation with ER status, and represents correctly the known ranked-order of breast cancer subtypes according to their genetic grading classification (Fig. 1A–D; [40,42,52,54,55]). These findings allow us to consider ABI1 transcription level as a functional score of indicating (a) this gene locus instability, (b) ER(-) status of the primary tumor, (c) histologic grading system estimator, and iv) a genetic variable that represents correctly known ranked-order of breast cancer subtypes that reflect genetic grading and drug sensitivity/resistance of the tumor subtypes/groups.

Additionally, multivariate testing *ABI1* expression variation as a random function of CNA, ER status, and tumor subtypes showed that CNA in basal-like tumor subtype samples provides a major explanatory contribution of *ABI1* expression variation in our data ($P < 1.00E-6$; two-way ANOVA, Statistica 13).

3.2. Survival prediction analysis identifies ABI1 as breast cancer metastasis prognostic marker and an important component of the multigene metastasis prognostic signature

We analyzed associations of survival data with microarray gene expression profiles of well-established

Fig. 1. *ABI1* expression alteration is associated with copy number alteration (CNA) and high-aggressive basal-like breast cancer. Box Plots: (A) Putative *ABI1* DNA copy number alteration (CNA) drives *ABI1* transcription level in subpopulations of primary breast cancer patients [42]. The gene expression, CNA, tumor samples, and clinical datasets representing 1904 primary breast cancer samples were downloaded from METABRIC dataset (<https://www.cbioportal.org/>). CNA categorization is the following: shallow deletion: 1 ($n = 166$), diploid: 2 ($n = 1554$), gain: 3 (146), and amplification: 4 ($n = 38$). One-way ANOVA test (Statistica 13) showed significant differences in the *ABI1* expression between the groups and also in the entire cohort ($P < 1.00E-9$). Furthermore, the transcription level of *ABI1* is highly significant and positively correlated with CNA ($r = 0.338$; $P < 1.00E-6$; estimated by Spearman). (B) *ABI1* transcription level positively correlated with histologic grades (univariate and bivariate linear regression models testing shown significance at $P < 1.00E-6$), however (C) negatively correlated with ER status. Bivariate linear regression models (Statistica 13) showed that both expression *ABI1* expression level and CNA are significant ($r = -0.278$; $P < 1.00E-6$ and $r = -0.207$, $P < 1.00E-6$ respectively); however, the *ABI1* expression provides a major contribution in the bivariate linear regression function). Correlate coefficients in (B) and (C) were calculated by Kendall. (D) *ABI1* overexpression is associated with basal-like and claudin-low breast cancer subtypes and aggressiveness of breast cancer scoring also by histologic grade. PAM50 (Basal-like, HER2(+), luminal B, luminal A, normal-like), and claudin-low subtypes were ranked-order according to the trend of decreasing of *ABI1* expression. One-way ANOVA test (Statistica 13) showed significant differences in the *ABI1* expression between basal-like, claudin-low subtypes and other subtypes ($P < 1.00E-6$). *ABI1* expression in the HER2 subtype was significantly higher than in luminal B or luminal A tumor subtypes ($P < 1.00E-6$) and higher but less significant than in the normal-like tumor subtype ($P = 0.01$). A negative trend in the *ABI1* expression across rank-ordered tumor subtypes was mostly defined by relative overexpression of Basal-like and claudin-low tumor subtypes; it was highly significant (one-way ANOVA; $P < 1.00E-9$; Statistica 13).



publicly available breast cancer datasets 39–41. These datasets were used to construct our Metadata and Rosetta microarray datasets (Methods, Supplementary Methods).

Firstly, we focused on the identification of the role of ABI1 expression in breast cancer survival associated with cancer progression/recurrence (defined DFS time) and metastatic process (DMFS time). Table S1

provides *ABI1* annotation and unique probe sets on Affymetrix U133 A&B and Rosetta microarray platforms utilized in our analysis. For stratification of the patients onto risk groups, we utilized 1D-DDg, which approximates patient risks by analyzing the survival time functions of two (or more) patient groups given by the prognostic variable cutoff value(s) estimated statistically in a given patient cohort (Methods, Supplementary Methods). The examples of implementation of 1D-DDg results for Rosetta and Metadata cohorts are presented in Figs. S3 and S4. Each figure shows the gene panels of two K-M plots of disease-free survival (DFS) (Fig. S3) and distant metastasis-free survival (DMFS) (Fig. S4), respectively. The groups of the patients assigned to relatively low-risk (step function line indicated by black color) and high-risk (step function line indicated by red color) K-M survival functions are defined by the gene expression cutoff value calculated by 1D-DDg. The group with higher mean survival time is labeled as 'low risk', while the group with lower mean survival time is labeled as 'high risk'. According to this classification, two possible relationships exist for the patients with lower and higher risks and the expression pattern of a given gene (higher expressed, lower expressed). In the case of a parallel pattern, 'higher risk – the higher the expression' or 'low risk – the lower the expression', the relatively higher prognostic gene expression level is associated with the poorer prognosis (a gene exhibits pro-oncogenic behavior). In the case of antiparallel pattern 'higher risk – the lower the expression' or 'lower risk – the higher the expression', the relatively higher prognostic gene expression level is associated with better prognosis (a gene exhibits tumor-suppressive-like behavior).

Importantly, the prognostic association 'higher risk – the higher the expression' of *ABI1* was statistically significant and reproducible over breast cancer cohorts (Figs. S3–S4). These results consist of our 1D-DDg analysis of the gene expression of the *ABI1* gene and *ABI1* protein in breast cancer patients found in RNA-seq and proteomics databases (Fig. S6).

We propose that the high-risk group of patients in our cohorts is associated with a higher frequency of metastatic events. Indeed, the metastasis event enrichment analysis (Table S3) showed that in both cohorts, the higher risk group was significantly enriched by metastatic events vs. the lower risk group. The fold change (FC) enrichment of metastatic event and *P*-value was calculated using the exact test of two binomial distributions that showed FC = 1.38, *P* = 0.05 in Rosetta and FC = 1.96, *P* = 0.033 in Metadata dataset, respectively.

Next, we used the results generated by the 1D-DDg survival prediction method, which automatically selects survival significant prognostic variables (survival significant genes represented by microarray probes), as the input data for the 2D-DDg [29,32] that identifies the interaction effect between paired prognostic variables (gene pairs) [29,32]. Fig. S7 shows the result of the implementation of 2D-DDg survival prediction to Rosetta data (DFS and DMFS, respectively). These results show that in most gene pairs *ABI1* improves the balance between risk groups and in some cases the bivariate partition of the patients provides more confident risk group differentiation. Similar results were observed for the Metadata data set (not shown).

Interestingly, in both our cohorts, *ABI1* expression is positively correlated with the expression of *BRK1*, *CYFP1*, and *NDELI*, but is not significantly correlated with the expression of *CYFP2*, *WASF3*, and *RAC1* (*P* < 0.05, Spearman). These findings in most cases consist of the correlation analysis of *ABI1* and other gene expression from the METABRIC datasets (Table S4). Note *WAVE3*, *CYFP1* prognostic models may be more data variation- and noise-sensitive.

Finally, the SWVg algorithm was used to construct a survival group prediction model based on the combinations of 1DDg-defined gene expression level models. Fig. 2A–D shows that in both Rosetta and Metadata cohorts, the method revealed a high confidence stratification of the patients onto three risk groups with high, intermediate, and low metastasis-free survival time, called the *ABI1*-based 7-gene prognostic signature. Similar results were obtained for DFS time (results are not shown). Overall, the genes of the *ABI1*-based 7-gene prognostic signature provide robust functional associations and high-confidence survival prediction properties.

3.3. *ABI1*-based prognostic signature as a predictive tool for a metastatic event of breast cancers

Importantly, the *ABI1*-based prognostic signature could serve as a predictive tool for a metastatic event of breast cancer patients. Indeed, Table S5 shows that in the case of DMFS the high- and intermediate-risk groups are highly enriched for patients with metastatic breast cancer events compared to the low-risk group (62% and 44% vs. 15% for Metadata and 79% and 34% vs 15% for Rosetta cohorts). The median and mean time values of metastatic events showed an inverse order in these risk groups. These findings suggest that our signature values defined in primary breast cancer samples can be used for the quantitative

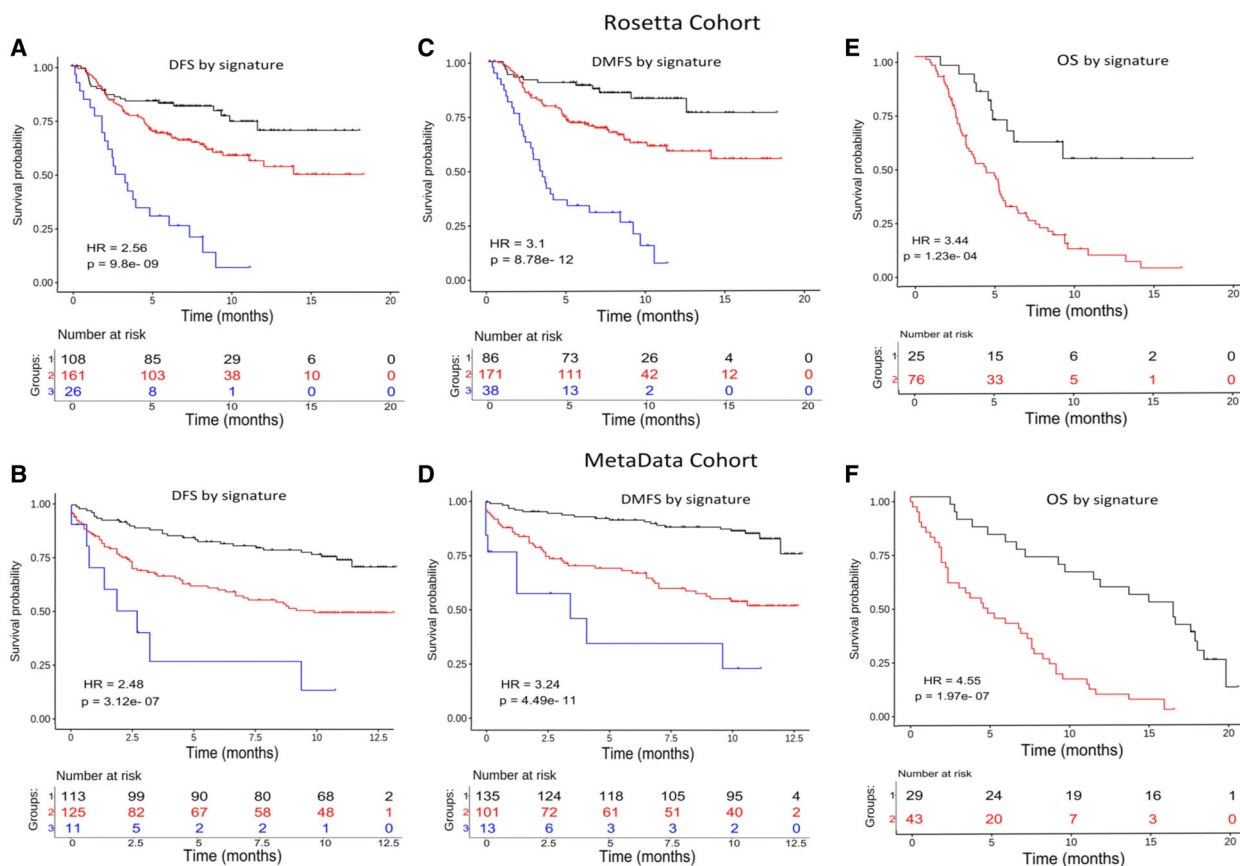


Fig. 2. ABI1-based prognostic signature predicts disease-free and metastatic-free survival risks. The disease-free survival (DFS) and disease metastasis-free survival (DMFS) of patients stratified based on the ABI1-associated signature derived by our survival prognostic analysis method (see Methods for details) is shown using Kaplan–Meier survival curves for Rosetta (A, C) and MetaData cohorts (B, D). The Wald statistic P-value and hazard ratio (HR) associated with the partitioning of the patients into distinct risk groups are also shown (see methods for details). Our method computationally categorizes each covariate (expression level of a gene) as a binarizing risk factor and stratifies each patient according to the multivariate expression pattern of the genes included in the signature (Table S1). In panels A, B, C, and D: black color line = ‘low-risk’, red = ‘intermediate risk’, blue = ‘high-risk’ groups. Panels E and F represent the overall survival (OS) time functions for the patients with metastasis detected after diagnostic and following surgical treatment. The black color line is associated with the group of patients with relatively better disease outcomes, while the red color is associated with patients with poor disease outcomes. The tables at the bottom of plots show the number of patients who survived in the predicted groups more than the given time point.

prediction of distant metastasis events and the time interval of metastatic event occurrence.

Using our ABI1-based prognostic signature genes and specifying their expression cutoff values as done before, we were able to further stratify patients with metastatic events into relatively lower and higher OS time risk groups (Fig. 2E–F). These results suggest that the ABI1 and other genes of the prognostic signature are involved in the progression toward metastatic disease and may be mechanistic regulators of a subset of metastatic breast cancers.

To compare the prognostic significance, we used Rosetta and Metadata cohort’s clinical and gene expression data and compared the ABI1-based 7-gene prognostic signature with commonly used clinical

markers: estrogen receptor (ESR) and lymph node (LN) status (Fig. S5). While ESR status shows significant differences in DMFS in the survival of the Rosetta cohort (low vs. high expression) (Fig. S5A), it was not a predictive factor in the Metadata cohort (Fig. S5C). An opposite prognostic pattern was observed for LN status: It is not significant in the Rosetta cohort (Fig. S5B) but shows prognostic significance in the Metadata cohort (Fig. S5D). Additionally, univariate and multivariate analyses showed that LN and ER status is insufficient for reliable prediction of 3 risk groups (not shown).

Overall, the ABI1-based prognostic signature provided robust, reproducible, and high confidence prediction models of DFS, DMFS, and OS (Figs. 1, 2,

Figs. S3-S4, Tables S2,S5) and demonstrates high performance across different cohorts (Fig. 2; Table S2). Reproducibility of risk stratification of the patients with metastases in the Rosetta and Metadata datasets based on OS time supports this statement. Furthermore, the *Abil*-based signature predicts distant metastatic events more accurately than commonly used clinical factors (Table S6).

3.4. Loss of *Abi1* does not grossly affect the long-term development of normal mammary glands

While implicated in breast tumor progression, the role of AB11 in normal mammary tissue remains unknown. To ensure that phenotypes that may be observed in our *Abil* knockout (KO) breast tumor model result from the effects of AB11 protein loss on tumor progression and not from an otherwise global effect on breast tissue, we conditionally deleted *Abil* from mammary epithelial cells of non-tumor-bearing animals. As with most mammals, mouse mammary gland development occurs postnatally [56]. Mice are born with rudimentary mammary fat pads that develop into functional mammary glands upon the onset of puberty. Beginning at 5 weeks of age, the ductal tree begins to penetrate the mammary fat pad and continues until sexual maturation. This dynamic tissue reconstruction allows for examination of classical mammary structures such as ductal branches and terminal end buds (TEBs) (for an extensive review of mammary gland development, refer to Inman *et al.* [56]). Expression of the mammary-specific CRE recombinase is under the control of the murine mammary tumor virus (MMTV) promoter and begins at ~21 days, allowing us to observe phenotypic changes in normal mammary gland tissue upon AB11 loss [57].

To determine the effects of AB11 loss on the development and structural integrity of normal mammary tissue, a whole mount analysis was performed on the inguinal mammary gland (see Materials and Methods). Gross examination revealed a modest impact of AB11 loss on mammary gland development (Fig. 3A). We examined changes in the total number of terminal end buds (TEBs) as well as the number of ductal branches and total ductal tree length. TEBs are highly proliferative, tear-shaped structures found at the distal end of the ductal tree that penetrate the mammary fat pad to facilitate ductal tree elongation and are involute upon completion of ductal tree extension [56]. Morphometry of mammary gland whole mounts showed a significant increase in the number of TEBs upon homozygous ablation of the *Abil* gene and a trend toward increased

branching, the latter of which did not reach significance (Fig. 3B,C); however, this does not seem to impact long-term gland development, as ductal tree elongation remained unaffected (Fig. 3D). Heterozygous *Abil* KO glands showed sustained TEB counts in 5- and 7-week-old whole mounts (Fig. 3B-D).

In addition to dynamic tissue reorganization, mammary glands also have classically defined ductal structures. Murine mammary ducts are defined as lumens lined by an inner layer of luminal epithelial cells and an outer layer of myoepithelial cells [56]. Thus, analysis of this cellular organization would indicate whether there are organizational defects within the mammary duct upon *Abil* deletion. Gross pathological examination of hematoxylin and eosin (H&E)-stained mammary gland sections show the unaltered organization of epithelial cells and connective tissue within ducts (Fig. 3E). Immunohistochemical staining for cytokeratins 8 and 14, which mark myoepithelial and luminal epithelial cells, respectively, shows similar staining patterns in control and *Abil*-null mice (Fig. 3F) [58]. Taken together, we show that AB11 loss does not affect the long-term mammary gland development of healthy mice.

3.5. AB11 protein level and gene dose regulate tumor growth in PyMT animals

AB11 overexpression has been implicated in promoting an aggressive breast cancer phenotype; however, its exact role in mammary tumor progression is still unclear [32–34]. First, we established that PyMT transgene induces expression of *Abil* in primary tumors vs. normal mammary gland epithelium of *Abil* floxed mice (Fig. 3G); therefore, we concluded that PyMT mouse recapitulates overexpression of *AB11* observed in human tissue, and thus, it is an appropriate model to examine the role of *AB11* in breast cancer tumor progression. To determine efficiency of *Abil* gene loss in our *Abil* KO PyMT animals, we performed deep RNA-seq analysis of representative primary tumors of each genotype (Table 1). We found that *Abil* gene expression follows gene dosage effect as expected: 15.4-fold in homozygotes and 2-fold in heterozygotes vs. their respective controls. Several members of the WAVE complex were modestly downregulated or retained their expression in *Abil* KO tumors vs. controls, while *Wave3* (*Wasf3*) was upregulated and *Cyfp2* was downregulated. An opposite effect on several WAVE complex genes expression in the heterozygous vs. homozygous animals was apparent (Table 1).

Interestingly, comparative analysis of the basal-like vs. luminal breast cancer cell types markers in AB11

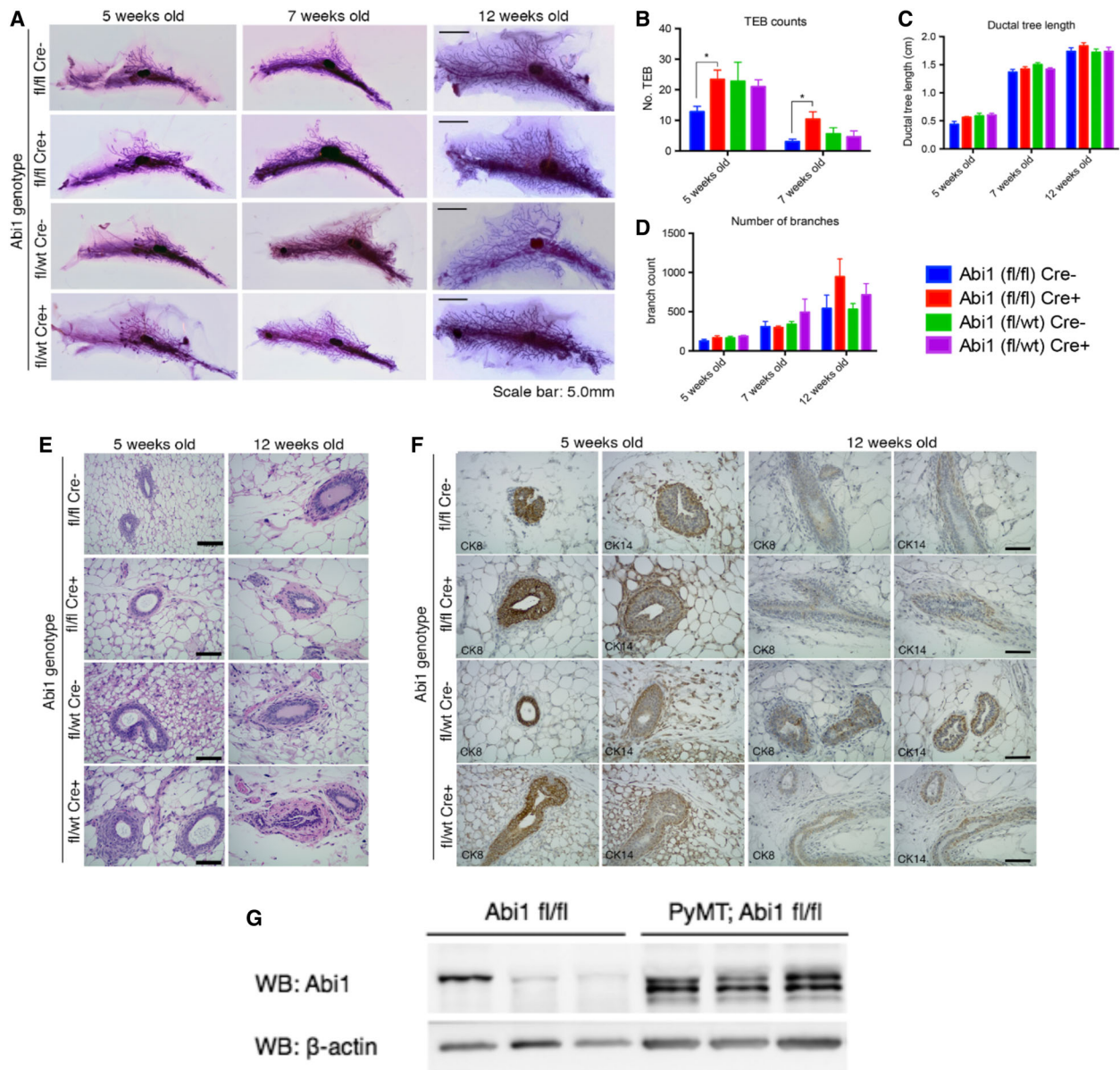


Fig. 3. *Abi1* loss does not impact the long-term development of healthy mouse mammary glands. (A) Whole-mount analysis of the inguinal mammary gland stained with Carmine Alum reveals no gross changes in gland anatomy at 5, 7, or 12 weeks of age after CRE-mediated deletion of *Abi1*. Morphometry of whole mounts reveals a significant increase in the number of terminal end buds in homozygous *Abi1* null glands (B); however, this does not affect the elongation of the ductal tree (C) or the number of ductal branches (D). Scale bar, 5.0 mm. (E) Histological staining of mammary gland sections reveals no changes in tissue organization after CRE-mediated loss of *Abi1*. Scale bar, 100 μ m. (F) Immunostaining of mammary sections using markers for luminal epithelial cells (CK8) and myoepithelial cells (CK14) reveals sustained organization of the ductal epithelium in both control and *Abi1* null mammary glands. Scale bar, 50 μ m. Error bars indicate SEM. (* indicates $P < 0.05$, Student's t -test; $n = 5$ animals/genotype). (G) WB analysis indicates enhanced expression of *Abi1* in mammary epithelium of *Abi1*(fl/fl) PyMT mice vs. *Abi1* floxed mice *Abi1* (fl/fl). Each lane represents one mammary gland (*Abi1* fl/fl) or tumor [PyMT: *Abi1* (fl/fl)], ($n = 3$ mice).

KO mice showed that the genes of basal-like cells (*Krt14*, *Vim*) are responded to *Abi1* depletion, however luminal cell type genes markers (*Krt8*, *Krt18*, *Sox9*, *Esr1*) do not (Table S7). The directionality of gene

expression of *Krt14* and *Vim* in heterozygous and homozygous mice was different.

We have shown that *Abi1* KO mouse embryonic fibroblasts reliably show downregulation of WAVE2

Table 1. Gene expression variability upon *Abi1* depletion in primary PyMT breast cancer tumors defined by RNA-seq.

Mouse ID Genotype		G144	G164	G184	G174	Fold change depletion			Treatment Effect ^a
		fl/wt	fl/wt	fl/fl	fl/fl	Heterozygous	Homozygous	Ratio	
Cre		-	+	-	+				
Gene expression (RNA-seq from primary tumors)	<i>Abi1</i>	14389	7343	15682	1018	1.96	15.40	7.86	Yes
	<i>Abi2</i>	8173	7189	8031	9719	1.14	0.83	0.73	No
	<i>Abi3</i>	146	271	343	195	0.54	1.76	3.26	Yes
	<i>Nckap1</i>	36435	34964	40591	38802	1.04	1.05	1.00	No
	<i>Wasf1</i>	88	66	85	118	1.33	0.72	0.54	No
	<i>Wasf2</i>	8563	10349	13894	11087	0.83	1.25	1.51	No
	<i>Wasf3</i>	261	151	95	221	1.73	0.43	0.25	Yes
	<i>Brk1</i>	10240	9103	11091	9920	1.12	1.12	0.99	No
	<i>Cytip1</i>	23448	22528	31708	23030	1.04	1.38	1.32	No
	<i>Cytip2</i>	521	1289	1633	580	0.40	2.82	6.97	Yes
	<i>Rac1</i>	25732	22621	27209	24431	1.14	1.11	0.98	No
	<i>Ndel1</i>	10923	10842	15082	13030	1.01	1.16	1.15	No

^aTreatment effect is 'positive' (yes) if the fold change of gene expression for heterozygous and homozygous mice changed more than 1.5 times (bold text) in any direction and 'negative' (no) in other cases. RNA-seq. expression profiles of WAVE complex, and *Rac1* and *Ndel1* genes (involved in WAVE complex stability and functionality) show the differences between heterozygous vs. homozygous *Abi1* KO PyMT mammary tumors.

[16]. Consistent with this finding, western blot analysis of *Abi1* KO breast tumors (tumor lysates from 3 mice/genotype) showed an appreciable reduction in WAVE2 expression in the absence of ABI1, recapitulating previously observed WAVE complex dynamics and dependence of complex stability on *Abi1* gene status (Fig. 4A) [16,17]. Interestingly, WAVE2 expression remains relatively stable in heterozygous *Abi1* KO tumors, suggesting that a single copy of *Abi1* is enough to sustain WAVE complex stability to some degree, noting that there is still a noticeable loss in WAVE2 expression. Also, densitometric analysis of our western blots revealed significant upregulation of ABI2, another member of the ABI family, only in homozygous *Abi1* KO animals, in agreement with our previous findings (Fig. 4B) [16].

Based on our western blot findings, we next examined whether altered WAVE complex expression in the absence of ABI1 was recapitulated by immunohistochemical staining of tumor tissue (Fig. 4D). Similar to our western blot results, *Abi1*-null tissue shows increased ABI2 expression in the cytosol, while WAVE2 shows moderate downregulation overall. WAVE1 is modestly expressed regardless of ABI1 status; therefore, it may not play a role in breast tumorigenesis in this model. Due to their ubiquitous expression, ABI1-WAVE2 complexes are considered canonical WAVE complexes that drive F-actin polymerization during cell processes [59]. As there is a concomitant loss of WAVE2 upon *Abi1* KO but sustained

tumor growth in PyMT mammary tumors, it is possible that other factors contribute to ARP2/3-mediated actin polymerization. Moreover, overall primary mammary tumor histopathology was not affected upon ABI1 loss (Fig. 4C; Table S8). While most of the primary tumors in either control or homozygous *Abi1* knockout animals remain in grades 3 or 4, some tumors in the heterozygous *Abi1* knockout appear to be in grade 2, further highlighting the impact of single copy *Abi1* deletion as opposed to homozygous deletion and suggesting other mechanisms may be induced in the complete genetic absence of *Abi1*.

3.6. ABI1 gene dose regulates primary PyMT tumors growth kinetics

To determine the impact of *Abi1* disruption on mammary tumor initiation and progression, we used our *Abi1* KO mouse model to study the impact of *Abi1* loss on tumor progression and characteristics in the PyMT-driven breast cancer. The PyMT model initiates spontaneous tumor formation with most mammary glands developing tumor nodes. Interestingly, KO mice *Abi1* do not significantly impact primary tumor latency (Fig. 5A). To determine the effects of *Abi1* expression on breast cancer progression, heterozygous and homozygous KO mice were used to study the growth kinetics (i.e., tumor volume changes over time) of sporadically occurring tumors. Tumor size was measured biweekly, starting from first day of tumor palpation.

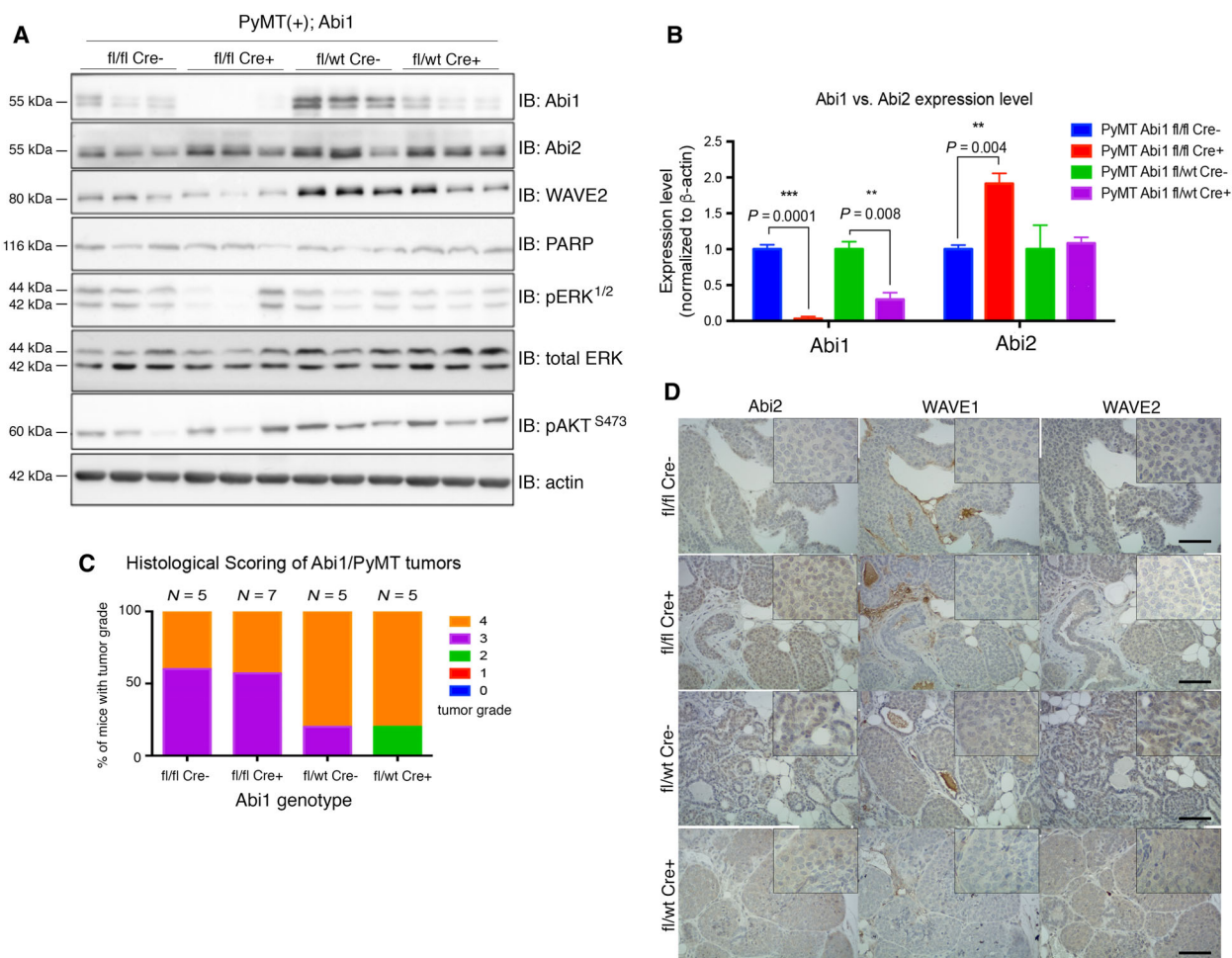


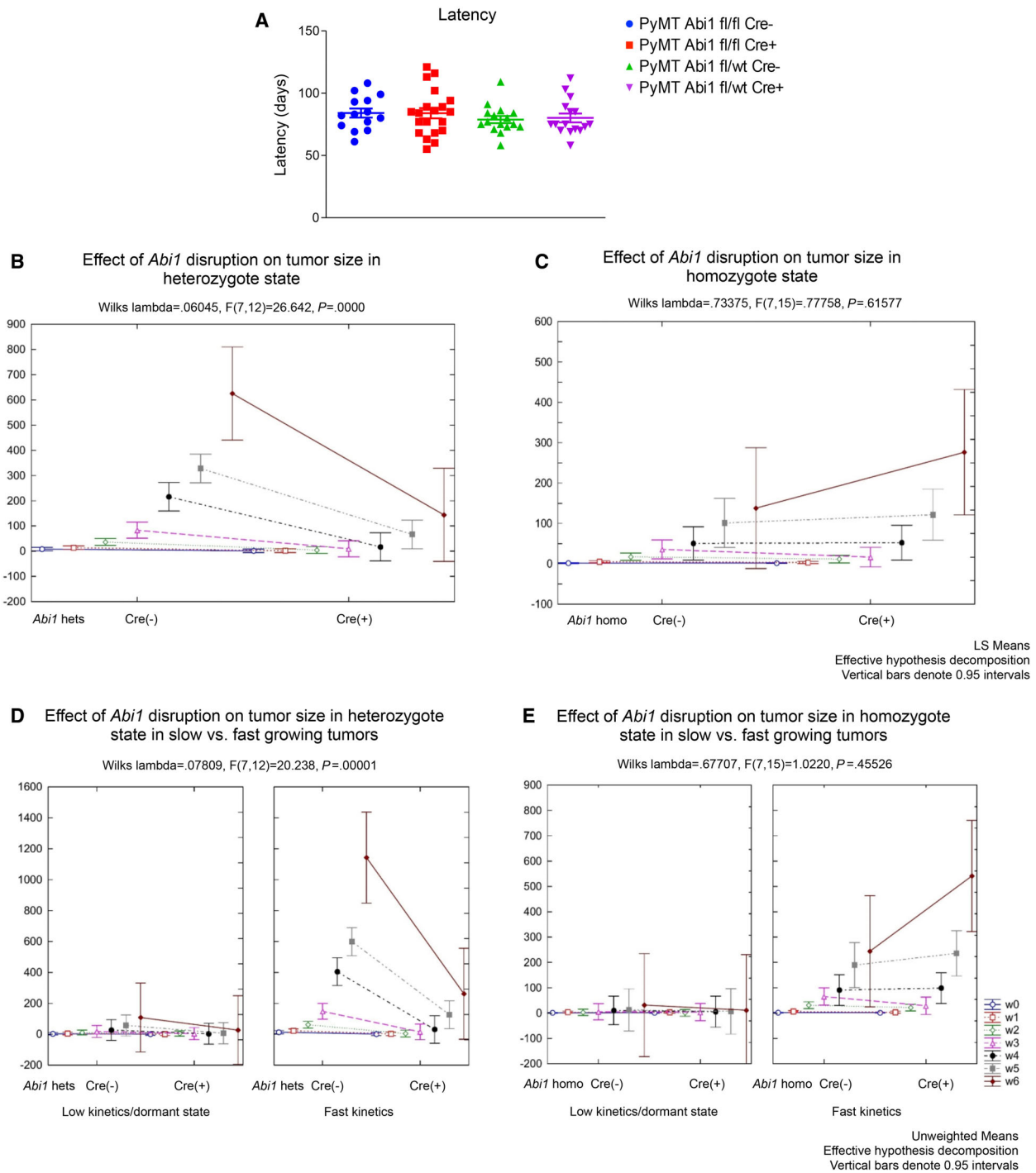
Fig. 4. *Abi1* KO severely impacts WAVE complex gene expression dynamics. (A) Western blot analysis of primary mammary tumors from *Abi1* KO PyMT mice shows significant depletion of ABI1 protein, but only in the homozygote *Abi1* null is there significant upregulation of ABI2 protein as indicated by densitometry (B). Each lane represents one mammary tumor isolated from one animal of that genotype. Error bars indicate SEM. ($P < 0.05$, *t*-test; $n = 3$ animals/genotype). (C) Analysis of primary tumor histology reveals no significant changes in tumor grade between controls and *Abi1* knockouts, ($P > 0.05$, *t*-test; $n \geq 5$ mice per genotype of age between 20 and 22 weeks were used for analysis, Table S8). Error bars indicate SEM). (D) Immunostaining with antibodies against WAVE complex proteins supports our findings that ABI2 is upregulated only in ABI1 null breast tumors. WAVE1 retained its low expression, while WAVE2 was concomitantly depleted with ABI1, in agreement with WB data, above. 20x magnification; inset, 40x magnification, Scale bar, 50 μ m.

We collected and analyzed datasets from *Abi1* homozygous Cre(+) ($n = 11$), *Abi1* heterozygous Cre(+) ($n = 11$), *Abi1* homozygous Cre(-) ($n = 13$), and *Abi1* heterozygous Cre(-) ($n = 13$) samples (Fig. 5B–E). The tumor kinetics showed two growth patterns. The analysis of tumor volume kinetic data in mice identified two tumor growth patterns, which are growing with very low or fast rates across all four genotypes (Fig. 5B–E; Tables S9). The fraction of tumors exhibiting the slower growth varied between 54% and 64% across the four experimental groups. Other breast tumor samples showed stable exponential growth with either moderate or high growth rates (Fig. 5B–E;

Table S9). Using the one-way ANOVA test, we found that in *Abi1* heterozygous KO model samples the tumor growth kinetics was strongly suppressed vs. control (Fig. 5D), while no significant effect was found in *Abi1* homozygous KO model tumor samples with some positive trend in the opposite direction in faster-growing tumors (Fig. 5E).

3.7. ABI1 promotes the number and size of lung metastases in a gene dose-dependent manner

Most *Abi1* KO PyMT mice demonstrated pulmonary metastasis within 6 months of the primary tumor



detection. We noted that mice with fast-growing tumors showed a positive trend for association with multiple metastatic events and large size metastatic foci in both Cre(-) control groups (Fig. 6). To elucidate the role of *Abi1* gene dosage effect in lung metastasis, we first analyzed the tumor kinetic rates of the

primary tumor growth vs. the largest tumor metastatic foci at 6 months within the same mice in *Abi1* KO homozygous and heterozygous tumor groups (Fig. 6A–B). Fig. 5A–B shows a weak gene dose effect in both homo- and heterozygous primary tumor kinetics. To be more conclusive, we estimated parameters

Fig. 5. Primary tumor growth kinetics analysis indicates *Abi1* gene dose effect in heterozygous mice. (A) Primary tumor latency in PyMT animals is not significantly affected upon *Abi1* KO. The X-axis of a panel (a) represents latency time comparison of the tumors in four treatment conditions defined on the upright corner of the panel (*Abi1* fl/fl Cre⁻, *n* = 14 mice; *Abi1* fl/fl Cre⁺, *n* = 20 mice, *Abi1* fl/wt Cre⁻, *n* = 16; *Abi1* fl/wt Cre⁺, *n* = 16 mice). (B-E) Treatment effects of *Abi1* disruption (fw Cre⁺) vs fw Cre⁻) and tumor kinetics of tumor size in heterozygous or homozygous mice. Graphical tools of Statistica-13 were used. Each plot on panels (B-E) shows tumor size at seven-time points (w0, w1, w3, w3, w4, w5, and w6 (see Methods)) (for *Abi1* fl/fl Cre⁺, or Cre⁻, *n* = 13 mice were used; for *Abi1* fl/wt Cre⁺, or Cre⁻, *n* = 11 mice were used). The line connects start (Cre⁻) with the endpoint (Cre⁺) tumor size datasets allowing the comparison of tumor kinetic observations to be easily followed; mean values of tumor size are linked by direct lines at the same detection time point. Wilks lambda statistics and Fisher test were used for estimation of treatment significance. Panels (B) and (C) represent a visualization of the treatment effect (Cre⁻) v.s. Cre⁺) of *Abi1* on tumor size in observed time points. Vertical bars indicate 0.95 intervals, CI. An effective decomposition method of Statistica-13 was used. The primary tumor size comparison in fastly growing mouse groups shows the exponential growth kinetics. (Methods, Table S10). To compare gene dosage effects within heterozygote and homozygote groups, mean values in 7 observed time points were compared (see Table S10 for details). Our results showed that in the cases of fast kinetics datasets, differences between the paired sample mean values were not significant for homozygote (*t*-test, *P* > 0.15) but significant for heterozygote state (*t*-test, *P* = 0.017). (See Methods and Table S10).

of the tumor volume kinetics using the exponential fit function $f(t; a, b) = (a - b) * \exp(ax)$, where parameter *a* is the rate of cell volume growth and (*a* - *b*) is the initial tumor volume.

No statistical differences between exponent rates in control and treatment were found (Materials and Methods). However, a comparison of the number and size of metastatic foci of *Abi1* KO animals indicated a strong gene dose effect (Fig. 6C–G). Fig. 6C–D shows the frequency distribution of pulmonary metastatic foci that exhibited the highest number of metastatic foci and largest metastasis size in the *Abi1* fl/fl and fl/wt lung tissues. In each lung sample, the frequency distribution of the metastatic foci size shows the skewed form with long tails. We found that for each case, the frequency distribution of pulmonary metastatic foci size is fitted well by a discrete analog of shifted log-normal distribution function (for better visualization the function approximated by continuous curves) (Table S9A–B and Methods). Estimated parameters of the distribution function we used to define significant differences between the shapes of the distribution functions shown in Fig. 6C–D (Table S9B). In particular, parameter x_0 estimates a mode of the frequency distribution function which is most frequent size of micro-metastasis foci. For *Abi1* fl/fl Cre⁻, fl/wt Cre⁻, fl/wt Cre⁺ data are varied between 6.3–8.6 μm^2 , but for fl/fl Cre⁺ focus size equals 1.3 μm^2 (Table S9B). A comparison of x_0 and the parameter *b* (basal (smallest) foci size at $x = 0$), Table S9B) of the best-fit distribution function draw in Fig. 6C,D suggests a significant reduction of the multiple metastatic foci size and their numbers in the treatment cases fl/wt Cre⁺ and fl/fl Cre⁺. Additionally, statistical testing using the Wilcoxon signed-rank method demonstrated significant differences between the observed frequency distributions of treatment v.s. control datasets (*P* < 0.0001).

Comparison of the frequency distributions of the treatment groups provided a significant difference (reduction of median value in fl/fl Cre⁺ vs median value in fl/wt Cre⁺) (*P* < 0.0001). These results indicate a strong *Abi1* gene dose effect promoting lung metastases in both homozygous and heterozygous PyMT models but the effect in homozygous mice was stronger. Similar results were observed for pulmonary metastasis foci size bins (50 μm^2) frequency distribution that includes all defined pulmonary metastatic foci datasets (Fig. 6E–F). Representative lung tumor images are shown in Fig. 6G.

4. Discussion

Here, for the first time, we demonstrate the metastasis driver role of *ABI1* in breast cancer tumor progression using the PyMT mouse model and clinical data from breast cancer patients. Our bioinformatics analyses revealed the significant role of human *ABI1* and a subset of the WAVE complex genes in the context of breast cancer progression and metastatic process.

In the Metadata and Rosetta cohorts, the high expression of *ABI1* demonstrated poor survival time patterns as indicated by survival time and is significantly associated with metastatic events. Moreover, in the large METABRIC cohort the *ABI1* expression is positively correlated with DNA CNA, histologic grade 3, and basal-like phenotype, but negatively correlated with ER status and does not correlate with LN status. We identified the high confidence and reproducible multigene survival prognosis signature comprised of *ABI1* and six other genes: *BRK1*, *CYFIP1*, *CYFIP2*, and *WASF3*, which are the genes encoding WAVE complex members; and *RAC1* and *NDEL1* genes, which are upstream interactors and regulators of the WAVE complex [5,14,60]. Both *RAC1* and *NUDEL*

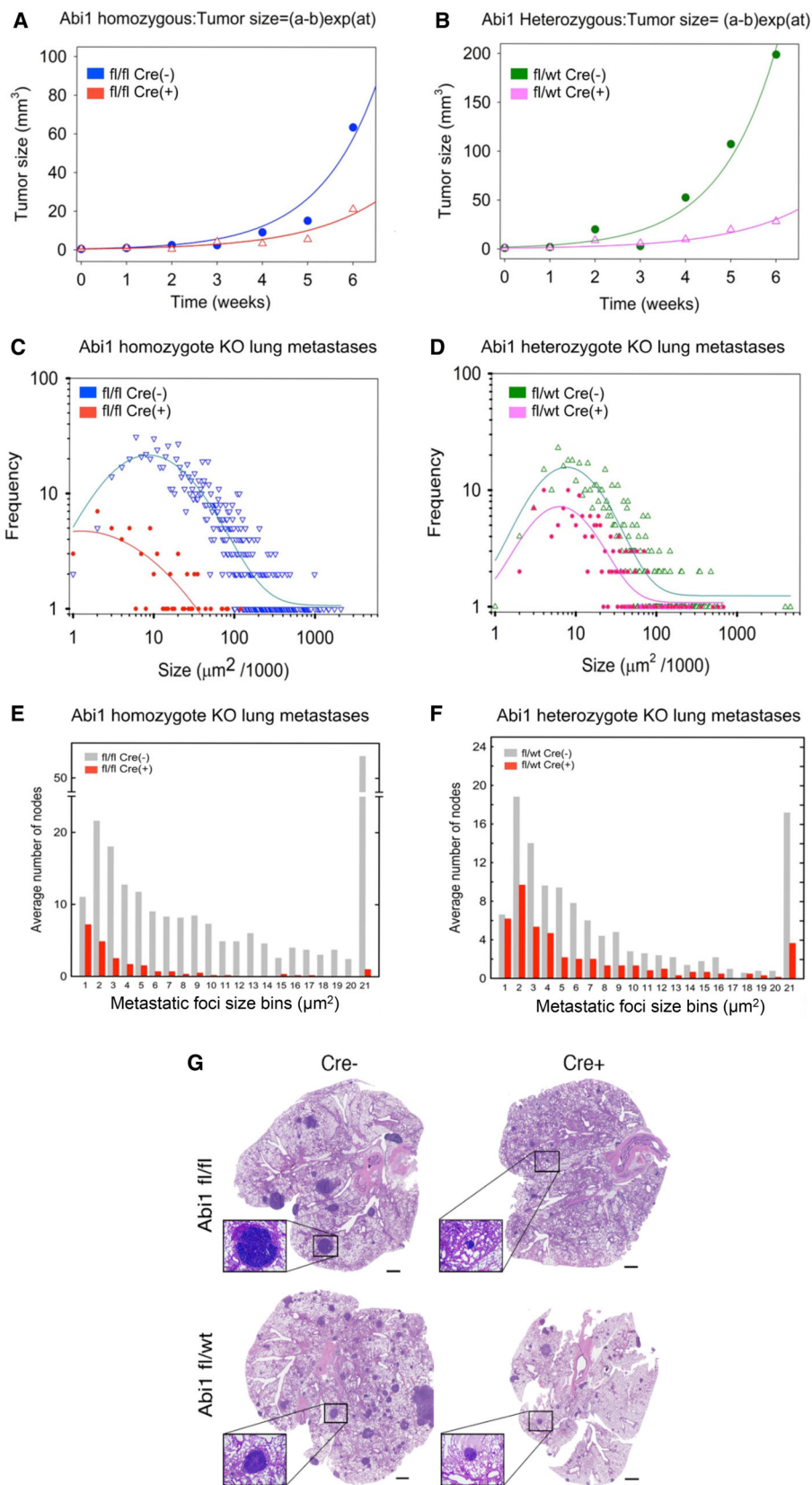


Fig. 6. *Abi1* gene knockout reduces metastatic burden in heterozygous and homozygous mice. Representative tumor kinetics of primary (panels a-b) vs metastatic tumors (panels c-d). Panel (A) Comparison of the primary tumor volume kinetics in *Abi1* homozygous KO mouse (fl/fl; Cre+) (G209, data: red triangle; best-fit function: red line) and the control *Abi1* (fl/fl Cre-) mouse (G184, data: blue circle; best-fit function: blue line). (B) Comparison of the primary tumor kinetics of *Abi1* KO heterozygous (fl/wt Cre+) mouse (G251, data: pink triangle; best-fit function: pink line) and the *Abi1* control (fl/wt; Cre-) mouse (G202, data: green circle; best-fit function: green line). Kinetics of mean values (A and B) were fitted by exponential curve $f(t, a, b) = (a - b) * \exp(at)$, where t is time, constant a is the rate of cell population growth and constant $(a - b)$ is the initial tumor population size. Each kinetic dataset includes seven time points (see also Table S10). The estimated parameters in *Abi1* fl/fl Cre (-) tumors: $a = 0.77 \pm 0.159$, t -test, $P = 0.0047$, $b = 0.2 \pm 0.678$, t -test, $P > 0.1$ and in *Abi1* fl/fl Cre (+) $a = 0.60 \pm 0.153$, t -test, $P = 0.0039$, $b = 0.1 \pm 0.586$, t -test, $P > 0.1$. Estimated parameters in *Abi1* fl/wt Cre (-) tumors: $a = 0.79 \pm 0.091$, t -test, $P = 0.001$, $b = -1.00 \pm 0.964$, t -test, $P > 0.1$, and in *Abi1* fl/wt Cre (+) $a = 0.589 \pm 0.110$, t -test, $P = 0.0031$, $b = -0.30 \pm 0.66$, t -test, $P > 0.1$. According to these results, differences between mean values of the tumor sizes in the studied groups in time are not significant. While primary tumor volume kinetics was not significantly different in these mice vs. their corresponding controls, (A, homozygous *ABI1* KO vs. control) and (B, heterozygous KO vs. control), the difference in metastatic tumor burden of the same mice within each mouse genotype was significant (C) and (D). Panels (C) and (D) show the frequency distributions of a lung metastatic foci size in the heterozygous and homozygous mice, which primary tumors kinetics showed on panels (A) and (B), respectively. Each Y-axis value shown in the histograms (C-D) represents a count of metastatic foci within a metastatic size normalized interval (a bin). The bin was defined by rounding the metastatic size divided by 1000 to the nearest integer, and the number of metastatic foci in each bin was counted. Based on our findings, the metastases size frequency distribution in the lung has skewed form with the long right tail. To provide a visualization of such frequency distribution, we used $\log_{10} - \log_{10}$ plot. We used the same color for dots of the empirical distributions and the fitting function lines, as was indicated in the figures. Such empirical frequency distribution was modeled and parameterized using the shifted log-normal distribution function:

$$f(x; y_0, x_0, a, b) = y_0 + a * \exp(-0.5 * (\ln(x/x_0)/b)^2),$$

where x is the node size and y_0, x_0, a, b unknown parameters. We estimated the parameters using the nonlinear curve fitting option of SigmaPlot-13 software. Datasets and detailed results of the parameterization of this function are presented in Table S9. (E-F) show histogram bar plots for the distribution of the average number of metastases foci size in the lungs of *Abi1* KO mice in comparison to their genetic controls. X-axis indicates binning for every 5000 μm^2 metastasis colony area size, with bin 1 representing 0–5000 μm^2 and bin 21 representing 100001 μm^2 and larger; Y-axis: count of the samples within given binning interval (\pm SEM). The size stratification of individual metastatic colonies shows that mice lacking *ABI1* still have relatively small metastatic colonies but they grow slowly or/and stay at dormant state and appear unable to establish macrometastases when compared to our controls ($P < 0.001$; Wilcoxon signed-rank test). Lung metastasis quantification was performed following fixation, paraffin embedding and sectioning: three 5 μm sections (sectioned every 50 μm) were collected from each mouse (*Abi1* fl/fl, Cre-, $n = 7$; *Abi1* fl/fl, Cre+, $n = 6$; *Abi1* fl/wt, Cre-, $n = 6$; *Abi1* fl/wt, Cre+, $n = 6$; animals per genotype, age 18–22 weeks), were stained with hematoxylin and eosin, and imaged using Omnyx digital pathology scanner (GE Healthcare). Images were quantified using ImageJ software (NIH). Results of panels (E) and (F) support the results presented in (C) and (D). (G) Histological staining of representative lung sections reveals severely diminished metastasis upon deletion of the *Abi1* gene. Scale bar, 1 mm. Inset, 4x magnification.

participate in the EMT pathway and play key roles in the metastatic migration of epithelial cells via the interaction with WAVE family proteins and the regulation of cancer-determined pathways [5,12,61–63].

Collectively, our tumor progression and metastatic prognostic signatures allow for the identification of optimal gene expression cutoff values to stratify patients on low-, moderate-, and high-risk subgroups based on DFS and DMFS times. Our survival prediction analyses establish the significance of *ABI1* gene expression as a pro-oncogenic factor of primary tumor formation and metastasis in breast cancer patients. These findings support the experimentally testable working hypothesis that genetic mechanisms of *ABI1* are key components in the metastatic breast cancer process.

Univariate and multivariate analyses and comparisons between Kaplan–Meier survival curves generated

with our prognostic signature and those generated with either estrogen receptor (ESR) or lymph node status reveal that our signature outperforms these clinically used variables and could lead to better personalized and predictable treatment selection. This conclusion is supported by our co-expression analysis between *ABI1* and other members of the *ABI1* survival (prediction) signature and the observed significant positive correlation between *ABI1* expression, CNA, histologic grades, and basal-like phenotype vs. ER(+) luminal cancer phenotype—the clinical markers of aggressiveness, metastasis, and drug resistance frequency.

The availability of a genetically engineered conditional *Abi1* KO mouse permitted us to investigate the role of *Abi1* downstream from the PyMT oncogene. By comparing the effects of one- and two-allele inactivation of the *Abi1* gene, we were able to determine that *ABI1* expression levels play an important pro-

oncogenic role in breast cancer tumor progression and metastatic disease. The two-allele inactivation of the gene (in *Abil* homozygote KO mice) and one-allele inactivation of *Abil* (in *Abil* heterozygote KO mice) led to lower metastatic burden in the lungs.

Disruption of *Abil* in normal mammary epithelium led to a significant increase in terminal end buds at weeks 5 and 7 (Fig. 3B), but beyond that time point, the development of mammary glands was not affected (Fig. 3B–D). The increase in the TEB number, as well as the trend toward increased branching in tissue with *Abil*-disruption, warrants further investigation to determine whether ABI1 or other ABI proteins play a role in normal murine mammary gland development. To corroborate the findings of *Abil*, the disruption of *Waf3* gene also demonstrated no significant effect on mammary gland development [22]. *WASF3* is part of *AB11* 7-gene signature.

We observed that complete loss of ABI1 yields no difference in primary mammary tumor growth kinetics (Fig. 5C,E) and that lung metastasis is severely abrogated in both homozygous and heterozygous *Abil* KO (Fig. 6C–F). Thus, our findings strongly suggest that ABI1 is critical for pulmonary metastasis of aggressive breast tumors due to its essential role in sustaining WAVE complex dynamics. The WAVE complex is assembled from intimate interactions of five obligatory components: a WAVE, an ABI, a CYFIP, an NAP, and BRK protein, which are altogether products of 11 genes [6,8,9]. The study by Kirschner's group demonstrated that the presence of all five WAVE complex proteins is required to form the functional WAVE complex *in vitro* [6]. Genetic inactivation of *Abil* led to overall WAVE complex downregulation in MEF cells, but deregulation of individual WAVE complex proteins was also evident. These included the relative upregulation of ABI2. Similarly, upregulation of ABI2 is observed in breast tissue lacking ABI1 (Fig. 4A,B). Despite their homology and similarities in function, upregulation of ABI2 cannot sustain pulmonary metastases in homozygous *Abil* KO animals (Fig. 6C, E,G), strongly indicating that ABI1 is critical for lung metastases in this model.

The lack of local effect on primary tumor growth in *AB11* homozygous mice is difficult to explain in the context of the effect on lung metastases but raises the possibility for potential tumor suppressor role for *AB11* in breast epithelial cells in some genetic contexts such as here downstream from the PyMT oncogene. ABI1 acts as tumor suppressor in several other tissues such as prostate [30].

Focus is a pathologic term describing cells that can grow as a colony and be seen only microscopically. In

this study, we quantified differences in the number of multiple metastatic foci and the sizes of the breast cancer metastases. We found essential differences for both characteristics in the breast cancer metastases in the *AB11* gene dosage-dependent manner. Our experimental model results demonstrate the important role of *AB11* gene dosage and expression in the lung metastasis process which may model metastatic potential of CNA and gene expression of *AB11* in patient's primary breast tumors (Fig. 1A), consistent with histologic high-aggressive breast cancers (Fig. 1B), and basal-like subtype (Fig. 1D)—hallmarks of high aggressive invasive breast cancer with polyclonal metastases potential. Also, our experimental findings consist of high ABI1 protein expression in human invasive breast carcinoma associated with high risks of tumor recurrence and overall survival (Fig. 2, Figs. S1 and S3, [32]).

It was observed that protein interaction combinations of *WASF3* with some members of WAVE complex and *RAC1* are responsible for breast cancer aggressiveness and metastasis [22]. In our study, we found an association of *WASF3* and some other WAVE complex components (that are part of the prognostic signature) with invasive breast cancer that molecular pattern is associated with aggressive (basal-like) breast cancer subtype. Interestingly, heterogeneity and instability of Wave complexes without *Abil* protein could contribute to the heterogeneity in latency, the size and number of lung metastatic lesions as observed in *Waf3* KO mice [22].

Our data adhere to previously published findings regarding the impact of ABI1 protein in driving aggressive mammary oncogenesis in mouse xenograft models of breast cancer [17,34]. ABI1 has been cited in several cancer types, such as ovarian cancer [29,64], hepatocellular carcinoma [65], and colorectal carcinoma [66]. Notably, all studies to date examined the role of ABI1 in breast cancer using cancer cell lines. This is the first genetic study examining the role of *Abil* *in vivo* using the mouse model of aggressive breast cancer. The critical role of *Abil* in the lung metastasis in the mouse not only provides preclinical evidence for the role of *Abil* in metastatic progression but also supports *AB11*-based 7-gene prognostic signature as both a prognostic marker and a prospective therapeutic target.

Univariate and multivariate analyses and comparison of Kaplan–Meier survival curves generated with our *AB11* gene expression signature to those generated with either estrogen receptor (ESR) or lymph node status reveal that our gene signature is indeed a more robust prognostic predictor than other clinically used variables and could lead to better treatment selection.

5. Conclusion

Our findings indicate the significant predictive value of the *ABI1*-based 7-gene prognostic signature derived from primary tumors in the metastatic risk of breast cancer patients. Moreover, targeting *ABI1* may provide a beneficial therapeutic effect in preventing metastases.

Acknowledgments

We are grateful to Patricia Numann, Beth Baldwin, Vince, and Judy Smith for the support of the study. We thank Jessica Ouderkirk-Pecone for assistance with tumor measurements and animal dissections. We thank Claudia Mondragon for the validation of quantifications included in this publication. We thank Dr. Jeffrey Ross for his careful reading of the manuscript; Disharee Das, Heidi Hehnly, and Patricia Kane and past and current members of the Drs Kotula and Kuznetsov laboratories for the insightful discussions. This work was supported in part by grants from the Carol Baldwin Foundation of CNY to LK and MK, Upstate Cancer Center Pilot Grant (Connolly Fund) and the National Cancer Institute (R01CA161018) to LK, and Peter T. Rowley Breast Cancer Projects program of the NY State Dept. of Health to MK. VAK study was supported by the New York Empire Innovative Program grant. We acknowledge the gift from Dawn K. (Smith) Steber Endowment for Cancer Research at the Upstate Foundation.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

LK designed and interpreted the experimental results of the study. VAK designed bioinformatic and statistical analyses and interpreted the results of this study; VAK and LK wrote the final version of the manuscript. AG, CP, and VAK performed the analyses. AR performed all animal experiments and cowrote the paper with LK and VAK. BAP performed qPCR analysis. TC completed the pathological assessment of tumors. IB analyzed TCGA breast tumor mutation data sets. MK helped with experimental design in mice. VAK, GB, and AS contributed to the interpretation and discussion of the clinical significance of the data presented in this paper.

Consent for publication

All authors have agreed to publish this manuscript.

Peer Review

The peer review history for this article is available at <https://publons.com/publon/10.1002/1878-0261.13175>.

Data accessibility

All data generated or analyzed during this study are included in this manuscript.

References

- 1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020;**70**:7–30. <https://doi.org/10.3322/caac.21590>.
- 2 Redig AJ, McAllister SS. Breast cancer as a systemic disease: a view of metastasis. *J Intern Med*. 2013;**274**:113–26. <https://doi.org/10.1111/joim.12084>.
- 3 Spence HJ, Timpson P, Tang HR, Insall RH, Machesky LM. Scar/WAVE3 contributes to motility and plasticity of lamellipodial dynamics but not invasion in three dimensions. *Biochem J*. 2012;**448**:35–42. <https://doi.org/10.1042/BJ20112206>.
- 4 Yokotsuka M, Iwaya K, Saito T, Pandiella A, Tsuboi R, Kohno N, et al. Overexpression of HER2 signaling to WAVE2-Arp2/3 complex activates MMP-independent migration in breast cancer. *Breast Cancer Res Treat*. 2011;**126**:311–8. <https://doi.org/10.1007/s10549-010-0896-x>.
- 5 Chen Z, Borek D, Padrick SB, Gomez TS, Metlagel Z, Ismail AM, et al. Structure and control of the actin regulatory WAVE complex. *Nature*. 2010;**468**:533–8. <https://doi.org/10.1038/nature09623>.
- 6 Gautreau A, Ho HY, Li J, Steen H, Gygi SP, Kirschner MW. Purification and architecture of the ubiquitous Wave complex. *Proc Natl Acad Sci USA*. 2004;**101**:4379–83.
- 7 Molinie N, Gautreau A. The Arp2/3 regulatory system and its deregulation in cancer. *Physiol Rev*. 2018;**98**:215–38. <https://doi.org/10.1152/physrev.00006.2017>.
- 8 Miki H, Suetsugu S, Takenawa T. WAVE, a novel WASP-family protein involved in actin reorganization induced by Rac. *EMBO J*. 1998;**17**:6932–41. <https://doi.org/10.1093/emboj/17.23.6932>.
- 9 Suetsugu S, Miki H, Takenawa T. Identification of two human WAVE/SCAR homologues as general actin regulatory molecules which associate with the Arp2/3 complex. *Biochem Biophys Res Commun*. 1999;**260**:296–302.
- 10 Iwaya K, Norio K, Mukai K. Coexpression of Arp2 and WAVE2 predicts poor outcome in invasive breast carcinoma. *Mod Pathol*. 2007;**20**:339–43. <https://doi.org/10.1038/modpathol.3800741>.
- 11 Litschko C, Linkner J, Bruhmann S, Stradal TEB, Reinl T, Jansch L, et al. Differential functions of WAVE regulatory complex subunits in the regulation of

- actin-driven processes. *Eur J Cell Biol.* 2017;**96**:715–27. <https://doi.org/10.1016/j.ejcb.2017.08.003>.
- 12 Eden S, Rohatgi R, Podtelejnikov AV, Mann M, Kirschner MW. Mechanism of regulation of WAVE1-induced actin nucleation by Rac1 and Nck. *Nature.* 2002;**418**:790–3. <https://doi.org/10.1038/nature00859>.
 - 13 Lebensohn AM, Kirschner MW. Activation of the WAVE complex by coincident signals controls actin assembly. *Mol Cell.* 2009;**36**:512–24. <https://doi.org/10.1016/j.molcel.2009.10.024>.
 - 14 Wu S, Ma L, Wu Y, Zeng R, Zhu X. Nudel is crucial for the WAVE complex assembly *in vivo* by selectively promoting subcomplex stability and formation through direct interactions. *Cell Res.* 2012;**22**:1270–84. <https://doi.org/10.1038/cr.2012.47>.
 - 15 Dubielecka PM, Cui P, Xiong X, Hossain S, Heck S, Angelov L, et al. Differential regulation of macropinocytosis by Abi1/Hssh3bp1 isoforms. *PLoS One.* 2010;**5**:e10430. <https://doi.org/10.1371/journal.pone.0010430>.
 - 16 Dubielecka PM, Ladwein KI, Xiong X, Migeotte I, Chorzalska A, Anderson KV, et al. Essential role for Abi1 in embryonic survival and WAVE2 complex integrity. *Proc Natl Acad Sci USA.* 2011;**108**:7022–7. <https://doi.org/10.1073/pnas.1016811108>.
 - 17 Innocenti M, Zucconi A, Disanza A, Frittoli E, Areces LB, Steffen A, et al. Abi1 is essential for the formation and activation of a WAVE2 signalling complex. *Nat Cell Biol.* 2004;**6**:319–27. <https://doi.org/10.1038/ncb1105>.
 - 18 Roffers-Agarwal J, Xanthos JB, Miller JR. Regulation of actin cytoskeleton architecture by Eps8 and Abi1. *BMC Cell Biol.* 2005;**6**:36. <https://doi.org/10.1186/1471-2121-6-36>.
 - 19 Sweeney MO, Collins A, Padrick SB, Goode BL. A novel role for WAVE1 in controlling actin network growth rate and architecture. *Mol Biol Cell.* 2015;**26**:495–505. <https://doi.org/10.1091/mbc.E14-10-1477>.
 - 20 Tang Q, Schaks M, Koundinya N, Yang C, Pollard LW, Svitkina TM, et al. WAVE1 and WAVE2 have distinct and overlapping roles in controlling actin assembly at the leading edge. *Mol Biol Cell.* 2020;**31**:2168–78. <https://doi.org/10.1091/mbc.E19-12-0705>.
 - 21 Loveless R, Teng Y. Targeting WASF3 Signaling in Metastatic Cancer. *Int J Mol Sci.* 2021;**22**:836. <https://doi.org/10.3390/ijms22020836>.
 - 22 Qin H, Lu S, Thangaraju M, Cowell JK. Wasf3 deficiency reveals involvement in metastasis in a mouse model of breast cancer. *Am J Pathol.* 2019;**189**:2450–8. <https://doi.org/10.1016/j.ajpath.2019.08.012>.
 - 23 Teng Y, Ngoka L, Cowell JK. Promotion of invasion by mutant RAS is dependent on activation of the WASF3 metastasis promoter gene. *Genes Chromosomes Cancer.* 2017;**56**:493–500. <https://doi.org/10.1002/gcc.22453>.
 - 24 Teng Y, Qin H, Bahassan A, Bendzun NG, Kennedy EJ, Cowell JK. The WASF3-NCKAP1-CYFIP1 complex is essential for breast cancer metastasis. *Cancer Res.* 2016;**76**:5133–42. <https://doi.org/10.1158/0008-5472.Can-16-0562>.
 - 25 Molinie N, Rubtsova SN, Fokin A, Visweshwaran SP, Rocques N, Poleskaya A, et al. Cortical branched actin determines cell cycle progression. *Cell Res.* 2019;**29**:432–45. <https://doi.org/10.1038/s41422-019-0160-9>.
 - 26 Chorzalska A, Morgan J, Ahsan N, Treaba DO, Olszewski AJ, Petersen M, et al. Bone marrow-specific loss of ABI1 induces myeloproliferative neoplasm with features resembling human myelofibrosis. *Blood.* 2018;**132**:2053–66. <https://doi.org/10.1182/blood-2018-05-848408>.
 - 27 Kumar S, Lu B, Dixit U, Hossain S, Liu Y, Li J, et al. Reciprocal regulation of Abl kinase by Crk Y251 and Abi1 controls invasive phenotypes in glioblastoma. *Oncotarget.* 2015;**6**:37792–807. <https://doi.org/10.18632/oncotarget.6096>.
 - 28 Steinestel K, Bruderlein S, Steinestel J, Markl B, Schwerer MJ, Arndt A, et al. Expression of Abelson interactor 1 (Abi1) correlates with inflammation, KRAS mutation and adenomatous change during colonic carcinogenesis. *PLoS One.* 2012;**7**:e40671. <https://doi.org/10.1371/journal.pone.0040671>.
 - 29 Zhang J, Tang L, Chen Y, Duan Z, Xiao L, Li W, et al. Upregulation of Abelson interactor protein 1 predicts tumor progression and poor outcome in epithelial ovarian cancer. *Hum Pathol.* 2015;**46**:1331–40. <https://doi.org/10.1016/j.humpath.2015.05.015>.
 - 30 Nath D, Li X, Mondragon C, Post D, Chen M, White JR, et al. Abi1 loss drives prostate tumorigenesis through activation of EMT and non-canonical WNT signaling. *Cell Commun Sig.* 2019;**17**:120. <https://doi.org/10.1186/s12964-019-0410-y>.
 - 31 Xiong X, Chorzalska A, Dubielecka PM, White JR, Vedvyas Y, Hedvat CV, et al. Disruption of Abi1/Hssh3bp1 expression induces prostatic intraepithelial neoplasia in the conditional Abi1/Hssh3bp1 KO mice. *Oncogenesis.* 2012;**1**:e26. <https://doi.org/10.1038/oncsis.2012.28>.
 - 32 Wang C, Tran-Thanh D, Moreno JC, Cawthorn TR, Jacks LM, Wang DY, et al. Expression of Abl interactor 1 and its prognostic significance in breast cancer: a tissue-array-based investigation. *Breast Cancer Res Treat.* 2011;**129**:373–86. <https://doi.org/10.1007/s10549-010-1241-0>.
 - 33 Sun X, Li C, Zhuang C, Gilmore WC, Cobos E, Tao Y, et al. Abl interactor 1 regulates Src-Id1-matrix metalloproteinase 9 axis and is required for invadopodia formation, extracellular matrix

- degradation and tumor growth of human breast cancer cells. *Carcinogenesis*. 2009;**30**:2109–16. <https://doi.org/10.1093/carcin/bgp251>.
- 34 Wang C, Navab R, Iakovlev V, Leng Y, Zhang J, Tsao MS, et al. Abelson interactor protein-1 positively regulates breast cancer cell proliferation, migration, and invasion. *Mol Cancer Res*. 2007;**5**:1031–9. <https://doi.org/10.1158/1541-7786.MCR-06-0391>.
- 35 Fluck MM, Schaffhausen BS. Lessons in signaling and tumorigenesis from polyomavirus middle T antigen. *Microbiol Mol Biol Rev*. 2009;**73**:542–63. <https://doi.org/10.1128/MMBR.00009-09>.
- 36 Lin EY, Jones JG, Li P, Zhu L, Whitney KD, Muller WJ, et al. Progression to malignancy in the polyoma middle T oncoprotein mouse breast cancer model provides a reliable model for human diseases. *Am J Pathol*. 2003;**163**:2113–26. [https://doi.org/10.1016/S0002-9440\(10\)63568-7](https://doi.org/10.1016/S0002-9440(10)63568-7).
- 37 Chen L, Jenjaroenpun P, Pillai AM, Ivshina AV, Ow GS, Efthimios M, et al. Transposon insertional mutagenesis in mice identifies human breast cancer susceptibility genes and signatures for stratification. *Proc Natl Acad Sci USA*. 2017;**114**:E2215–24. <https://doi.org/10.1073/pnas.1701512114>.
- 38 Kulkarni S, Augoff K, Rivera L, McCue B, Khoury T, Groman A, et al. Increased expression levels of WAVE3 are associated with the progression and metastasis of triple negative breast cancer. *PLoS One*. 2012;**7**:e42895. <https://doi.org/10.1371/journal.pone.0042895>.
- 39 van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;**415**:530–6. <https://doi.org/10.1038/415530a>.
- 40 Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*. 2006;**66**:10292–301. <https://doi.org/10.1158/0008-5472.Can-05-4414>.
- 41 Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA*. 2005;**102**:13550–5. <https://doi.org/10.1073/pnas.0506230102>.
- 42 Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;**486**:346–52. <https://doi.org/10.1038/nature10983>.
- 43 de Carvalho LP, Tan SH, Ow G-S, Tang Z, Ching J, Kovalik J-P, et al. Plasma ceramides as prognostic biomarkers and their arterial and myocardial tissue correlates in acute myocardial infarction. *JACC: basic to translational*. *Science*. 2018;**3**:163–75. <https://doi.org/10.1016/j.jacbts.2017.12.005>.
- 44 Kuznetsov VA, Tang Z, Ivshina AV. Identification of common oncogenic and early developmental pathways in the ovarian carcinomas controlling by distinct prognostically significant microRNA subsets. *BMC Genom*. 2017;**18**:692. <https://doi.org/10.1186/s12864-017-4027-5>.
- 45 Motakis E, Ivshina AV, Kuznetsov VA. Data-driven approach to predict survival of cancer patients: estimation of microarray genes' prediction significance by Cox proportional hazard regression model. *IEEE Eng Med Biol Mag*. 2009;**28**:58–66. <https://doi.org/10.1109/memb.2009.932937>.
- 46 Enderlein G. Cox, D. R.; Oakes, D.: Analysis of Survival Data. Chapman and Hall, London – New York 1984, 201 S., £ 12,–. *Biom J*. 1987;**29**:114. <https://doi.org/10.1002/bimj.4710290119>.
- 47 Breslow N. Covariance analysis of censored survival data. *Biometrics*. 1974;**30**:89–99.
- 48 Euhus DM, Hudd C, LaRegina MC, Johnson FE. Tumor measurement in the nude mouse. *J Surg Oncol*. 1986;**31**:229–34. <https://doi.org/10.1002/jso.2930310402>.
- 49 Thorpe LM, Spangle JM, Ohlson CE, Cheng H, Roberts TM, Cantley LC, et al. PI3K-p110alpha mediates the oncogenic activity induced by loss of the novel tumor suppressor PI3K-p85alpha. *Proc Natl Acad Sci USA*. 2017;**114**:7095–100. <https://doi.org/10.1073/pnas.1704706114>.
- 50 Plante I, Stewart MK, Laird DW. Evaluation of mammary gland development and function in mouse models. *J vis Exp*. 2011;**53**:2828. <https://doi.org/10.3791/2828>.
- 51 Roarty K, Serra R. Wnt5a is required for proper mammary gland development and TGF-beta-mediated inhibition of ductal growth. *Development*. 2007;**134**:3929–39. <https://doi.org/10.1242/dev.008250>.
- 52 Wali VB, Gilmore-Hebert M, Mamillapalli R, Haskins JW, Kurppa KJ, Elenius K, et al. Overexpression of ERBB4 JM-a CYT-1 and CYT-2 isoforms in transgenic mice reveals isoform-specific roles in mammary gland development and carcinogenesis. *Breast Cancer Res*. 2014;**16**:501. <https://doi.org/10.1186/s13058-014-0501-z>.
- 53 Frugtniet B, Jiang WG, Martin TA. Role of the WASP and WAVE family proteins in breast cancer invasion and metastasis. *Breast Cancer*. 2015;**7**:99–109. <https://doi.org/10.2147/BCTT.S59006>.
- 54 Aswad L, Yenamandra SP, Siong Ow G, Grinchuk O, Ivshina AV, Kuznetsov VA. Genome and transcriptome delineation of two major oncogenic pathways governing invasive ductal breast cancer development. *Oncotarget*. 2015;**6**:36652–74.
- 55 Kuznetsov VA, Motakis E & Ivshina AV Low- and high- aggressive genetic breast cancer subtypes and significant survival gene signatures, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 4151–4156

- 56 Inman JL, Robertson C, Mott JD, Bissell MJ. Mammary gland development: cell fate specification, stem cells and the microenvironment. *Development*. 2015;**142**:1028–42. <https://doi.org/10.1242/dev.087643>.
- 57 Wagner K, McAllister K, Ward T, Davis B, Wiseman R, Hennighausen L. Spatial and temporal expression of the Cre gene under the control of the MMTV-LTR in different lines of transgenic mice. *Transgenic Res*. 2001;**10**:545–53.
- 58 Ouderkirk-Pecone JL, Goreczny GJ, Chase SE, Tatum AH, Turner CE, Krendel M. Myosin 1e promotes breast cancer malignancy by enhancing tumor cell proliferation and stimulating tumor cell differentiation. *Oncotarget*. 2016;**7**:46419–32. <https://doi.org/10.18632/oncotarget.10139>.
- 59 Kurisu S, Takenawa T. The WASP and WAVE family proteins. *Genome Biol*. 2009;**10**:226. <https://doi.org/10.1186/gb-2009-10-6-226>.
- 60 Niethammer M, Smith DS, Ayala R, Peng J, Ko J, Lee MS, et al. NUDEL is a novel Cdk5 substrate that associates with LIS1 and cytoplasmic dynein. *Neuron*. 2000;**28**:697–711. [https://doi.org/10.1016/s0896-6273\(00\)00147-1](https://doi.org/10.1016/s0896-6273(00)00147-1).
- 61 Fang D, Chen H, Zhu JY, Wang W, Teng Y, Ding HF, et al. Epithelial-mesenchymal transition of ovarian cancer cells is sustained by Rac1 through simultaneous activation of MEK1/2 and Src signaling pathways. *Oncogene*. 2017;**36**:1546–58. <https://doi.org/10.1038/ncr.2016.323>.
- 62 Kaneto N, Yokoyama S, Hayakawa Y, Kato S, Sakurai H, Saiki I. RAC1 inhibition as a therapeutic target for gefitinib-resistant non-small-cell lung cancer. *Cancer Sci*. 2014;**105**:788–94. <https://doi.org/10.1111/cas.12425>.
- 63 Zhang W, Xing L, Xu L, Jin X, Du Y, Feng X, et al. Nudel involvement in the high-glucose-induced epithelial-mesenchymal transition of tubular epithelial cells. *Am J Physiol Renal Physiol*. 2019;**316**:F186–94. <https://doi.org/10.1152/ajprenal.00218.2018>.
- 64 Yu X, Liang C, Zhang Y, Zhang W, Chen H. Inhibitory short peptides targeting EPS8/ABI1/SOS1 tri-complex suppress invasion and metastasis of ovarian cancer cells. *BMC Cancer*. 2019;**19**:878. <https://doi.org/10.1186/s12885-019-6087-1>.
- 65 Wang JL, Yan TT, Long C, Cai WW. Oncogenic function and prognostic significance of Abelson interactor 1 in hepatocellular carcinoma. *Int J Oncol*. 2017;**50**:1889–98. <https://doi.org/10.3892/ijo.2017.3920>.
- 66 Steinestel K, Bruderlein S, Lennerz JK, Steinestel J, Kraft K, Propper C, et al. Expression and Y435-phosphorylation of Abelson interactor 1 (Abi1) promotes tumour cell adhesion, extracellular matrix degradation and invasion by colorectal carcinoma cells. *Mol Cancer*. 2014;**13**:145. <https://doi.org/10.1186/1476-4598-13-145>.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig S1. Annotation discrepancies and cluster analysis.

Fig S2. Example of KS-weighted means batch effect correction and its effect on survival analysis.

Fig S3. Risk-predicting ability of individual members of the ABI1-WAVE signature in disease-free survival (DFS).

Fig S4. Risk-predicting ability of individual members of the ABI1-WAVE signature in distant metastasis-free survival (DMFS).

Fig S5. Commonly used clinical variables are insufficient for robust patient risk stratification.

Fig S6. Survival predictive analysis (RFS time) at transcription and protein level suggests a pro-oncogenic role of ABI1 in BC progression and outcome.

Fig S7. The implementation of 2D-DDg survival prediction to Rosetta data (DFS and DMFS).

Table S1. Gene list for prognostic signatures and associated probes/probsets represented by Rosetta (Merk) and Affymetrix U133-A & B microarrays.

Table S2. Results of 1D DDg survival prediction (DFS and DMFS) for Rosetta and Metadata sets.

Table S3. 1D-DDg by ABI1 expression level suggests significant metastasis events enrichment in a high-risk group (DMFS time).

Table S4. Survival significant prognostic genes encoding WAVE complex and *NDEL* gene are correlated with expression of ABI1. METABRIC breast cancer Dataset ($n = 1904$).

Table S5. Significance of association between the survival stratification grouping (by DMFS) and metastatic events (A, B) and our estimations of the probability of metastasis risk events (C,D).

Table S6. Abi1-based signature predicts distant metastatic events more accurately than commonly used clinical factors. Cox univariate and multivariate hazards proportional models analysis compares the ABI1-based 7-gene metastasis risk classifier (low, moderate, high risks by SVWg), ESR status (ER(+), ER(-)), and lymph node status (LN(+), LN(-)) to predict metastatic events in the Rosetta cohort.

Table S7. Basal-like vs. luminal cell type markers in primary breast tumors of Abi1 KO mice.

Table S8. Mouse breast pathology Ki67 and grading data.

Table S9. Data and statistical analysis of pulmonary node metastases size characteristics.

Table S10. Primary tumor size kinetic data.