Molecular Profiles of Matched Primary and Metastatic Tumor Samples Support an Evolutionary Model of Breast Cancer

Runpu Chen[§], Steve Goodison[‡*], Yijun Sun[¶§†*]

[§]Department of Computer Science and Engineering
[¶]Department of Microbiology and Immunology
[†]Department of Biostatistics
The State University of New York at Buffalo, Buffalo, NY 14203
[‡]Department of Health Sciences Research
Mayo Clinic, Jacksonville, FL 32224

Abstract

The interpretation of accumulating genomic data with respect to tumor evolution and cancer progression requires integrated models. We developed a computational approach that enables the construction of disease progression models using static sample data. Application to breast cancer data revealed a linear, branching evolutionary model with two distinct trajectories for malignant progression. Here, we use the progression model as a foundation to investigate the relationships between matched primary and metastasis breast tumor samples. Mapping paired data onto the model confirmed that molecular breast cancer subtypes can shift during progression, and supported directional tumor evolution through luminal breast cancer subtypes to increasingly malignant states. Cancer progression modeling through the analysis of available static samples represents a promising breakthrough. Further refinement of a breast cancer progression roadmap will provide a foundation for the development of improved cancer diagnostics, prognostics and targeted therapeutics.

Keywords— breast cancer, disease progression, gene expression, machine learning

[*]Please address all correspondence to Dr. Yijun Sun (yijunsun@buffalo.edu) and Dr. Steve Goodison (Goodison.Steven@mayo.edu).

1 Introduction

Through the accumulation of a series of genetic alterations, the progression of human cancer to malignancy can be viewed as a multistep evolutionary process driven by selective pressure imposed by the microenvironment [11, 34]. Tumor cells that accumulate changes in molecular pathways that control key cellular processes can come to dominate and drive the tumor towards malignancy, manifest in the dissemination of tumor cells and metastasis in secondary organs. The delineation of tumor evolution would provide invaluable insights into tumor biology and provide a foundation for the development of improved cancer diagnostics, prognostics and targeted therapeutics.

To derive models of dynamic progression it is optimal to obtain time-series data, but this is not typically possible for human cancer because of the need for timely surgical intervention upon diagnosis, and it is certainly not feasible on the large scale required for interpretation across the spectrum of disease. On the other hand, the wealth of excised tissue samples available from human cancer cases, coupled with the rapid development of molecular profiling techniques, has enabled the analysis of many thousands of tumors and the cataloging of cancer-associated molecular changes continues apace. In parallel, the computational handling of high-dimensional and multi-platform data is developing to assist with the interpretation and translation of the findings for clinical application.

Given the human tumor sampling limitation, we previously devised a computational strategy [28] to derive pseudo time-series data from static samples (excised tissue specimens). The design is based on the idea that each sample can provide a snapshot of the disease process, and if the number of samples is sufficiently large we can recover a visualization of disease progression. We demonstrated the utility of the developed pipeline - referred to as CancerMapp - by applying it to the analysis of gene expression data from over 3,100 breast tissue samples available from the TCGA and METABRIC consortiums [29, 6]. The analysis revealed defined data structures supporting a linear, branching model of breast cancer with two distinct trajectories for malignant progression [28]. The pattern of progression was confirmed by analysis in a series of smaller independent breast cancer datasets, and the conceptual validity of the model was validated by the mapping of additional clinical and molecular data available from the METABRIC and TCGA data sets. Although these factors were not included in the mathematical modeling, DNA mutation rate and genomic instability were clearly associated with progression towards malignancy, and disease-specific survival data aligned significantly with the paths to the most malignant model termini [28].

These findings showed that while cancer is a genetically and clinically heterogeneous disease, tumor samples are distributed on a low-dimensional data structure manifold, and suggest that disease subtypes are not hard-wired and can shift. As ever more tumor tissue profile data becomes available, we can continue to refine and improve the resolution of the progression model and investigate the conceptual validity with testable hypotheses. One hypothesis in this context would be that primary and metastatic tissue samples recovered from the same patient represent two time-points in a continuum of disease progression. The mapping of the molecular profiles from these samples on to the model would enable the visualization of the inter-relationship between paired samples, and may provide insight into the trajectory of disease progression. Here, we used gene expression data obtained from 123 matched primary and metastatic tumor samples [3] to investigate the progression relationships. The analyses supported the idea that molecular subtypes can shift within the same individual, and suggested that the shift is unidirectional along a continuum of disease state towards malignancy.

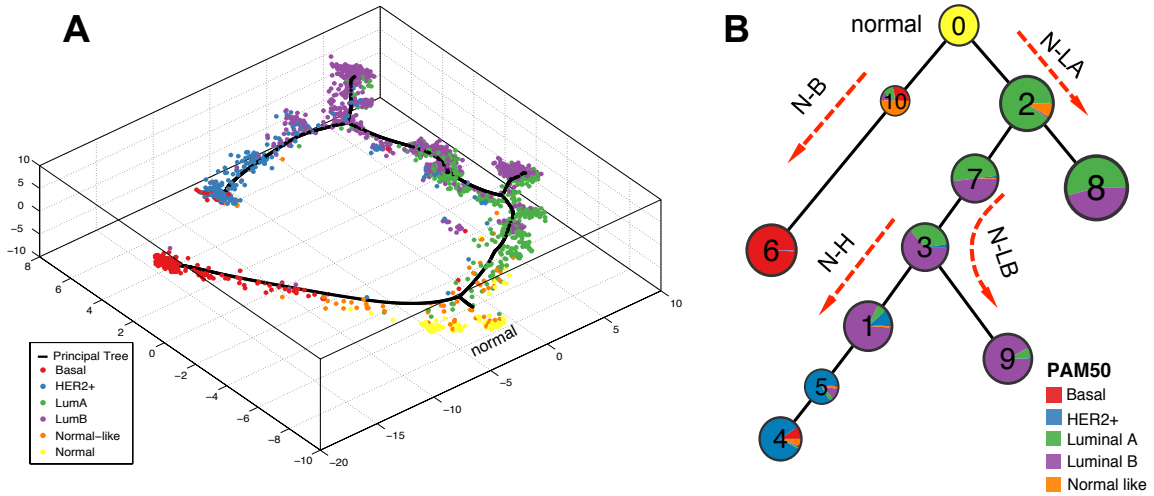

Figure 1: Breast cancer progression modeling analysis performed on METABRIC gene expression data (n = 1,989). (A) Data visualization analysis provides a general view of sample distribution. (B) Constructed progression model of breast cancer. The analysis revealed four major progression paths, referred to as N-B (normal to basal), N-H (normal through luminal A/B to HER2+), N-LB (normal through luminal A to luminal B terminus), and N-LA (normal to luminal A terminus). Each node represents an identified cluster, and the pie chart in each node depicts the percentage of the samples in the node belonging to one of the five PAM50 subtypes.

2 Results

2.1 Breast Cancer Progression Modeling

We applied the developed bioinformatics pipeline to the METABRIC data [6] to construct a progression model of breast cancer. The dataset contains the expression profiles of 25,160 genes obtained from 1,989 surgically excised breast tumor samples. Briefly, we first performed supervised learning by using the breast cancer subtypes as class labels and selected 359 disease related genes. Then, we performed a clustering analysis on the expression measures of the selected genes to detect genetically homogeneous groups. By using gap statistic [30] and consensus clustering [20], ten distinct clusters were identified. Finally, we constructed a progression model and represented it as an undirected graph, by using the centroids of the identified clusters as the vertices and connecting them based on the progression trend inferred from the analysis. To visualize the sample distribution, we also developed a graph model-based method for dimensionality reduction for structured data [19]. Figs. 1A-B show the distribution of the breast tumor samples in a three-dimensional space supported by the selected genes and the constructed model, respectively. The dataset also contains 144 normal breast tissue samples, which we used as the baseline to represent the origin of cancer progression. To help with visualization and to put the result into context by referring to previous classification systems, we added the PAM50 subtype labels [21] to the model. Our analysis identified a bifurcating progression model describing two distinct trajectories to malignancy, either directly to the basal-like subtype with little deviation, or a stepwise, more indolent path through the luminal subtypes to the HER2+ subtype. The two trajectory termini (i.e., HER2+ and basal) represent the two most aggressive breast tumor subtypes [25, 26]. Significant side-branches are also evident for both luminal A/B subtypes. Our findings support that molecular subtypes are not

hardwired, and genotypes and phenotypes can shift over time [11], and show that cancer development follows common progression paths, as posited in [29].

We performed a series of interrogations that provided substantial support for the constructed model, and showed the utility of such a model for testing and generating hypotheses and providing novel insights into previous observations from the cancer-evolution perspective [28]. First, we demonstrated that the same progression pattern was repeatedly observed in additional 26 breast cancer datasets, which together with the METABRIC dataset comprise almost all of the breast tumor samples assayed over the past 15 years ($> 9,000$ samples). Given the total number and diversity of samples, and the range of analytical platforms included in the combined datasets, the above analysis provides strong evidence suggesting that the data pattern presented in Fig. 1A-B is unlikely to be an artifact but universally present in breast cancer. Next, we performed a survival data analysis to examine the relationship between the clinical outcomes of the identified subgroups and the modeled disease progression paths to malignancy (Fig. 1C). Remarkably, Kaplan-Meier plots of disease-specific survival for the ten identified clusters revealed a clear trend of worsening survival function along the major trajectories to malignancy through normal to basal (node 10 to node 6), or to luminal A dead-end (node 2 to node 8), or to luminal B dead-end (node 2 through nodes 7, 3 to node 9), or through luminal types to HER2+ (node 2 through nodes 7, 3, 1, 5 to node 4). Finally, we mapped relevant clinical and genetic variables onto the constructed model to investigate how they correlate with identified progression paths. Examples with implications for cancer progression reported in the literature that we can test include deviation of morphological traits of tumor cells from normal cells [24] and the accumulation of genetic alterations [11, 34, 22, 27]. Figs. 1D-F show that molecular grade, genome instability, and overall mutation rate are highly correlated with disease progression along the normal-HER2+ path. More results and discussions can be found in [28].

2.2 Mapping of Matched Primary and Metastatic Tumors Samples

To further validate the constructed progression model, we investigated the progression relationships of 246 matched primary and metastasis (P/M) tumor samples collected from 123 breast cancer patients [3]. Each pair can be considered samples collected at two time points. As with other RNA-Seq data [9], the P/M dataset contains a large amount of missing data and outlier samples, which could complicate the downstream progression analysis. To address the issue, we performed a series of data pre-processing to remove genes and samples containing $\geq 20\%$ missing values, to impute missing data in the remaining samples, and to identify outlier data (see Material and Methods section). After data pre-processing, a total of 210 samples and 87 genes were retained for further analysis. By using the PAM50 classifier [21] implemented in the genefu package [8], we stratified 210 samples into five intrinsic molecular subtypes, including 3 normal-like, 79 luminal A, 78 luminal B, 33 HER2+ and 17 basal tumors, respectively. It has been reported that normal-like samples could be a technical artifact from high contamination with normal tissue [10]. We thus removed the three pairs that contain normal-like samples. Fig. 2A presents a Sankey plot showing the subtype change between each pair of primary and metastasis tumors.

We next performed a quantitative analysis of progression relationships of the primary and metastasis tumors by using three different strategies. First, we attempted to build a model by using the P/M data directly. However, due to the lack of normal samples and a large number of tumor samples (the prerequisite of progression modeling analysis is that there are a large number of tumor samples that populate progression paths), while the sample distribution of the P/M data is highly consistent with the METABRIC data, we were not able to construct the basal path (Supplementary Fig. X). To address the data scarcity issue, we mapped the P/M data back onto the METABRIC model. Since data sources are

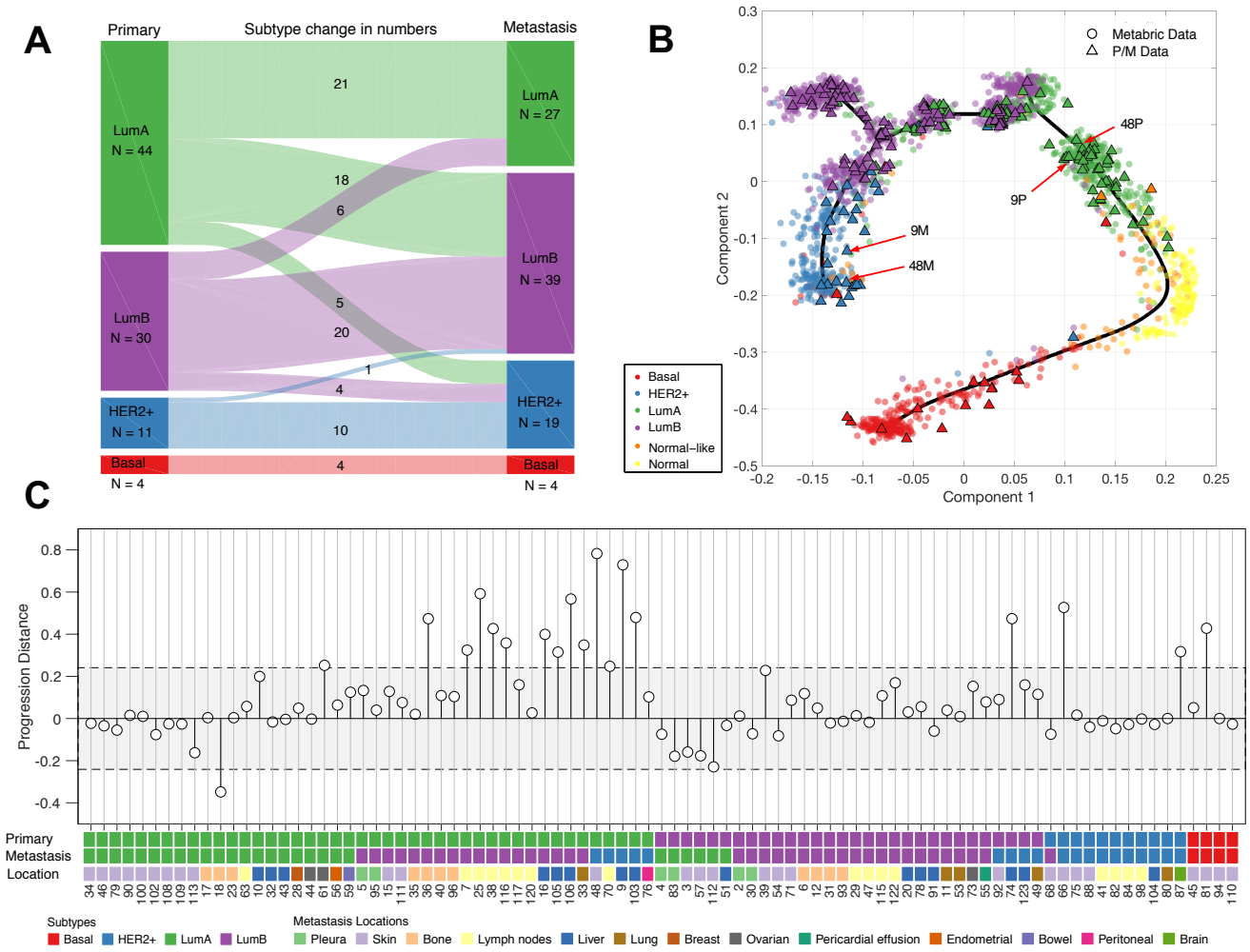

Figure 2: Progression analysis of 89 pairs of matched primary and metastasis (P/M) tumor samples. (A) Sankey diagram showing the subtype changes of matched tumor pairs. (B) Data visualization of the P/M cohort mapped onto the METABRIC progression model. Each sample was color-coded based on its PAM50 subtype. (C) Progression distances of the 89 pairs of matched tumors. A cut-off was estimated to identify tumor pairs with small progression distances that are likely due to random variations (shaded region). A total of 18 and 1 pairs were identified that had significant positive and negative progression distances, respectively.

not entirely compatible (the METABRIC data was obtained through the classic microarray technology [6], and the METABRIC model was built by using 350 genes [28]), we first performed batch effect correction by using the COMBAT method [15] and then proposed a novel strategy that mapped the P/M data onto a 350-dimensional space (see Methods Section). Fig. 2B visualizes the sample distributions of both P/M and METABRIC data, where each sample was color-coded based on its PAM50 label. By mapping the P/M sample onto the progression path and using the normal samples as the baseline to represent the origin of cancer progression, we calculated the progression distance of each pair of primary and metastasis tumors (see Material and Methods section) and reported the results in Fig. 2C. We noticed that many pairs had small distances, which are likely due to random variations. By assuming that progression distances of the tumor pairs that experienced little disease progression follow a zero-mean Gaussian distribution, we estimated the trimmed standard deviation σ of the distances using the values between 10% and 90% quantiles and identified tumor pairs with a progression distance outside of the $\pm 2\sigma$ band as those undergoing significant disease progression. A total of 18 and 1 pairs were identified that had significant positive and negative progression distances, respectively. Finally, we performed a joint progression modeling analysis by combining together the P/M and METABRIC data after data normalization described above. The topology structure of the constructed model (Supplementary Fig. X) is almost identical to that of the METABRIC model.

From Fig. 2A, we observed that most primary and metastatic tumors (62%) remain to be the same subtype, that basal cancer is a distinct disease entity, and that molecular subtypes are not hardwired, and can shift over time [11], which verifies the hypothesis proposed in [11]. However, the Sankey plot also suggested that cancer evolution is bi-directional, that is, luminal A and B can convert to each other and so is luminal B and HER2+. Since the PAM50 subtyping system can only provide a rough stratification of breast cancer [12, 4, 5], the P/M data measured only the expression levels of 87 genes, and RNA-Seq data usually contains considerable measurement errors, a qualitative analysis of subtype changes tells only part of the story. Indeed, as clearly shown in Fig. 2B, luminal A and B do not have a clear-cut boundary and so is luminal B and HER2+. This also explains why in some large-scaled benchmark studies it has been observed that existing molecular subtyping methods could only achieve moderate concordance, particularly when classifying luminal A and B tumors [14, 18, 32]. Through a close examination of the progression distances (Fig. 2C), we found that all six luminal B primary tumors that converted to luminal A metastatic tumors and one HER2+ primary tumor that changed to a luminal B metastatic tumor have very small progression distances, and that all 18 pairs with large positive progression distances are either luminal A $\rightarrow$ luminal B (9 pairs), luminal A $\rightarrow$ HER2+ (4), luminal B $\rightarrow$ HER2+ (3), or basal $\rightarrow$ basal (2). This is in full agreement with the METABRIC model. As two examples, we present in Fig. 2B the locations of two pairs (9P/9M, 48P/48M) that underwent evident disease progression. Interestingly, we observed that there is one pair of tumors (18P/18M), which were both classified as luminal A and had a large negative progression distance. One plausible explanation is that metastatictumor cells was spread from a primary site before the primary tumor was detected and the rate of cancer evolution was slower than that in the primary tumor.

Since the proposal of the breast cancer subtypes [25, 26], some fundamental issues regarding whether subtypes are biologically independent entities or there is a progressive relationship among them are still under debate [5, 1, 23, 13, 11, 33]. Our result suggested that while basal is biologically independent entity, other subtypes can shift over time and thus should be considered different stages of the same disease. Our study also suggested that cancer evolution is a unidirectional process and the direction is from normal $\rightarrow$ luminal A $\rightarrow$ luminal B $\rightarrow$ HER2+.

3 Discussion

Major efforts continue to catalog the genomic events associated with human cancers, but interpretation of the accumulating data to reveal insights into to the dynamic aspects of the disease are lacking. Here, we present a breast cancer progression model derived from METABRIC data [6] using an unbiased computational approach to the analysis of static tumor sample data [28]. The approach is based on the rationale that if the number of samples is sufficiently large, the genetic footprints of individual samples populate progression trajectories, enabling us to recover a representation of the dynamic disease process from static samples using a computational approach. The analytical strategy is equally applicable to model other cancers and other human progressive diseases where the lack of time-series data to study system dynamics is an unavoidable problem. Here, the application of the approach to a large-scale breast cancer genomic data set identified a bifurcating progression model describing two distinct pathways to breast cancer malignancy (Fig. 1A, B), either directly to the aggressive basal-like subtype(s) or a stepwise, more indolent path through the luminal subtypes to the HER2+ phenotype. The model supports the notion that genotypes and phenotypes can shift over time [11, 16, 2], and suggest that breast cancer development follows limited, common progression paths [29].

The concept of cancer evolution has long been proposed, and numerous studies have been conducted that significantly expanded our understanding of the concept [31, 35]. While there is currently no established progression model derived from human breast tumor tissue data, conceptual models regarding the origins of breast cancer subtypes have been proposed [1]. One model proposes a distinct-path scenario where each subtype follows a path of initiation and progression independently of the others. The alternative is a linear evolution model, which proposes that tumors gradually evolve from normal cells to malignant states through the accumulation of genetic alterations. While both models allow cancer evolution, one model suggests that the subtypes are different diseases, and the other suggests that tumor subtypes are different stages of the same disease. Clarifying this issue could have impacts on both research and clinical management strategies. Our modeling uses a generic approach that makes no model assumption *a priori*, and the bifurcation structure revealed in our model supports a linear evolution process of disease progression.

Using the proposed model of breast cancer progression as a foundation, we investigated the interrelationship between paired primary and metastatic samples with respect to molecular subtype and directional progression. Qualitative analysis revealed that most of the paired samples (62%) were of the same molecular subtype (Fig. 2A). But those that did change subtype, primarily shifted from a lesser to a more malignant phenotype, i.e., luminal A to B or to HER2+ subtype). However, there were several examples of a luminal B primary subtype shifting to a luminal A metastasis subtype. Given the approximation of the PAM50 subtyping system [12, 4, 5], the limited overlap of progression model and P/M data, and the error rate of RNASeq data, this may not be surprising. As shown in Fig. 1, luminal A and B do not have a clear-cut boundary on the progression model, as is the case for luminal B and HER2+ clusters. This boundary overlap of approximate subtypes explains why in some large-scaled benchmark studies it has been observed that existing molecular subtyping methods only achieve moderate concordance, particularly when classifying luminal A and B tumors [18]. Through a quantitative examination of the progression distances revealed by mapping to the model (Fig. 2B, C), we found that all six luminal B pairs and the one HER2+ primary tumor that changed to a luminal B metastatic tumor had progression distances within the standard deviation. Conversely, all 18 pairs with significant positive progression distances were either luminal A to luminal B (9 pairs), luminal A to HER2+ (4), luminal B to HER2+ (3), or basal to basal (2). As examples, Fig. 2B shows the locations of two pairs (9P/9M, 48P/48M) that underwent

evident disease progression.

Interestingly, one pair of tumor samples (i.e., 18P/18M) both classified as luminal A did have a significant negative progression distance (Fig. 2C). There are several plausible, but speculative explanations for this phenomenon in a background of the observed 'downstream' P-to-M trend. Although the probability of tumor cell dissemination and survival as metastases increases as the primary tumor grows and becomes more genetically complex, tumor cells can escape the primary and survive in a secondary location at early stages of primary tumor development. The evolutionary time index may be very different for these foci if the metastasis lesion was dormant for a period [7], and the local selective pressures that the primary and the established secondary lesion could differentially impact the evolution of those separate but associated lesions. A more mundane explanation may allude to the pathological evaluation of tumor tissue samples. A relatively small area of the primary tumor may be sourced for subtyping, and the result may not represent the bulk of the tumor. Such limitations point to the need for improved tumor evaluation tools and approaches for accurate prognostication. Given these caveats, the clear trend downstream towards malignant model termini is strengthened.

Our findings support a linear, branching model of breast cancer evolution. The interpretation from the model is that the basal subtype is distinct from the luminal subtypes, and that the luminal subtypes do shift and may be considered different stages of the same disease. The mapping on to the progression model of the paired primary/metastasis samples further supports the overall model, and the concept that the evolutionary process is unidirectional through luminal subtypes towards the most malignant state.

Future studies using higher-resolution genomic methods, such as single-cell sequencing and tissue microdissection and guided by a working model can provide data for the refinement of a roadmap of breast cancer progression. An accurate progression model can provide a foundation for the visualization of key progressive molecular events and facilitate the identification of pivotal driver genes and pathways, the most relevant and robust biomarker signatures for patient evaluation, and potential points of susceptibility for therapeutic intervention.

4 Methods

4.1 Progression Modeling Analysis of METABRIC Data

We have previously developed CancerMapp, a comprehensive bioinformatics pipeline for cancer progression modeling using static tumor sample data [28]. For the purpose of this study, we reconstructed a progression model of breast cancer by analyzing data from the METABRIC project [6]. The dataset contains the expression profiles of 25,160 genes obtained from 1,989 breast tumor samples. Briefly, we first performed supervised learning by using the breast cancer subtypes as class labels to select disease related genes. Then, we performed a clustering analysis on the expression measures of the selected genes to detect genetically homogeneous groups. By using gap statistic [30] and consensus clustering [20], ten distinct clusters were identified. Finally, we constructed a progression model and represented it as an undirected graph, by using the centroids of the identified clusters as the vertices and connecting them based on the progression trend inferred from the analysis. To visualize the sample distribution, we also developed a graph model-based method for dimensionality reduction for structured data. The dataset also contains 144 normal breast tissue samples, which we used as the baseline to represent the origin of cancer progression.

4.2 Progression Analysis of Matched Primary and Metastasis Data

The gene expression dataset was downloaded from Gene Expression Omnibus (GSE92977). It contains 246 matched primary and metastasis (P/M) tumor samples from 123 breast cancer patients [3]. For each sample, 105 breast cancer-related genes and 5 house-keeping genes were sequenced. The data was log-transformed (base 2) and normalized using the house-keeping genes.

4.2.1 Data Filtering, Missing Value Imputation and Outlier Removal

The P/M dataset contains a large amount of missing data. We first filtered out 11 genes and 13 samples containing $> 20\%$ missing values, and then performed missing data imputation on the remaining samples. Specifically, for each sample, we found its 10 nearest neighbors (NNs) and replaced the missing data in each gene in the sample by the average of the observed values of the corresponding gene in the identified NNs. As with other RNAseq data [9], the P/M dataset also contains a considerable number of outlier samples, which would complicate downstream analysis. Intuitively, an outlier sample should be dissimilar to its NNs. We exploited this to identify outliers by calculating the average distance between each sample and its 10-NNs and removing those with the top 10% average distances.

4.2.2 Subtype Classification and Gene Selection

By using the PAM50 classifier implemented in the genefu package [8], we classified the remaining tumor samples into the five intrinsic molecular subtypes (luminal A, luminal B, HER2+, basal and normal-like). A Sankey diagram was generated to visualize the shifts of subtypes between the primary and metastasis tumors. Based on the obtained subtype information, we also performed a one-way ANOVA analysis on individual genes and removed 12 genes that contain little information in discriminating the five molecular subtypes (p-value > 0.1). After data pre-processing, 87 genes and 210 samples including 92 pairs were retained for the further analysis.

4.2.3 Progression Modeling Analysis

We investigated the progression relationships between the matched primary and metastasis tumors by mapping the P/M data back onto the METABRIC model. Due to the use of different platforms in generating gene expression data (the METABRIC study used a microarray technique whereas the P/M data was generated through RNAseq), and the fact that the METABRIC model was constructed using 359 genes [28] while the P/M data contains only the expression data of 87 genes, the data source is not entirely compatible. To remove the platform-induced data deviations, we first mapped the 87 genes in the P/M cohort back to the 25,160 genes in METABRIC data and identified 85 genes present in both datasets. Then, we performed batch-effect correction using ComBat [15] implemented in the SVA package [17]. To address the second issue, we designed a novel procedure that projects the P/M samples onto a high-dimensional space spanned by the 359 genes used to build the METABRIC model. Specifically, we assumed that the projection relationships between the 85 and 359 genes were shared by both cohorts, learned a projection function by using the METABRIC data, and applied the function to the P/M data.

Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ and $\mathbf{X}' = [\mathbf{x}'_1, \cdots, \mathbf{x}'_N]$ be the METABRIC data in the 85- and 359-dimensional spaces, respectively, and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_M]$ and $\mathbf{Y}' = [\mathbf{y}'_1, \cdots, \mathbf{y}'_M]$ be the P/M data in the 85- and 359-dimensional spaces, respectively. We formulated the projection relationship as:

$$f(\mathbf{x}_n) = \mathbf{x}'_n + \boldsymbol{\epsilon}_x, 1 \leq n \leq N \ , \tag{1}$$

$$f(\mathbf{y}_m) = \mathbf{y}'_m + \boldsymbol{\epsilon}_y, 1 \leq m \leq M \ , \tag{2}$$

where $f(\cdot)$ is a mapping function that projects low-dimensional data to a high-dimensional space, and $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_y$ are residual terms. In principle, $f(\cdot)$ can be a non-linear function to represent a complex relationship between two gene sets. However, in this study, we found that a linear form suffices for our purposes. By using a linear projection, Eqs. (1-2) can be respectively re-written as

$$\mathbf{W}\mathbf{x}_n + \mathbf{b} = \mathbf{x}'_n + \boldsymbol{\epsilon}_x, 1 \leq n \leq N \ , \tag{3}$$

$$\mathbf{W}\mathbf{y}_m + \mathbf{b} = \mathbf{y}'_m + \boldsymbol{\epsilon}_y, 1 \leq m \leq M \ , \tag{4}$$

where $\mathbf{W}$ is a projection matrix and $\mathbf{b}$ is a bias term, which can be estimated by solving the following least-squares optimization problem:

$$(\mathbf{W}^*, \mathbf{b}^*) = \underset{(\mathbf{W},\mathbf{b})}{\arg\min} \sum_{n=1}^{N} \|\mathbf{W}\mathbf{x}_n + \mathbf{b} - \mathbf{x}'_n\|_2^2 \ . \tag{5}$$

Once we obtained $(\mathbf{W}^*, \mathbf{b}^*)$, $\mathbf{Y}'$ can be estimated as

$$\bar{\mathbf{y}}'_m = \mathbf{W}^*\mathbf{y}_m + \mathbf{b}^*, 1 \leq m \leq M \ . \tag{6}$$

After we obtained the high-dimensional prediction of the P/M data, we projected each sample onto the progression paths identified in the METABRIC model. Here, the projection of a sample is defined as a point on the progression path that is the closest to the sample. By using the mean of the normal samples as the root to represent the origin of cancer progression, we calculated the progression distances of the matched primary and metastasis samples and reported the results in Fig. 2C. We noticed that many pairs had very small distances, likely due to random variations. By assuming that progression distances of the tumor pairs that experienced little disease progression follow a Gaussian distribution, we estimated the trimmed standard deviation of the distances using the values between 10% and 90% quantiles and identified tumor pairs with a progression distance outside of the $\pm 2\sigma$ band as those undergoing significant disease progression.

References

[1] W. F. Anderson, P. S. Rosenberg, A. Prat, C. M. Perou, and M. E. Sherman. How many etiological subtypes of breast cancer: two, three, four, or more? *Journal of the National Cancer Institute*, 106(8):165, 2014.

[2] L. Bao, R. D. Cardiff, P. Steinbach, K. S. Messer, and L. G. Ellies. Multipotent luminal mammary cancer stem cells model tumor heterogeneity. *Breast Cancer Research*, 17(1):137, 2015.

[3] J. M. Cejalvo, E. M. de Dueñas, P. Galvan, S. García-Recio, O. B. Gasión, L. Paré, S. Antolin, R. Martinello, I. Blancas, B. Adamo, et al. Intrinsic subtypes and gene expression profiles in primary and metastatic breast cancer. *Cancer Research*, 77(9):2213–2221, 2017.

[4] G. Ciriello, R. Sinha, K. A. Hoadley, A. S. Jacobsen, B. Reva, C. M. Perou, C. Sander, and N. Schultz. The molecular diversity of luminal a breast tumors. *Breast Cancer Research and Treatment*, 141(3):409–420, 2013.

[5] C. J. Creighton. The molecular profile of luminal B breast cancer. *Biologics: Targets & Therapy*, 6:289, 2012.

[6] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[7] J. Dittmer. Mechanisms governing metastatic dormancy in breast cancer. *Seminars in Cancer Biology*, 44:72–82, 2017.

[8] D. M. Gendoo, N. Ratanasirigulchai, M. S. Schroder, L. Pare, J. S. Parker, A. Prat, and B. Haibe-Kains. *genefu: computation of gene expression-based signatures in breast cancer*, 2015. R package version 2.4.2.

[9] N. I. George, J. F. Bowyer, N. M. Crabtree, and C.-W. Chang. An iterative leave-one-out approach to outlier detection in RNA-Seq data. *PLoS ONE*, 10(6):e0125224, 2015.

[10] A. Goldhirsch, W. C. Wood, A. S. Coates, R. D. Gelber, B. Thürlimann, H.-J. Senn, and P. Members. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology*, 22(8):1736–1747, 2011.

[11] M. Greaves and C. C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[12] S. Guiu, S. Michiels, F. Andre, J. Cortes, C. Denkert, A. Di Leo, B. Hennessy, T. Sorlie, C. Sotiriou, N. Turner, M. Van de Vijver, G. Viale, S. Loi, and J. S. Reis-Filho. Molecular subclasses of breast cancer: how do we define them? The IMPAKT 2012 Working Group Statement. *Annals of Oncology*, 23(12):2997–3006, 2012.

[13] B. Gusterson. Do 'basal-like' breast cancers really exist? *Nature Reviews Cancer*, 9(2):128–134, 2009.

[14] B. Haibe-Kains, C. Desmedt, S. Loi, A. C. Culhane, G. Bontempi, J. Quackenbush, and C. Sotiriou. A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of National Cancer Institute*, 104(4):311–325, 2012.

[15] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[16] S. Kimbung, A. Kovács, A. Danielsson, P.-O. Bendahl, K. Lövgren, M. F. Stolt, N. P. Tobin, L. Lindström, J. Bergh, Z. Einbeigi, et al. Contrasting breast cancer molecular subtypes across serial tumor progression stages: biological and prognostic implications. *Oncotarget*, 6(32):33306, 2015.

[17] J. T. Leek, W. E. Johnson, H. S. Parker, E. J. Fertig, A. E. Jaffe, and J. D. Storey. *SVA: surrogate variable analysis*, 2016. R package version 3.20.0.

[18] A. Mackay, B. Weigelt, A. Grigoriadis, B. Kreike, R. Natrajan, R. A'Hern, D. S. Tan, M. Dowsett, A. Ashworth, and J. Reis-Filho. Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *Journal of National Cancer Institute*, 103(8):662–673, 2011.

[19] Q. Mao, L. Wang, S. Goodison, and Y. Sun. Dimensionality reduction via graph structure learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 765–774, 2015.

[20] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003.

[21] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, 2009.

[22] O. Podlaha, M. Riester, S. De, and F. Michor. Evolution of the cancer genome. *Trends in Genetics*, 28(4):155–163, 2012.

[23] K. Polyak. Breast cancer: origins and evolution. *The Journal of Clinical Investigation*, 117(11):3155–3163, 2007.

[24] E. A. Rakha, M. E. El-Sayed, A. H. Lee, C. W. Elston, M. J. Grainge, Z. Hodi, R. W. Blamey, and I. O. Ellis. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. *Journal of Clinical Oncology*, 26(19):3153–3158, 2008.

[25] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A.-L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[26] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A.-L. Børresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423, 2003.

[27] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[28] Y. Sun, J. Yao, L. Yang, R. Chen, N. Nowak, and S. Goodison. Computational approach for deriving cancer progression roadmaps from static sample data. *Nucleic Acids Research*, 45(9):e69, 2017.

[29] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

[30] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63(2):411–423, 2001.

[31] Y. Wang, J. Waters, M. L. Leung, A. Unruh, W. Roh, X. Shi, K. Chen, P. Scheet, S. Vattathil, H. Liang, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 2014.

[32] B. Weigelt, A. Mackay, R. A'hern, R. Natrajan, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology*, 11(4):339–349, 2010.

[33] B. Weigelt and J. S. Reis-Filho. Histological and molecular types of breast cancer: is there a unifying taxonomy? *Nature Reviews Clinical Oncology*, 6(12):718–730, 2009.

[34] L. R. Yates and P. J. Campbell. Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806, 2012.

[35] L. R. Yates, S. Knappskog, D. Wedge, J. H. Farmery, S. Gonzalez, I. Martincorena, L. B. Alexandrov, P. Van Loo, H. K. Haugland, P. K. Lilleng, et al. Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell*, 32(2):169–184, 2017.