A data-driven interactome of synergistic genes improves network-based cancer outcome prediction

Allahyar, Amin; Ubels, Joske; de Ridder, Jeroen

DOI
10.1371/journal.pcbi.1006657

Publication date
2019

Document Version
Final published version

Published in
PLoS Computational Biology

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A data-driven interactome of synergistic genes improves network-based cancer outcome prediction

Amin Allahyar[1,2], **Joske Ubels**[1,3,4], **Jeroen de Ridder**[1]*

1 Department of Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands, **2** Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands, **3** Skyline DX, Rotterdam, **4** Department of Hematology, Erasmus MC Cancer Institute, Rotterdam

* J.deRidder-4@umcutrecht.nl

Citation: Allahyar A, Ubels J, de Ridder J (2019) A data-driven interactome of synergistic genes improves network-based cancer outcome prediction. PLoS Comput Biol 15(2): e1006657. https://doi.org/10.1371/journal.pcbi.1006657

Editor: Edwin Wang, University of Calgary Cumming School of Medicine, CANADA

Received: May 13, 2018

Accepted: November 20, 2018

Published: February 6, 2019

Data Availability Statement: The normalized and batch effect removed cohort can be downloaded from SyNet.deRidderLab.nl. This collection excludes METABRIC expressions. METABRIC requires access approval through synapse.org, which is implemented by the original data collectors to protect the privacy and confidentiality of participants in this study. Additional clinical variables for samples collected in this study can be downloaded from S1 Table. The full matrix of SyNet in binary format as well as top gene pairs (including their tri-score used to calculate their fitness) is available for download in tab-delimited

Abstract

Robustly predicting outcome for cancer patients from gene expression is an important challenge on the road to better personalized treatment. Network-based outcome predictors (NOPs), which considers the cellular wiring diagram in the classification, hold much promise to improve performance, stability and interpretability of identified marker genes. Problematically, reports on the efficacy of NOPs are conflicting and for instance suggest that utilizing random networks performs on par to networks that describe biologically relevant interactions. In this paper we turn the prediction problem around: instead of using a given biological network in the NOP, we aim to identify the network of genes that truly improves outcome prediction. To this end, we propose SyNet, a gene network constructed ab initio from synergistic gene pairs derived from survival-labelled gene expression data. To obtain SyNet, we evaluate synergy for all 69 million pairwise combinations of genes resulting in a network that is specific to the dataset and phenotype under study and can be used to in a NOP model. We evaluated SyNet and 11 other networks on a compendium dataset of >4000 survival-labelled breast cancer samples. For this purpose, we used cross-study validation which more closely emulates real world application of these outcome predictors. We find that SyNet is the only network that truly improves performance, stability and interpretability in several existing NOPs. We show that SyNet overlaps significantly with existing gene networks, and can be confidently predicted (~85% AUC) from graph-topological descriptions of these networks, in particular the breast tissue-specific network. Due to its data-driven nature, SyNet is not biased to well-studied genes and thus facilitates post-hoc interpretation. We find that SyNet is highly enriched for known breast cancer genes and genes related to e.g. histological grade and tamoxifen resistance, suggestive of a role in determining breast cancer outcome.

format from SyNet.deRidderLab.nl. Moreover, all scripts used for preparation of data and figures in this manuscript are available for download from github.com/UMCUGenetics/SyNet. To ensure the complete reproducibility of our results, the indices utilized for training and testing of all models (including inner and outer cross-validations) are also available for download through Mendeley data repository http://dx.doi.org/10.17632/c55f2v9dzj.1.

Funding: Part of computations required for this work was carried out on the Dutch national e-infrastructure (e-infra160001) with the support of the SURF Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Cancer is caused by disrupted activity of several pathways. Therefore, to predict cancer patient prognosis from gene expression profiles, it may be beneficial to consider the cellular interactome (e.g. the protein interaction network). These so-called Network based Outcome Predictors (NOPs) hold the potential to facilitate identification of dysregulated pathways and delivering improved prognosis. Nonetheless, recent studies revealed that compared to classical models, neither performance nor consistency (in terms of identified markers across independent studies) can be improved using NOPs. In this work, we argue that NOPs can only perform well when supplied with suitable networks. The commonly used networks may miss associations specially for under-studied genes. Additionally, these networks are often generic with low coverage of perturbations that arise in cancer. To address this issue, we exploit ~4100 samples and infer a disease-specific network called SyNet linking synergistic gene pairs that collectively show predictivity beyond the individual performance of genes. Using a thorough cross-validation, we show that a NOP yields superior performance and that this performance gain is the result of the wiring of genes in SyNet. Due to simplicity of our approach, this framework can be used for any phenotype of interest. Our findings confirm the value of network-based models and the crucial role of the interactome in improving outcome prediction.

Introduction

Metastases at distant sites (e.g. in bone, lung, liver and brain) is the major cause of death in breast cancer patients [1]. However, it is currently difficult to assess tumor progression in these patients using common clinical variables (e.g. tumor size, lymph-node status, etc.) [2]. Therefore, for 80% of these patients, chemotherapy is prescribed [3]. Meanwhile, randomized clinical trials showed that at least 40% of these patients survive without chemotherapy and thus unnecessarily suffer from the toxic side effect of this treatment [3, 4]. For this reason, substantial efforts have been made to derive molecular classifiers that can predict clinical outcome based on gene expression profiles obtained from the primary tumor at the time of diagnosis [5, 6].

An important shortcoming in molecular classification is that 'cross-study' generalization is often poor [7]. This means that prediction performance decreases dramatically when a classifier trained on one patient cohort is applied to another one [8]. Moreover, the gene signatures found by these classifiers vary greatly, often sharing only few or no genes at all [9–11]. This lack of consistency casts doubt on whether the identified signatures capture true 'driver' mechanisms of the disease or rather subsidiary 'passenger' effects [12].

Several reasons for this lack of consistency have been proposed, including small sample size [11, 13, 14], inherent measurement noise [15] and batch effects [16, 17]. Apart from these technical explanations, it is recognized that traditional models ignore the fact that genes are organized in pathways [18]. One important cancer hallmark is that perturbation of these pathways may be caused by deregulation of disparate sets of genes which in turn complicates marker gene discovery [19, 20].

To alleviate these limitations, the classical models are superseded by Network-based Outcome Predictors (NOP) which incorporate gene interactions in the prediction model [21]. NOPs have two fundamental components: aggregation and prediction. In the aggregation step, genes that interact, belong to the same pathway or otherwise share functional relation are aggregated (typically by averaging expressions) into so called "meta-genes" [22]. This step is

guided by a supporting data source describing gene-gene interactions such as cellular pathway maps or protein-protein interaction networks. In the consequent prediction step, meta-genes are selected and combined into a trained classifier, similar to a traditional classification approach. Several NOPs have been reported to exhibit improved discriminative power, enhanced stability of the classification performance and signature and better representation of underlying driving mechanisms of the disease [18, 23–25].

In recent years, a range of improvements to the original NOP formulation has been proposed. In the prediction step, various linear and nonlinear classifiers have been evaluated[26, 27]. Problematically, the reported accuracies are often an overestimation as many studies neglected to use cross-study evaluation scheme which more closely resembles the real-world application of these models [7]. Also for the aggregation step, which is responsible for forming meta-genes from gene sets, several distinct approaches are proposed such as clustering [23] and greedy expansion of seed genes into subnetworks [18]. Moreover, in addition to simple averaging, alternative means by which genes can be aggregated, such as linear or nonlinear embeddings, have been proposed [17, 28]. Most recent work combines these steps into a unified model [8, 29]. Recent efforts that extend these concepts to sequencing data by exploiting the concept of cancer hallmark networks have also been proposed [30].

Despite these efforts and initial positive findings, there is still much debate over the utility of NOPs compared to classical methods, with several studies showing no performance improvement [21, 31, 32]. Perhaps even more striking is the finding that utilizing a permuted network [32] or aggregating random genes [10] performs on par with networks describing true biological relationships. Several meta-analyses attempting to establish the utility of NOPs have appeared with contradicting conclusions. Notably, Staiger et al. compared performance of nearest mean classifier [33] in this setting and concluded that network derived meta-genes are not more predictive than individual genes [21, 32]. This is in contradiction to Roy et al. who achieved improvements in outcome prediction when genes were ranked according to their t-test statistics compared to their page rank property [34] in PPI network [28, 35]. It is thus still an open question whether NOPs truly improve outcome prediction in terms of predictive performance, cross-study robustness or interpretability of the gene signatures.

A critical—yet often neglected—aspect in the successful application of NOPs is the contribution of the biological network. In this regard, it should be recognized that many network links are unreliable [36, 37], missing [38] or redundant [39] and considerable efforts are being made to refine these networks [38, 40–42]. In addition, many links in these networks are experimentally obtained from model organisms and therefore may not be functional in human cells [43–45]. Finally, most biological networks capture only a part of a cell's multifaceted system [46]. This incomplete perspective may not be sufficient to link the wide range of aberrations that may occur in a complex and heterogeneous disease such as breast cancer [47, 48]. Taken together, these issues raise concerns regarding the extent to which the outcome predictors may benefit from inclusion of common biological networks in their models.

In this work, we propose to construct a network *ab initio* that is specifically designed to improve outcome prediction in terms of cross-study generalization and performance stability. To achieve this, we will effectively turn the problem around: instead of using a given biological network, we aim to use the labelled gene expression datasets to identify the network of genes that truly improves outcome prediction (see Fig 1 for a schematic overview).

Our approach relies on the identification of *synergistic gene pairs*, i.e. genes whose joint prediction power is beyond what is attainable by both genes individually [49]. To identify these pairs, we employed grid computing to evaluate all 69 million pairwise combinations of genes. The resulting network, called SyNet, is specific to the dataset and phenotype under study and can be used to infer a NOP model with improved performance.

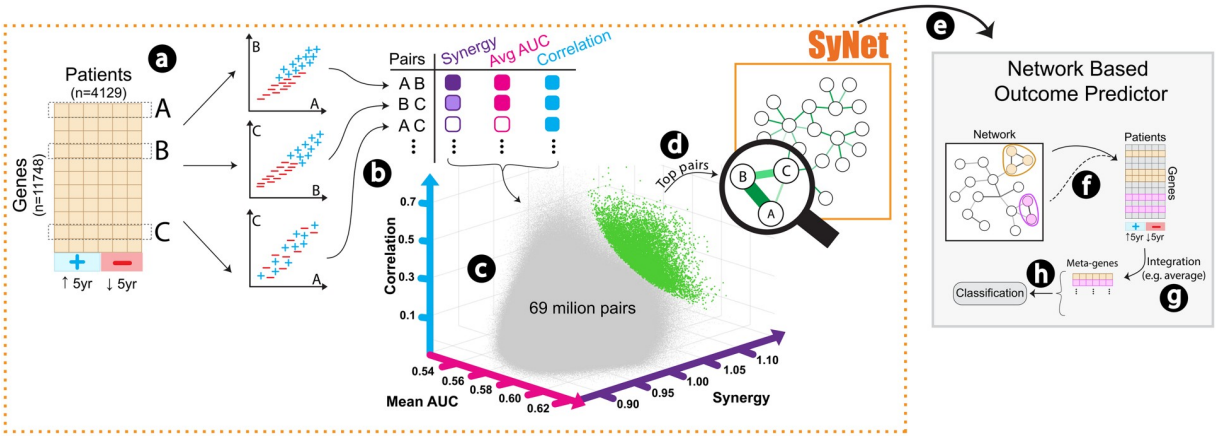

Fig 1. Schematic overview of SyNet inference and NOP training. For every 69 million combinations of gene pairs **(a)** we compute three criteria including synergy (S_{ij}, purple), average AUC (M_{ij}, pink), and correlation (C_{ij}, blue) **(b)**. These three criteria form a three-dimensional space **(c)** from which Fitness (F_{ij}) can be calculated for each pair. Top pairs (green dots) in this space collectively form SyNet **(d)**. SyNet is subsequently used in a NOP **(e)**, in which the links in SyNet guide the construction of "meta-genes". Within a NOP, groups of genes are formed **(f)** and then integrated into meta-genes (typically using averaging) **(g)**. The constructed meta-genes are then used as regular features to train standard classifiers **(h)**. The phenotype of interest is patient outcome (i.e. 5-year survival).

https://doi.org/10.1371/journal.pcbi.1006657.g001

To obtain SyNet, and allow for rigorous cross-study validation, a dataset of substantial size is required. For this reason, we combined 14 publicly available datasets to form a compendium encompassing 4129 survival labeled samples. To the best of our knowledge, the data combined in this study represents the largest breast cancer gene expression compendium to date. Further, to ensure unbiased evaluation, sample assignments in the inner as well as the outer cross-validations folds are kept equal across all assessments throughout the paper.

In the remainder of this paper, we will demonstrate that integrating genes based on SyNet provides superior performance and stability of predictions when these models are tested on independent cohorts. In contrast to previous reports, where shuffled versions of networks also performed well, we show that the performance drops substantially when SyNet links are shuffled (while containing the same set of genes), suggesting that SyNet connections are truly informative. We further evaluate the content and structure of SyNet by overlaying it with known gene sets and existing networks, revealing marked enrichment for known breast cancer prognostic markers. While overlap with existing networks is highly significant, the majority of direct links in SyNet is absent from these networks explaining the observed lack of performance when NOPs are guided by the phenotype-unaware networks. Interestingly, SyNet links can be reliably predicted from existing networks when more complex topological descriptors are employed. Taken together, our findings suggest that compared to generic gene networks, phenotype-specific networks, which are derived directly from labeled data, can provide superior performance while at the same time revealing valuable insight into etiology of breast cancer.

Results

SyNet improves NOP performance

We first evaluated NOP performance for three existing methods (Park, Chuang and Taylor) and the Group Lasso (GL) when supplied with a range of networks, including generic networks, tissue-specific networks and SyNet. As a baseline model, we used a Lasso classifier trained using all genes in our expression dataset (n = 11748) without network guidance. The

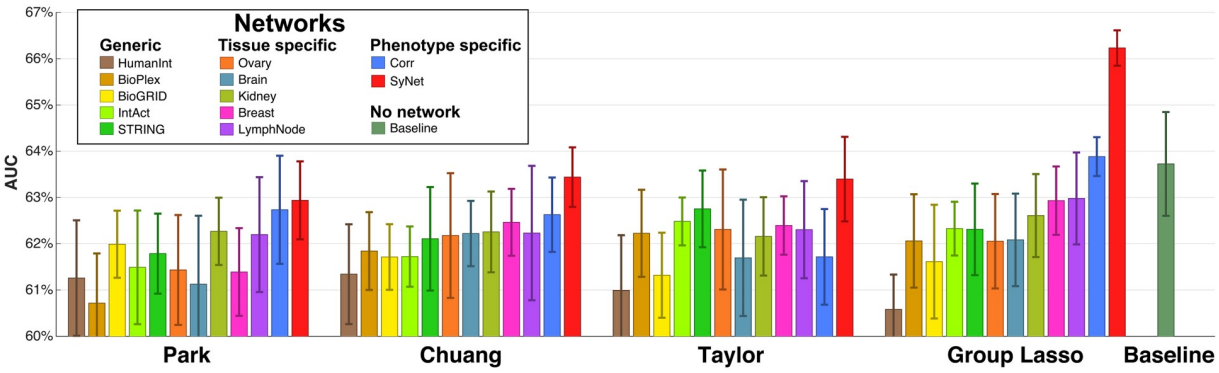

Fig 2. Performance comparison of NOPs for 4 methods and 12 networks including SyNet. Bars represent the averaged performance in terms of the AUC and error bars represent the standard deviation of performances across 10 repeats. The rightmost bar represents the performance of standard Lasso which considers all individual genes as features (i.e. no network is used in this model).

https://doi.org/10.1371/journal.pcbi.1006657.g002

Lasso exhibits superior performance among many linear and non-linear classifiers evaluated on our expression dataset (see **S3** for details).

The AUC of the four NOPs, presented in Fig 2, clearly demonstrates that SyNet improves the performance of all NOPs, except for the Park method in which it performs on par to the Correlation (Corr) network. Notably, SyNet is inferred using training samples only, which prevents "selection bias" in our assessments [50]. Furthermore, comparison of baseline model performance (i.e. Fig 2, rightmost bar) and other NOPs supports previous findings that many existing NOPs do not outperform regular classifiers that do not use networks [8, 21, 32].

The GL clearly outperforms all other methods, in particular when it exploits the information contained in SyNet. This corroborates our previous finding [8] that existing methods which construct meta-genes by averaging are suboptimal (see **S1** for a more extensive analysis). The GL using the Corr network also outperforms the baseline model, albeit non-significantly ($p \sim 0.6$), which is in line with previous reports [23]. It should be noted that across all these experiments an identical set of samples is used to train the models so that any performance deviation must be due to differences in (i) the set of utilized genes or (ii) the integration of the genes into meta-genes. In the next two sections, we will investigate these factors in more details.

SyNet provides feature selection capabilities

Networks only include genes that are linked to at least one other gene. As a result, networks can provide a way of ranking genes based on the number and weight of their connections. One explanation for why NOPs can outperform regular classifiers is that networks provide an a priori gene (feature) selection [32]. To test this hypothesis and determine the feature selection capabilities of SyNet, we compare classification performances obtained using the baseline classifier (i.e. Lasso) that is trained using enclosed genes in each network. While this classifier performs well compared to other standard classifiers that we investigated (see **S3** for details), it cannot exploit information contained in the links of given network. So, any performance difference must be due to the genes in the network. The number of genes in each network under study is optimized independently by varying the threshold on the weighted edges in the network and removing unconnected genes (see section "Regular classifiers and Network based prediction models" for network size optimization details). The edge weight threshold and the Lasso regularization parameter were determined simultaneously using a grid search cross-validation scheme (see **S5** for details). Fig 3 provides the optimal performances for 12 distinct

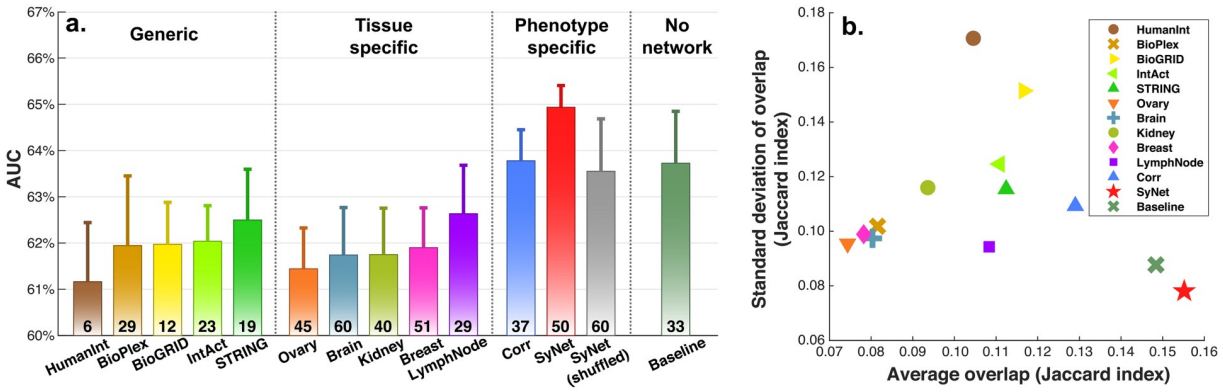

Fig 3. Performance comparison between networks when interconnections are ignored and genes contained in each network are utilized to train Lasso. a. Performance (AUC) of Lasso classification using individual genes in 12 networks. Numbers below each bar represent the median number of non-zero coefficients after training Lasso across 10 repeats and 14 folds. **b**. Stability of identified signatures measured by overlap between identified gene sets using Jaccard index. X and y-axis represent average and standard deviation of the Jaccard index measured across 10 repeats and 14 folds.

https://doi.org/10.1371/journal.pcbi.1006657.g003

networks along with number of genes used in the final model (i.e. genes with non-zero Lasso coefficients). We also included the baseline model where all genes (n = 11748) are utilized to train Lasso classifier (rightmost bar).

The results presented in Fig 3a demonstrate that SyNet is the only network that performs markedly better than the baseline model which is trained on all genes. Interestingly, we observe that SyNet is the top performing network while utilizing a comparable number of genes to other networks. The second-best network is the Corr network. We argue that superior performance of SyNet over the Corr network stems from the disease specificity of genes in SyNet which helps the predictor to focus on the relevant genes only. It should be noted that the data on which SyNet and the Corr network are constructed are completely independent from the validation data on which the performance is based due to our multi-layer cross-validation scheme (see Methods and **S5**) which avoids selection bias [50]. We conclude that dataset-specific networks, in particular SyNet which also exploits label information, provides a meaningful feature selection that is beneficial for classification performance.

Our result show that none of the tissue-specific networks outperform the baseline. Despite the modest performance, it is interesting to observe that performance for these networks increases as more relevant tissues (e.g. breast and lymph node networks) are utilized in the classification. Additionally, we observe that tissue-specific networks do not outperform the generic networks. This may be the result of the fact that generic networks predominantly contain broadly expressed genes with fundamental roles in cell function which may still be relevant to survival prediction. A similar observation was made for GWAS where SNPs in these widely-expressed genes can explain a substantial amount of missed heritability [51].

In addition to classifier performance, an important motivation for employing NOPs is to identify stable gene signatures, that is, the same genes are selected irrespective of the study used to train the models. Gene signature stability is necessary to confirm that the identified genes are independent of dataset specific variations and therefore are true biological drivers of the disease under study. To measure the signature consistency, we assessed the overlap of selected genes across all repeats and folds using the Jaccard Index. Fig 3b shows that a Lasso trained using genes preselected by SyNet, identifies more similar genes across folds and studies compared to other networks. Surprisingly, despite the fact that the expression data from which SyNet is inferred changes in each classification fold, the signature stability for SyNet is

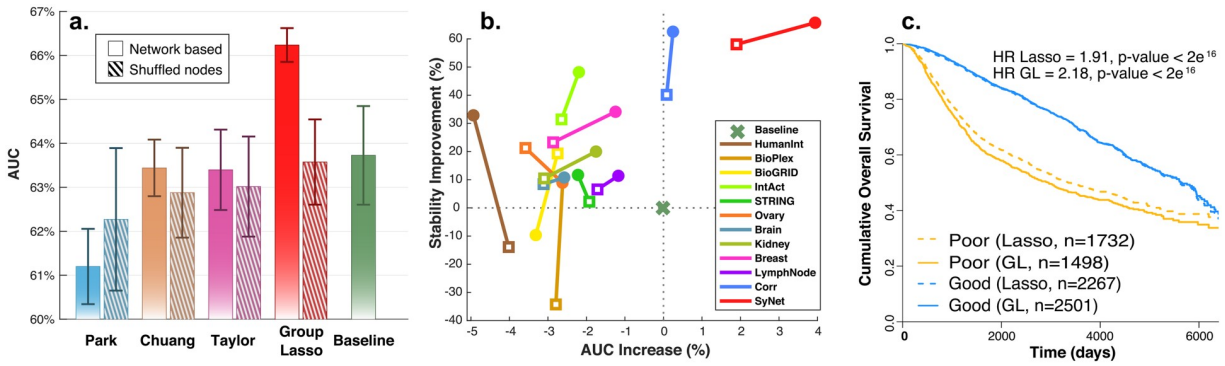

Fig 4. Performance of NOP models trained using SyNet compared to a shuffled version of this network (i.e. the same genes are present but randomly connected while keeping their degree intact). a. Bars indicate average performance of models across repeats and error bars denote the corresponding standard deviation. Solid bars represent average performance of models trained using SyNet. Dashed bars denote performance of the same model using shuffled SyNet. **b**. Improvement of performance (x-axis) and stability (in terms of the standard deviation of the AUC; y-axis) compared to the baseline model. Square and circle markers represent performance obtained using genes only (i.e. Lasso) and the network (i.e. GL), respectively. **c**. Kaplan-Meier plot for patients predicted to have good or poor prognosis. Dashed lines represent the Lasso prediction and solid lines the Group Lasso (GL) prediction.

https://doi.org/10.1371/journal.pcbi.1006657.g004

markedly better than for generic or tissue-specific networks that use a fixed set of genes across folds. Therefore, our results demonstrate that synergistic genes in SyNet truly aid the classifier to robustly select signatures across independent studies.

SyNet connections are beneficial for NOP

The ultimate goal of employing NOPs compared to classical models that do not use network information is to improve prognosis prediction by harnessing the information contained in the links of the given network. Therefore, we next aimed to assess to what extent also connections between the genes, as captured in SyNet and other networks, can help NOPs to improve their performance beyond what is achievable using individual genes. As before, we utilized identical datasets (in terms of genes, training and test samples) in inner and outer cross-validation loops to train all four NOPs as well as the baseline model which uses Lasso trained using all genes (n = 11748). Our results presented in Fig 4a, clearly demonstrate that compared to other NOPs under study, GL guided by SyNet achieves superior prognostic prediction for unseen patients selected from an independent cohort. To confirm that NOP performance using SyNet is the result of the network structure, we also applied the GL to a shuffled version of SyNet (Fig 4a). We observe a substantial deterioration of the AUC, supporting the conclusion that not only the genes, but also links contained in SyNet are important to achieve good prediction. Moreover, this observation rules out that the GL by itself is able to provide enhanced performance compared to standard Lasso. The result of a similar assessment for the Corr network is given in **S12**.

Additionally, we found that SyNet remains predictive even when the dataset is down sampled to 25% of samples (see **S13** for details). We also evaluated a recently developed set of subtype-specific networks for breast cancer [52] and found that SyNet markedly outperforms these networks in predictive performance (see **S18** for details).

We next assessed the performance gain of the network-guided model compared to a Lasso model that cannot exploit network information. To this end, the GL was trained based on each network whereas the Lasso is was trained based on the genes present in the network. Fig 4b demonstrates the results of this analysis. We find that the largest gain in GL performance is achieved when using SyNet (Fig 4b, x-axis), indicating that the links between genes in SyNet

truly aid classification performance beyond what is obtained as a result of the feature selection capabilities of Lasso.

Fig 4c provides the Kaplan-Meier plot when each patient is assigned to a good or poor prognostic class according to frequency of predicted prognosis across 10 repeats (ties are broken by random assignment to one of the classes) for Lasso as well as Group Lasso. Result of this analysis suggests that superior performance of the GL compared to the Lasso is mostly stemming from GLs ability to better discern the patients with poor prognosis.

An important property of an outcome predictor is to exhibit constant performance irrespective of the dataset used for training the model (i.e. performance stability). This is a highly desirable quality, as concerns have been raised regarding the highly variable performances of breast-cancer classifiers applied to different cohorts [7, 53]. To measure performance stability, we calculated the standard deviation of the AUC for Lasso and GL. The y-axis in Fig 4b represents the average difference of standard deviation for Lasso and GL across all evaluated folds and repeats (14 folds and 10 repeats). Based on this figure, we conclude that a NOP model guided by SyNet not only provides superior overall performance, it also offers improved stability of the classification performance.

Finally, we investigated the importance of hub genes in SyNet (genes with >4 neighbors) and observe that a comparable performance can be obtained with a network consisting of hub genes exclusively at the cost of reduced performance stability (see **S14** for details). Moreover, we did not observe performance gain for a model that is governed by combined links from multiple networks (either by intersection or unification, see **S15** for details). We further confirmed that the performance gain of the network-guided GL is preserved when networks are restricted to have equal number of links (see **S7** for details), or when links with lower confidence are included in the network (see **S16** for details). We also considered the more complex Sparse Group Lasso (SGL), which offers an additional level of regularization (see S1 Text for details). No substantial difference between GL and SGL performance was found (see **S8** for details). Likewise, we did not observe substantial performance differences when the number of genes, group size and regularization parameters were simultaneously optimized in a grid search (see **S9** for details). Together, these findings can be considered as the first unbiased evidence of true classification performance improvement in terms of average AUC and classification stability by a NOP.

Gene enrichment analysis for SyNet

Many curated biological networks suffer from an intrinsic bias since genes with well-known roles are the subject of more experiments and thus get more extensively and accurately annotated [54]. Post-hoc interpretation of the features used by NOPs, often by means of an enrichment analysis, will therefore be affected by the same bias. SyNet does not suffer from such bias, as its inference is purely data driven. Moreover, since SyNet is built based on gene pairs that contribute to the prediction of clinical outcome, we expect that the genes included in SyNet not only relate to breast cancer; they should play a role in determining how aggressively the tumor behaves, how advanced the disease is or how well it responds to treatment.

To investigate the relevance of genes contained in SyNet in the development of breast cancer and, more importantly, clinical outcome, we ranked all pairs according to their median Fitness (F_{ij}) across 14 studies and selected the top 300 genes (encompassing 3544 links). This cutoff was frequently chosen by the GL as the optimal number of genes in SyNet (see section "SyNet improves NOP performance"). Fig 5 visualizes this network revealing three main subnetworks and a few isolated gene pairs. We performed functional enrichment for all genes as

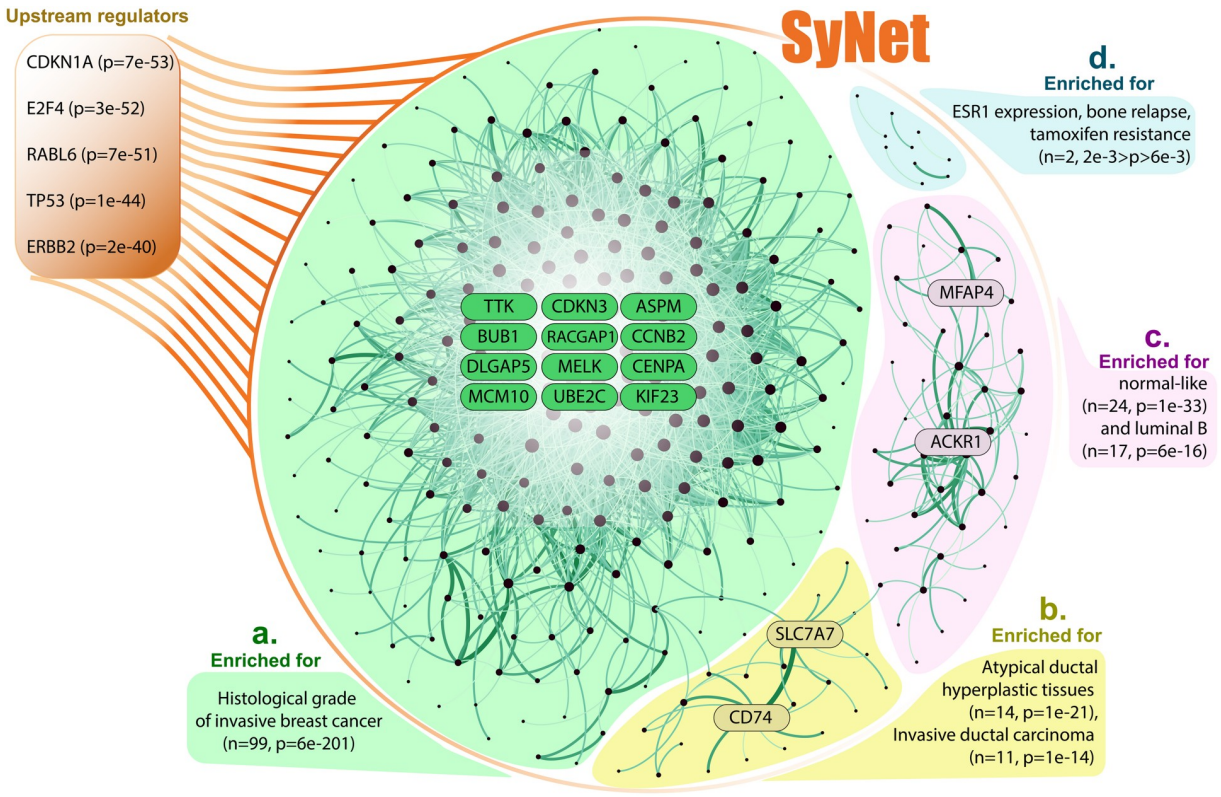

Fig 5. Visualization of SyNet. SyNet consists of three main subnetworks (**a**, **b** and **c**) and five separated gene pairs (**d**). Node size represents degree of node and link thickness indicates fitness of the corresponding pair. **a**. The largest subnetwork encompassing 223 genes is enriched for histologic grade of invasive breast cancer tumors. **b**. The second subnetwork is directly connected to the first cluster and contains risk factors for developing breast cancer. **c**. The third cluster is enriched for genes upregulated in normal-like subtype of breast cancer. **d**. Out of five pairs, only TFF3 and TFF1 pair is enriched for genes up-regulated in early primary breast tumors.

https://doi.org/10.1371/journal.pcbi.1006657.g005

well as for the subcomponents of the three large subnetworks in SyNet using Ingenuity Pathway Analysis (IPA) [55].

IPA reveals that out of 300 genes in SyNet, 287 genes have a known relation to cancer (2e-06<p<1e-34) of which 222 are related to reproductive system disease (2e-06<p<1e-34). Furthermore, according to IPA analysis, the top five upstream regulators of genes in SyNet (orange box, Fig 5) are CDKN1A, E2F4, RABL6, TP53 and ERBB2, all of which are well known players in the development of breast cancer[56–60]. The mean degree of the 300 genes in SyNet is 24, but there are 12 genes which have a degree of 100 or above: ASPM [61], BUB1 [62], CCNB2 [63], CDKN3 [64], CENPA [65], DLGAP5 [66], KIF23 [67], MCM10 [68], MELK [69], RACGAP1 [70], TTK [71] and UBE2C [72]. All these genes play a vital role in progression through the cell cycle and mitosis, by ensuring proper DNA replication, correct formation of the mitotic spindle and proper attachment to the centromere.

In addition to a clear involvement of genes linked to breast cancer generically, IPA also finds clear indications that the genes in SyNet are relevant to clinical outcome and prognosis of the disease. For instance, the most highly enriched cluster (Fig 5; green cluster) is found by IPA to be associated to histological grade of the tumor (p = 6e-201). The histological grade, which is based on the morphological characteristics of the tumor, has been shown to be informative for the clinical behavior of the tumor and is one of the best-established prognostic markers [73]. Notably, the blue cluster is enriched for genes involved in tamoxifen resistance (p<2e-3), one of the important treatments of ER-positive breast cancer.

Two other sub-clusters (yellow and purple in Fig 5), contain genes from distinctly different biological processes than the main cluster. In these clusters we also observe clear hub genes: SLC7A7 and CD74 in the yellow and ACKR1 and MFAP4 in the purple cluster. ACKR1 is a chemokine receptor involved in the regulation of the bio-availability of chemokine levels and MFAP4 is involved in regulating cell-cell adhesion. The recruitment of cells, as regulated by chemokines, and reducing cell-cell adhesion both play an important role in the process of metastasis. CD74 has also been linked to metastasis in triple negative breast cancer [74]. Metastasis, and not the primary tumor, is the main cause of death in breast cancer [3].

IPA highly significantly identifies the SyNet genes as upstream regulators of canonical pathways implicated in breast cancer (Fig 5), such as Cell Cycle Control of Chromosomal Replication (8e-18), Mitotic Roles of Polo-Like Kinase (4e-15), Role of CHK Proteins in Cell Cycle Checkpoint Control (6e-12), Estrogen-mediated S-phase Entry (2e-11), and Cell Cycle: G2/M DNA Damage Checkpoint Regulation (5e-10). Although all cancer cells deregulate cell cycle control, the degree of dysregulation may contribute to a more aggressive phenotype. For instance, it is recognized that the downregulation of certain checkpoint regulators is related to a worse prognosis in breast cancer[75, 76].

In summary, SyNet predominantly appears to contain genes relevant to two main processes in the progression of breast cancer: increased cell proliferation and the process of metastasis. Although many genes have not previously been specifically linked to breast cancer prognosis, their role in regulating different stages of replication and mitosis points to a genuine biological role in the progression and prognosis of breast cancer.

Similarity of SyNet to existing biological networks

We next sought to investigate the similarity between SyNet and existing biological networks that directly or indirectly capture biological interactions. To enable a comparison with networks of different sizes, we compare the observed overlap (both in terms of genes as well as links) to the distribution of expected overlap obtained by shuffling each network 1000 times (while keeping the degree distribution intact). Overlap is determined for varying network sizes by thresholding the link weights such that a certain percentage of genes or links remains. Results are reported in terms of a z-score in Fig 6.

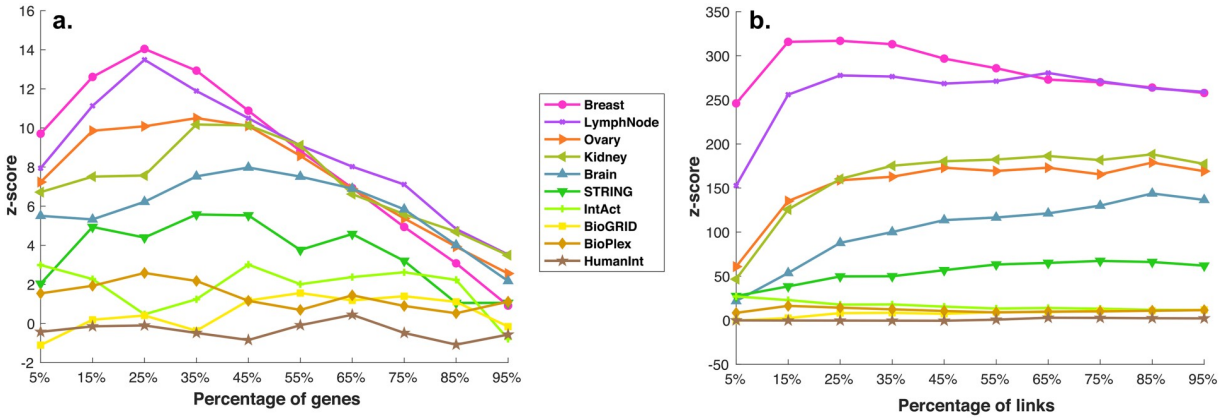

Fig 6. Similarity of existing biological networks to SyNet in terms of genes (a.) and links (b.). The x-axis represents the percentage of top gene/links used, the y-axis the z-score of observed vs. expected number of gene/links. The z-score is calculated by relating the observed number of SyNet gene/links that are present in existing biological networks to the expected distribution. To calculate the expected distribution, genes in biological networks are shuffled.

https://doi.org/10.1371/journal.pcbi.1006657.g006

Fig 6a shows that for the majority of networks a significantly higher than expected number of SyNet genes is contained in the top of each network. The overlap is especially pronounced for the tissue-specific networks, in particular the Breast-specific and Lymph node-specific networks, supporting our observation that SyNet contains links that are relevant for breast cancer. The enrichment becomes even more significant when considering the overlap between the links (Fig 6b). In this respect, SyNet is also clearly most similar to the Breast-specific and Lymph node-specific networks. We confirmed that these enrichments are not only driven by the correlation component of SyNet by repeating this analysis with a variant of the SyNet network without the correlation component (i.e. only average and synergy of gene pairs are used for pair-ranking; see **S10** for details). It should moreover be noted that, although a highly significant overlap is observed, the vast majority of SyNet genes and links are not present in the existing networks, explaining the improved performance obtained with NOPs using SyNet. Specifically, out of the 300 genes in SyNet, only 142 are contained within the top 25% of genes (n = 1005) in the Breast-specific network, and 151 in the top 25% of genes (n = 1290) in the Lymph node-specific network. Similarly, out of the 3544 links in SyNet, only 1182 are contained within the top 25% of links (n = 12500) in the Breast-specific network, and 617 in the top 25% (n = 12500) of the Lymph node-specific network (see **S11** for details). We further confirmed that the overall trend in observed overlaps between SyNet and other networks does not change when the size of these networks (in terms of the number of links) are increased or reduced (see **S17** for details).

Higher order structural similarity of SyNet and existing biological networks

In addition to direct overlap, we also aimed to investigate if genes directly connected in SyNet may be indirectly connected in existing networks. To assess this for each pair of genes in SyNet, we computed several topological measures characterizing their (indirect) connection in the biological networks. We included degree (Fig 7a), shortest path (Fig 7b) and Jaccard (Fig 7c) (see S1 Text for details). To produce an edge measure from degree and page rank (which are node based), we computed the average degree and page rank of genes in a pair respectively. Furthermore, we produced an expected distribution for each pair by computing the same topological measures for one of the genes and another randomly selected gene. The results from this analysis supports our previous observation that the information contained in the links of SyNet is markedly—yet only partially—overlapping with the information in the existing networks. Notably, the similarity increases for networks of increased relevance to the tissue in which the gene expression data is measured (i.e. breast tissue).

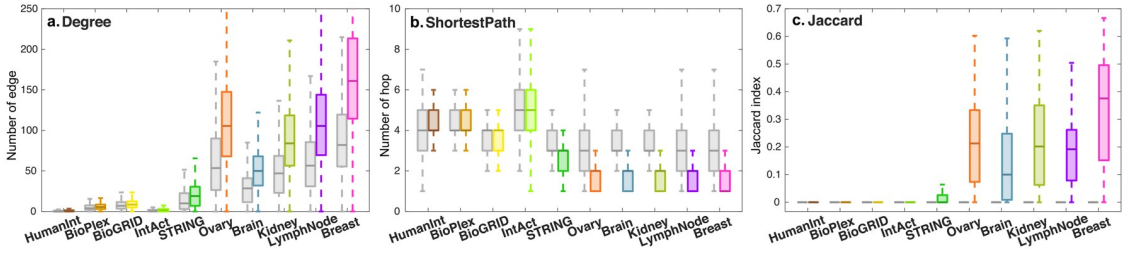

Fig 7. Comparison of three topological measures calculated over biological networks a. degree, b. shortest path and c. Jaccard index. Each color represents a network. Gray boxes represent the same topological measures calculated on the shuffled network.

https://doi.org/10.1371/journal.pcbi.1006657.g007

Predicting SyNet links from biological networks

Encouraged by the overlap with existing biological networks, we next asked whether links in SyNet can be predicted from the complete collection of topological measures calculated based on existing networks. To this end, we characterized each gene-pair by a set of 12 graph-topological measures that describe local and global network structure around each gene-pair. In addition to the degree, shortest path and Jaccard, we included several additional graph-topological measures including direct link, page rank (with four betas), closeness centrality, clustering coefficient and eigenvector centrality (see S1 Text for details). While converting node-based measures to edge based measures, in addition to using the average, we also used the difference between the score for each gene in the pair, similar to our previous work [77]. We applied these measures to all 10 networks in our collection yielding a total of 210 features. The gene-pairs are labeled according to their presence or absence in SyNet. Inspection of this dataset using the t-SNE [78] reveals that the links in SyNet occupy a distinct part of the 2D embedding obtained (Fig 8a).

We trained a Lasso and assessed classification performance in a 50-fold cross validation scheme where in each fold 1/50 of pairs in SyNet is kept hidden and the rest of pairs is utilized to train the classifier. To avoid information leakage in this assessment, we removed gene pairs from the training set in case one of the genes is present in the test set. Based on this analysis we find that a simple linear classifier can reach ~85% accuracy in predicting the synergistic gene relationships from SyNet (Fig 8b, rightmost bar). The contribution from generic networks is notably smaller than for the tissue-specific networks. In particular the networks relevant to breast cancer are highly informative, to the extent that combining multiple networks no longer improves prediction performance. Further investigation of feature importance revealed that the page rank topological measure was commonly used as a predictive marker across folds. Apparently, while direct overlap between SyNet and existing networks is modest, the topology of the relevant networks (i.e. breast-specific and lymph node-specific networks) are highly informative for the links contained in SyNet. This corroborates findings from Winter et al. in which the page rank topological measure was proposed to identify relevant genes in outcome prediction [34, 35, 79].

Discussion

Although the principle of using existing knowledge of the cellular wiring diagram to improve performance, robustness and interpretability of gene expression classifiers appears attractive, contrasting reports on the efficacy of such approach have appeared in literature [21, 28, 35]. Consensus in this field has particularly been frustrated by an evaluation of a limited set of sub-optimal classifiers [21, 23, 28, 35], small sample size [18, 24, 26], or the use of standard K-fold cross-validation instead of cross-study evaluation schemes, which results in inflated performance estimates [24, 26]. For this reason, it remained unclear if network-based classification, and in particular network-based outcome prediction, is beneficial. Here, we present a rigorously cross-validated procedure to train and evaluate Group Lasso-based NOPs using a variety of networks, including tissue-specific networks in particular, which have not been evaluated in the context of NOPs before.

Based on our analyses, we conclude that none of the existing networks achieve improved performance compared to using properly regularized classifiers trained on all genes. In this work we therefore present a novel gene network, called SyNet, which is computationally derived directly from the survival-labeled samples. The links in SyNet connect synergistic gene pairs. We followed a cross-validation procedure in which the inference of SyNet and validation of its utility in a NOP is strictly separated. We find that SyNet-based NOPs yields superior

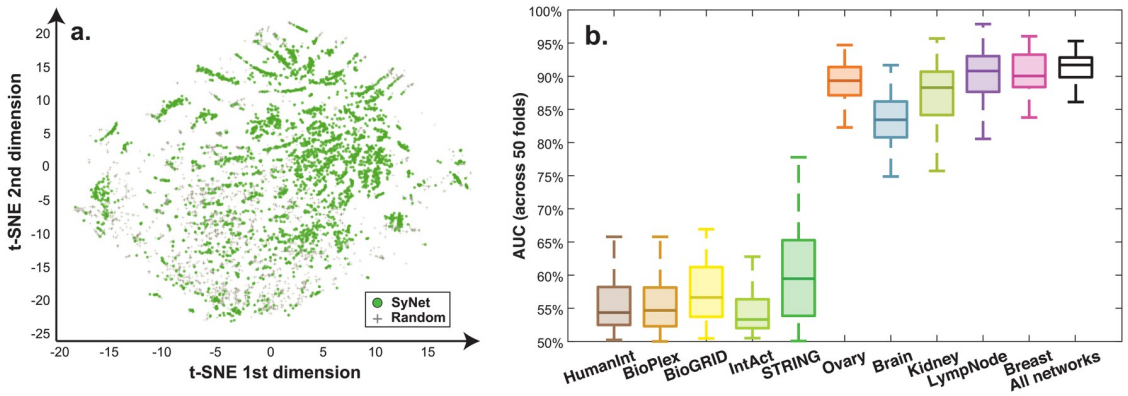

Fig 8. Characterizing SyNet links by a range of graph topological measures. a. t-SNE (unsupervised) visualization of the combined 180 topological measures. Each dot represents one gene pair. Green dots indicate SyNet links while gray markers represent an equal number of random pairs. **b**. Performance of Lasso model trained over all topological measure for different networks and all networks combined (rightmost bar).

https://doi.org/10.1371/journal.pcbi.1006657.g008

performance with higher stability across the folds compared to both the baseline model trained on all genes as well as models that use other existing gene networks. We therefore conclude that at least in outcome prediction problem, network guidance can improve model performance, but only if this network is phenotype-specific. Supporting this conclusion, we also show that a correlation network, which is dataset-specific but not phenotype specific, also improved performance but much less compared to SyNet.

A major benefit of SyNet over manually curated gene networks is that its inference is purely data driven, and therefore not biased to well-studied genes. Post-hoc interpretation of the genes selected by a NOP that utilized SyNet is therefore expected to provide a more unbiased interpretation of the important molecular players underlying breast cancer and patient survival. Analysis of the genes contained in SyNet shows strong enrichment for genes with known relevance to breast cancer. More importantly, the largest subcomponent of SyNet is strongly linked to patient prognosis as it includes many genes with a known relation to the histological grade of the tumor.

To investigate if SyNet captures known biological gene interactions, we extensively compared SyNet with existing networks. We find highly significant overlaps between links, indicating that SyNet connects genes that also have a known biological interaction. Despite this significant overlap, the majority of the SyNet links are not recapitulated by direct links in the existing networks. However, we find that accurate prediction of links in SyNet are possible if more complex graph topological descriptions of the indirect connections in the existing networks are employed. Interestingly, accurate predictions are only obtained when using the breast specific networks. Apparently, although the information contained in SyNet is similar to other gene interaction networks, the wiring of SyNet much better supports GL-based classification. This might explain why using existing biological networks in NOPs directly is unsuccessful and why graph topological measures have been successful in identifying relevant genes in outcome prediction [34, 35, 79].

Taken together, our results underline that network-based outcome prediction is a promising approach to improving patient prognosis prediction and therefore can provide an important contribution towards more personalized healthcare. At the same time, the SyNet approach provides an unbiased interactome which makes the NOP more amenable for model interpretation, thus providing important insights into the etiology of the disease under study.

Materials and methods

Inferring a synergistic network (SyNet)

We hypothesized that, in order to improve outcome prediction by network-based classification, interconnections in the network should correspond to gene pairs for which integration yields a performance beyond what is attainable by either of the individual genes (i.e. synergy). Accordingly, we formulated the synergy S_{ij} between gene *i* and gene *j* as

$$S_{ij} = \frac{A_{ij}}{Max(A_i, A_j)}$$

where A_i, A_j and A_{ij} respectively represent the Area Under Curve (AUC) of gene *i*, the AUC of gene *j* and the AUC of meta-gene *ij* formed by aggregation of gene *i* and gene *j*. Meta-gene formation is carried out by a linear regression model which demonstrated superior performance in our experiments (see **S1** for details). Cross-validation performance of the linear regression (see section "Cross validation design" for details) is obtained and the median of 65 AUCs (13 folds and 5 repeats) is used as the final score A_{ij} for each pair. The AUC of the individual genes (i.e. A_i and A_j) is obtained in a similar fashion.

Defining the synergy as a function of AUC yields a phenotype-specific (i.e. label-specific) measure which effectively ignores extraneous relationships between gene pairs that are not relevant in outcome prediction. The synergy measure S_{ij} depends on the performance of individual genes where poorly performing genes tend to achieve higher degree of synergy compared to two predictive genes (see **S2** for corresponding analysis). In order to account for this effect, the average AUC of individual genes is included as a second criterion. Furthermore, our preliminary tests confirmed previous findings [8, 23, 80], that integrating highly correlated genes (which reduces meta-gene noise) may improve survival prediction. For this reason, we added correlation of pairs as a third criterion. To combine these three measures, each measure is normalized independently between [0, 1] and then combined into an overall fitness score F_{ij} for gene pair *ij*:

$$F_{ij} = -\sqrt{(1-\overline{S_{ij}})^2 + (1-\overline{M_{ij}})^2 + (1-\overline{C_{ij}})^2}$$

Here, M_{ij} and C_{ij} represent mean AUC and absolute spearman correlation of gene *i* and *j* respectively. Bars above letters indicate that the corresponding values are normalized to the [0, 1] interval. Employing the Dutch grid infrastructure, we quantified the fitness for all 69 million possible pairs of genes (n = 11748). Fig 1c visualizes the fitness of all pairs in a three-dimensional space. Finally, the top 50,000 pairs with highest fitness are considered as SyNet.

Expression data

Accurately estimating survival risk and identifying markers relevant for progression of a complex disease such as breast cancer requires a large number of samples [11]. To this end, samples from METABRIC [81] (n = 1981) are combined with 12 studies collected in ACES [21] (n = 1606) as well as samples from the TCGA breast invasive carcinoma dataset [82] (n = 532) (see S1 Text for details). Collectively, these datasets, spanning 14 distinct studies, form a compendium encompassing 4129 samples. To the best of our knowledge, the data combined in this paper represents the largest breast cancer gene expression compendium to date. As a result, our compendium should capture a large portion of the biological heterogeneity among breast cancer patients, as well as technical biases originating from the variability in platforms and study-specific sample preparations [83]. This variability will assist the trained models to

achieve better generalization which is crucial in real world application of the final classification model [9, 13, 84]. To correct for technical variations that may arise during the library preparation, initially the expression data within each study is quantile normalized and then batch-effect corrected using Combat [85] where the outcome of patients was modeled as an additional covariate to maintain the variance associated with the prognostics. This procedure was shown to perform well among many batch effect removal methods [86, 87]. Successful removal of batch effects was confirmed using t-SNE visualization [78] (See **S4** for details). The label for each patient corresponds to overall survival time (or recurrence free survival if available) with respect to a 5-year threshold (good vs. poor outcome).

Regular classifiers and network based prediction models

Ascertaining the relevance of networks in outcome prediction should be performed using a robust predictor capable of providing adequate performance in prognostic prediction. Previous assessments in this regard have been limited to only few classifiers [21, 23, 28, 35]. To identify the optimal predictor, we have compared performance of wide range of linear and nonlinear classifiers (see **S3** for details). Supporting our previous findings [8], this evaluation demonstrates that simple linear classifiers outperform the more complex ones, with the regularized linear classifier (Lasso) reaching the highest AUC. This classifier supports both classical and well as network-based prediction by its derivative called Group Lasso (GL) [88]. The GL is structurally analogous to standard Lasso with the exception of the way in which the regularization is performed; Lasso applies regularization to genes while GL enforces selection of groups of genes (See S1 Text for details). In order to incorporate network information in the GL, similar to our previous work [8], each gene in the corresponding network is considered as seed gene and together with its K neighbors the group structure provided to the GL. Priority of neighbor selection is determined by edge weights between each neighbor and corresponding seed gene. The hyperparameters for each classifier (e.g. K in the GL) are determined by means of a grid search in the inner cross validation loop (see **S5** for schematic overview).

For comparison, we include three well-known NOPs in our analysis. Park et al. utilized hierarchical clustering to group highly correlated genes [23]. Each group is summarized into a meta-gene by averaging the expression profile of the genes in that group. These meta-genes are then employed as regular features to train a Lasso classifier. The optimal cluster size for hierarchical clustering is identified by iterative application of Lasso in an inner cross-validation. Chuang et al. employs a greedy search to define subnetworks [18]. This is done by iteratively expanding a sub-network initiated from a seed gene guided by a supervised performance criterion which halts when performance no longer increases (in the training set). After groups are formed, the meta-genes are constructed by averaging expression of each gene within each group similar to Park et al. Finally, Taylor et al. focus on hubs (i.e. highly connected genes, degree$>$5) in a network [24]. To identify dysregulated subnetworks, the change in correlation between each hub and its direct neighbors across two classes of outcome (poor vs. good) is assessed. Meta-genes are formed from candidate subnetworks similar to the procedure employed by Park et al.

Networks

In addition to SyNet, we considered a range of publicly available networks, including generic networks (HumanInt, BioPlex, BioGRID, IntAct and STRING) as well as a correlation network (Corr) which was previously shown to be an effective network in outcome prediction [8, 23].

Additionally, we assessed five tissue-specific networks (including brain, kidney, ovary, breast, lymph node) that are recently introduced by Greene et al. [44]. These tissue-specific networks are inferred by integrating protein-protein interactions collected from Human Protein Reference Database [89] and tissue-specific information from BRENDA tissue ontology [90] and then filtered using expert-selected Gene Ontology (GO) terms. The tissue-specificity of each network is then validated by a comprehensive collection of expression and interaction datasets encompassing about 38000 conditions collected from approximately 14000 publications. To the best of our knowledge, our study is the first to evaluate tissue-specific networks in the context of NOPs.

To maintain a reasonable network size, we utilized only the top 50,000 links (based on the link weight) in each network (similar to number of links in SyNet). For the only unweighted network, HumanInt [38], all interactions (n = ~14k) were included and links were weighted according to the average degree of the two interacting genes. Moreover, a randomized version of each network is constructed by shuffling nodes in the network which destroys the biological information of the links while preserving the overall network structure (see S1 Text for full details on preparation of networks).

Cross validation design

In order to ascertain if network information truly aids outcome prediction, the evaluation should be based on a rigorous cross-validation that closely resembles the real-world application of these models. To this end, we perform cross-study validation in order to mimic a realistic situation in which a classifier is applied to data from a different hospital than it was trained on [7]. Briefly, one study is taken out for validation of the final performance (outer loop test set). SyNet inference and NOP training are carried out on the 13 remaining studies (outer loop training set). Within each fold of the outer loop training set, again one study is left out to obtain the inner loop test set and the rest of studies for inner loop training set. The inner loop training set is sub sampled (with replacement) to 70% and regression is performed for every gene as well as gene pairs (identical set of samples are used across all genes and pairs). The AUC scores (A_i, A_j and A_{ij}) are calculated on the inner loop test set. This is repeated 5 times. To train a NOP for this fold, a new inner loop training set is formed by redrawing 70% of the samples from the outer loop training set. This set is also used to infer correlation network. To assess the final performance of the NOP the outer loop test set is used (see **S5** for a detailed schematic). Our initial experiments showed a large variation of performance across studies (see **S6** for details). To prevent this variation from influencing our comparisons, assignment of samples to folds in both inner and outer cross-validation loops are kept identical across all comparisons throughout the paper. We used Area Under the ROC Curve (AUC) as the main measure of performance in this paper.

Supporting information

S1 Text. This section provides further details about analyses performed in this paper.
(DOCX)

S1 Table. GEO accession and number of samples per study in ACES.
(XLSX)

S2 Table. Additional clinical variables of patients used in this study. In order of appearance this variables include: survival time, prognostic status (used in this study as sample label), age, breast cancer subtype, ER status, Her2 status, PGR status, histological grade, lymph node status, tumor size, microarray measurement platform, relapse-free survival event, relapse-free

survival time, distant metastasis free survival event, distant metastasis free survival time, overall survival event, overall survival time, patient id in the study, treatment, study source of the sample and study index.
(XLSX)

Author Contributions

Conceptualization: Amin Allahyar, Joske Ubels, Jeroen de Ridder.

Data curation: Amin Allahyar.

Formal analysis: Amin Allahyar.

Funding acquisition: Jeroen de Ridder.

Investigation: Amin Allahyar, Joske Ubels, Jeroen de Ridder.

Methodology: Amin Allahyar, Joske Ubels, Jeroen de Ridder.

Project administration: Jeroen de Ridder.

Software: Amin Allahyar.

Supervision: Jeroen de Ridder.

Validation: Amin Allahyar, Joske Ubels.

Visualization: Amin Allahyar, Joske Ubels.

Writing – original draft: Amin Allahyar, Joske Ubels, Jeroen de Ridder.

Writing – review & editing: Amin Allahyar, Joske Ubels, Jeroen de Ridder.

References

1. Fantozzi A. and Christofori G., Mouse models of breast cancer metastasis. Breast Cancer Research, 2006. 8(4): p. 212. https://doi.org/10.1186/bcr1530 PMID: 16887003
2. Shapiro C.L. and Recht A., *Side Effects of Adjuvant Treatment of Breast Cancer*. 2001. 344(26): p. 1997–2008.
3. Weigelt B., Peterse J.L., and van't Veer L.J., Breast cancer metastasis: markers and models. Nature Reviews Cancer, 2005. 5: p. 591. https://doi.org/10.1038/nrc1670 PMID: 16056258
4. Cardoso F., et al., *70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer*. 2016. 375(8): p. 717–729.
5. van't Veer L.J., et al., Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002. 415: p. 530. https://doi.org/10.1038/415530a PMID: 11823860
6. van de Vijver M.J., et al., *A Gene-Expression Signature as a Predictor of Survival in Breast Cancer*. 2002. 347(25): p. 1999–2009.
7. Bernau C., et al., Cross-study validation for the assessment of prediction algorithms. Bioinformatics, 2014. 30(12): p. i105–i112. https://doi.org/10.1093/bioinformatics/btu279 PMID: 24931973
8. Allahyar A. and de Ridder J., FERAL: network-based classifier with application to breast cancer outcome prediction. Bioinformatics, 2015. 31(12): p. i311–i319. https://doi.org/10.1093/bioinformatics/btv255 PMID: 26072498
9. Ransohoff D.F., Bias as a threat to the validity of cancer molecular-marker research. Nature Reviews Cancer, 2005. 5: p. 142. https://doi.org/10.1038/nrc1550 PMID: 15685197
10. Venet D., Dumont J.E., and Detours V., Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLOS Computational Biology, 2011. 7(10): p. e1002240. https://doi.org/10.1371/journal.pcbi.1002240 PMID: 22028643
11. Ein-Dor L., Zuk O., and Domany E., *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer*. 2006. 103(15): p. 5923–5928.

12. Akavia U.D., et al., An Integrated Approach to Uncover Drivers of Cancer. Cell, 2010. 143(6): p. 1005–1017. https://doi.org/10.1016/j.cell.2010.11.013 PMID: 21129771
13. Stretch C., et al., Effects of Sample Size on Differential Gene Expression, Rank Order and Prediction Accuracy of a Gene Signature. PLOS ONE, 2013. 8(6): p. e65380. https://doi.org/10.1371/journal.pone.0065380 PMID: 23755224
14. Hua J., Tembe W.D., and Dougherty E.R., Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition, 2009. 42(3): p. 409–424.
15. Bryant P.A., et al., Technical Variability Is Greater than Biological Variability in a Microarray Experiment but Both Are Outweighed by Changes Induced by Stimulation. PLOS ONE, 2011. 6(5): p. e19556. https://doi.org/10.1371/journal.pone.0019556 PMID: 21655321
16. Parker Hilary S. and Leek Jeffrey T., The practical effect of batch on genomic prediction, in *Statistical Applications in Genetics and Molecular Biology*. 2012.
17. Alcaraz N., et al., De novo pathway-based biomarker identification. Nucleic Acids Research, 2017. 45 (16): p. e151–e151. https://doi.org/10.1093/nar/gkx642 PMID: 28934488
18. Chuang H.-Y., et al., *Network-based classification of breast cancer metastasis*. 2007. 3(1): p. 140.
19. Hanahan D. and Weinberg Robert A., Hallmarks of Cancer: The Next Generation. Cell, 2011. 144(5): p. 646–674. https://doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230
20. Hanahan D. and Weinberg R.A., The Hallmarks of Cancer. Cell, 2000. 100(1): p. 57–70. PMID: 10647931
21. Staiger C., et al., *Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis*. 2013. 4(289).
22. Cheng W.-Y., Yang T.-H.O., and Anastassiou D., Biomolecular Events in Cancer Revealed by Attractor Metagenes. PLOS Computational Biology, 2013. 9(2): p. e1002920. https://doi.org/10.1371/journal.pcbi.1002920 PMID: 23468608
23. Park M.Y., Hastie T., and Tibshirani R., Averaged gene expressions for regression. Biostatistics, 2007. 8(2): p. 212–227. https://doi.org/10.1093/biostatistics/kxl002 PMID: 16698769
24. Taylor I.W., et al., Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nature Biotechnology, 2009. 27: p. 199. https://doi.org/10.1038/nbt.1522 PMID: 19182785
25. Zhang W., et al., Network-based machine learning and graph theory algorithms for precision oncology. npj Precision Oncology, 2017. 1(1): p. 25. https://doi.org/10.1038/s41698-017-0029-7 PMID: 29872707
26. Popovici V., et al., *Effect of training-sample size and classification difficulty on the accuracy of genomic predictors*. 2010. 12(1): p. R5.
27. Wessels L.F.A., et al., A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics, 2005. 21(19): p. 3755–3762. https://doi.org/10.1093/bioinformatics/bti429 PMID: 15817694
28. Roy J., Winter C., and Schroeder M., *Meta-analysis of Cancer Gene Profiling Data, in Cancer Gene Profiling: Methods and Protocols*, Grützmann R. and Pilarsky C., Editors. 2016, Springer: New York, NY. p. 211–222.
29. Dutkowski J. and Ideker T., Protein Networks as Logic Functions in Development and Cancer. PLOS Computational Biology, 2011. 7(9): p. e1002180. https://doi.org/10.1371/journal.pcbi.1002180 PMID: 21980275
30. Wang E., et al., Predictive genomics: A cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. Seminars in Cancer Biology, 2015. 30: p. 4–12. https://doi.org/10.1016/j.semcancer.2014.04.002 PMID: 24747696
31. Cun Y. and Fröhlich H., Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics. PLOS ONE, 2013. 8(9): p. e73074. https://doi.org/10.1371/journal.pone.0073074 PMID: 24019896
32. Staiger C., et al., A Critical Evaluation of Network and Pathway-Based Classifiers for Outcome Prediction in Breast Cancer. PLOS ONE, 2012. 7(4): p. e34796. https://doi.org/10.1371/journal.pone.0034796 PMID: 22558100
33. Alpaydin E., *Introduction to Machine Learning*. 2004: MIT Press.
34. Winter C., et al., Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes. PLOS Computational Biology, 2012. 8(5): p. e1002511. https://doi.org/10.1371/journal.pcbi.1002511 PMID: 22615549
35. Roy J., et al., Network information improves cancer outcome prediction. Briefings in Bioinformatics, 2014. 15(4): p. 612–625. https://doi.org/10.1093/bib/bbs083 PMID: 23255167
36. Cusick M.E., et al., Literature-curated protein interaction datasets. Nature Methods, 2008. 6: p. 39.

37. von Mering C., et al., Comparative assessment of large-scale data sets of protein–protein interactions. Nature, 2002. 417: p. 399. https://doi.org/10.1038/nature750 PMID: 12000970

38. Rual J.-F., et al., Towards a proteome-scale map of the human protein–protein interaction network. Nature, 2005. 437: p. 1173. https://doi.org/10.1038/nature04209 PMID: 16189514

39. Mahdavi M.A. and Lin Y.-H.J.B.B., *False positive reduction in protein-protein interaction predictions using gene ontology annotations*. 2007. 8(1): p. 262.

40. Rolland T., et al., A Proteome-Scale Map of the Human Interactome Network. Cell, 2014. 159(5): p. 1212–1226. https://doi.org/10.1016/j.cell.2014.10.050 PMID: 25416956

41. Huttlin E.L., et al., The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell, 2015. 162(2): p. 425–440. https://doi.org/10.1016/j.cell.2015.06.043 PMID: 26186194

42. Huttlin E.L., et al., Architecture of the human interactome defines protein communities and disease networks. Nature, 2017. 545: p. 505. https://doi.org/10.1038/nature22366 PMID: 28514442

43. Greene C.S. and Troyanskaya O.G., Chapter 2: Data-Driven View of Disease Biology. PLOS Computational Biology, 2012. 8(12): p. e1002816. https://doi.org/10.1371/journal.pcbi.1002816 PMID: 23300408

44. Greene C.S., et al., Understanding multicellular function and disease with human tissue-specific networks. Nature Genetics, 2015. 47: p. 569. https://doi.org/10.1038/ng.3259 PMID: 25915600

45. Yeger-Lotem E. and Sharan R., *Human protein interaction networks across tissues and diseases*. 2015. 6(257).

46. Zhang S., et al., *Discovering functions and revealing mechanisms at molecular level from biological networks*. 2007. 7(16): p. 2856–2869.

47. Kotlyar M., et al., Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Research, 2016. 44(D1): p. D536–D541. https://doi.org/10.1093/nar/gkv1115 PMID: 26516188

48. de Anda-Jáuregui G., et al., *Transcriptional Network Architecture of Breast Cancer Molecular Subtypes*. 2016. 7(568).

49. Watkinson J., et al., *Identification of gene interactions associated with disease from gene expression data using synergy networks*. 2008. 2(1): p. 10.

50. Ambroise C. and McLachlan G.J., *Selection bias in gene extraction on the basis of microarray gene-expression data*. 2002. 99(10): p. 6562–6566.

51. Boyle E.A., Li Y.I., and Pritchard J.K., An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell, 2017. 169(7): p. 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038 PMID: 28622505

52. Zaman N., et al., Signaling Network Assessment of Mutations and Copy Number Variations Predict Breast Cancer Subtype-Specific Drug Targets. Cell Reports, 2013. 5(1): p. 216–223. https://doi.org/10.1016/j.celrep.2013.08.028 PMID: 24075989

53. Castaldi P.J., Dahabreh I.J., and Ioannidis J.P.A., An empirical assessment of validation practices for molecular classifiers. Briefings in Bioinformatics, 2011. 12(3): p. 189–202. https://doi.org/10.1093/bib/bbq073 PMID: 21300697

54. Khatri P., Sirota M., and Butte A.J., Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLOS Computational Biology, 2012. 8(2): p. e1002375. https://doi.org/10.1371/journal.pcbi.1002375 PMID: 22383865

55. Krämer A., et al., Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics, 2014. 30(4): p. 523–530. https://doi.org/10.1093/bioinformatics/btt703 PMID: 24336805

56. Khaleel S.S., et al., *E2F4 regulatory program predicts patient survival prognosis in breast cancer*. 2014. 16(6): p. 486.

57. Gasco M., Shami S., and Crook T.J.B.C.R., *The p53 pathway in breast cancer*. 2002. 4(2): p. 70.

58. Wei C.-Y., et al., *Expression of CDKN1A/p21 and TGFBR2 in breast cancer and their prognostic significance*. 2015. 8(11): p. 14619.

59. Tan M. and Yu D., Molecular Mechanisms of ErbB2-Mediated Breast Cancer Chemoresistance, in *Breast Cancer Chemosensitivity*, Yu D. and Hung M.-C., Editors. 2007, Springer: New York, NY. p. 119–129.

60. Montalbano J., et al., *RBEL1 Is a Novel Gene That Encodes a Nucleocytoplasmic Ras Superfamily GTP-binding Protein and Is Overexpressed in Breast Cancer*. 2007. 282(52): p. 37640–37649.

61. Fish J.L., et al., *Aspm specifically maintains symmetric proliferative divisions of neuroepithelial cells*. 2006. 103(27): p. 10438–10443.

62. Skoufias D.A., et al., *Mammalian mad2 and bub1/bubR1 recognize distinct spindle-attachment and kinetochore-tension checkpoints*. 2001. 98(8): p. 4492–4497.

63. Draetta G., et al., cdc2 protein kinase is complexed with both cyclin A and B: Evidence for proteolytic inactivation of MPF. Cell, 1989. 56(5): p. 829–838. PMID: 2538242

64. Nalepa G., et al., *The tumor suppressor CDKN3 controls mitosis*. 2013. 201(7): p. 997–1012.

65. Foltz D.R., et al., The human CENP-A centromeric nucleosome-associated complex. Nature Cell Biology, 2006. 8: p. 458. https://doi.org/10.1038/ncb1397 PMID: 16622419

66. Tsou A.-P., et al., Identification of a novel cell cycle regulated gene, HURP, overexpressed in human hepatocellular carcinoma. Oncogene, 2003. 22: p. 298. https://doi.org/10.1038/sj.onc.1206129 PMID: 12527899

67. Pavicic-Kaltenbrunner V., et al., *Cooperative Assembly of CYK-4/MgcRacGAP and ZEN-4/MKLP1 to Form the Centralspindlin Complex*. 2007. 18(12): p. 4992–5003.

68. Ricke R.M. and Bielinsky A.-K., Mcm10 Regulates the Stability and Chromatin Association of DNA Polymerase-&#x3b1. Molecular Cell, 2004. 16(2): p. 173–185. https://doi.org/10.1016/j.molcel.2004.09.017 PMID: 15494305

69. Nakano I., et al., *Maternal embryonic leucine zipper kinase (MELK) regulates multipotent neural progenitor proliferation*. 2005. 170(3): p. 413–427.

70. Lee K.-Y., et al., Direct interaction between centralspindlin and PRC1 reinforces mechanical resilience of the central spindle. Nature Communications, 2015. 6: p. 7290. https://doi.org/10.1038/ncomms8290 PMID: 26088160

71. Fisk H.A., Mattison C.P., and Winey M., *Human Mps1 protein kinase is required for centrosome duplication and normal mitotic progression*. 2003. 100(25): p. 14875–14880.

72. Hao Z., Zhang H., and Cowell J.J.T.B., *Ubiquitin-conjugating enzyme UBE2C: molecular biology, role in tumorigenesis, and potential as a biomarker*. 2012. 33(3): p. 723–730.

73. Rakha E.A., et al., *Breast cancer prognostic classification in the molecular era: the role of histological grade*. 2010. 12(4): p. 207.

74. Greenwood C., et al., Stat1 and CD74 overexpression is co-dependent and linked to increased invasion and lymph node metastasis in triple-negative breast cancer. Journal of Proteomics, 2012. 75(10): p. 3031–3040. https://doi.org/10.1016/j.jprot.2011.11.033 PMID: 22178447

75. Catzavelos C., et al., Decreased levels of the cell-cycle inhibitor p27Kip1 protein: Prognostic implications in primary breast cancer. Nature Medicine, 1997. 3: p. 227. PMID: 9018244

76. Craig C., et al., A recombinant adenovirus expressing p27Kip1 induces cell cycle arrest and loss of cyclin-Cdk activity in human breast cancer cells. Oncogene, 1997. 14: p. 2283. https://doi.org/10.1038/sj.onc.1201064 PMID: 9178904

77. Hulsman M., Dimitrakopoulos C., and de Ridder J., Scale-space measures for graph topology link protein network architecture to function. Bioinformatics, 2014. 30(12): p. i237–i245. https://doi.org/10.1093/bioinformatics/btu283 PMID: 24931989

78. van der Maaten L. and Hinton G.J.M.L., *Visualizing non-metric similarities in multiple maps*. 2012. 87 (1): p. 33–55.

79. Newman M.E.J., Analysis of weighted networks. Physical Review E, 2004. 70(5): p. 056131.

80. Wu G. and Stein L., A network module-based method for identifying cancer prognostic signatures. Genome Biology, 2012. 13(12): p. R112. https://doi.org/10.1186/gb-2012-13-12-r112 PMID: 23228031

81. Curtis C., et al., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 2012. 486: p. 346. https://doi.org/10.1038/nature10983 PMID: 22522925

82. The Cancer Genome Atlas, N., et al., Comprehensive molecular portraits of human breast tumours. Nature, 2012. 490: p. 61. https://doi.org/10.1038/nature11412 PMID: 23000897

83. Allott E.H., et al., *Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification*. 2016. 18(1): p. 68.

84. Kourou K., et al., Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 2015. 13: p. 8–17. https://doi.org/10.1016/j.csbj.2014.11.005 PMID: 25750696

85. Johnson W.E., Li C., and Rabinovic A., Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 2007. 8(1): p. 118–127. https://doi.org/10.1093/biostatistics/kxj037 PMID: 16632515

86. Müller C., et al., Removing Batch Effects from Longitudinal Gene Expression—Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. PLOS ONE, 2016. 11(6): p. e0156594. https://doi.org/10.1371/journal.pone.0156594 PMID: 27272489

87. Chen C., et al., Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLOS ONE, 2011. 6(2): p. e17238. https://doi.org/10.1371/journal.pone.0017238 PMID: 21386892

88. Yuan M. and Lin Y., *Model selection and estimation in regression with grouped variables*. 2006. 68(1): p. 49–67.

89. Keshava Prasad T.S., et al., Human Protein Reference Database—2009 update. Nucleic Acids Research, 2009. 37(suppl_1): p. D767–D772.

90. Gremse M., et al., The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Research, 2011. 39(suppl_1): p. D507–D513.