Turning straw into gold : building robustness into gene signature inference

Goh, Wilson Wen Bin; Wong, Limsoon

2018

Goh, W. W. B., & Wong, L. (2019). Turning straw into gold : building robustness into gene signature inference. Drug Discovery Today, 24(1), 31-36. doi:10.1016/j.drudis.2018.08.002

https://hdl.handle.net/10356/137542

https://doi.org/10.1016/j.drudis.2018.08.002

Turning straw into gold: building robustness into gene signature inference

Wilson Wen Bin Goh [1*], Limsoon Wong [2,3*]

1. School of Biological Sciences, Nanyang Technological University, Singapore
2. Department of Computer Science, National University of Singapore, Singapore
3. Department of Pathology, National University of Singapore, Singapore

*Corresponding Author: Wilson Wen Bin Goh, wilsongoh@ntu.edu.sg; Limsoon Wong, wongls@comp.nus.edu.sg

Address for correspondence/proofs:

Wilson Wen Bin Goh, PhD

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

Email: wilsongoh@ntu.edu.sg, Tel: +65-65162902

Limsoon Wong, PhD

Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417

Email: wongls@comp.nus.edu.sg, Tel: +65-65162902

Teaser (35 words)

Naïve reliance on basic statistics leads to a lack of gene signature reproducibility and generalizability. To improve analytical outcome, we may leverage on existing knowledge (based on meta-analysis), systematically evaluating confounders, and performing generalizability tests.

Abstract (104 words)

Reproducible and generalizable gene signatures are essential for clinical deployment, but are hard to come by. The primary issue is insufficient mitigation of confounders: ensuring that hypotheses are appropriate, test statistics and null distributions are appropriate, and so on. To further improve robustness, additional good analytical practices (GAPs) are needed: 1/ leveraging on existing data and knowledge; 2/ careful and systematic evaluation of gene sets, even if they overlap with known sources of confounding; and 3/ rigorously testing inferred signatures against as many published datasets as possible. Using a re-examination of a breast cancer dataset and 48 published signatures, we illustrate the value of adopting these GAPs.

Keywords

Bioinformatics; Statistics; Feature selection; Biomarker; Prediction; Confounder

Introduction

Statistical feature selection on -omics data is a practical means of deriving signatures for predictive purposes. While the exact conditions for deriving a successful signature are not easily defined, it is known that statistical significance can arise for a variety of confounders (e.g. sampling bias, presence of hidden subpopulations, and batch effects), besides biological relevance [1]. This is known as the Anna Karenina Principle [2,3].

Naïve reliance on basic statistics therefore leads to a lack of signature reproducibility (getting a similar signature on a different dataset) [4-6] and signature generalizability (able to correctly predict phenotype on a different dataset) [7]. Addressing confounders are important but not necessarily practicable (assuming it is even possible to correctly identify every possible confounder). Some key points covered previously include developing more reasonable hypothesis statements and ensuring the correct test statistics and reference distributions are used [1]. Broadly, these constitute good analytical practices (GAPs) in the context of general analysis. But more robustness can be introduced for the purpose of signature inference. Using a re-examination of the dataset of Venet et al. [7], we illustrate here the following GAPs: 1/ the importance of meta-analysis, 2/ systematic evaluation of confounders, and 3/ generalizability tests.

The case study

In the study by Venet et al., they evaluated 48 published breast cancer signatures on an independent dataset [7]. A good signature is one that is associated significantly with outcome or phenotype. But in this study, it was found that most published signatures do not outperform randomly generated signatures, and even irrelevant signatures derived from other phenotypes do well. That is, statistical significance alone cannot prove relevance.

Suspected confounders include 1/ use of inappropriate null distribution, where large fractions of randomly generated signatures are significant under the nominal p-value of 0.05, far exceeding the expected 5%; 2/ the statistical tests do not account for the fact that cancer-associated genes are deeply confounded with the proliferation signature, of which many genes are part of; and 3/ inappropriate test-statistic which produces highly unreliable p-values: randomly generated signatures are used as null samples but it is unclear what the appropriate test-statistic should be. Although the nominal p-value of Cox's analysis is used as the test-statistic, this test-statistic is likely to exhibit large fluctuations on different sets of patients, which in turn causes large fluctuations in the corresponding p-value [1].

The importance of meta-analysis

Meta-analysis is the comparative evaluation of independent studies covering the same subject matter (e.g. breast cancer versus normal patients). In Venet et al., they evaluated 48 independently published breast cancer signatures against the NKI benchmark dataset (see Supplementary) [7], which revealed that these signatures are not only very different from each other, they also perform variably on the benchmark.

Each signature may be considered an independent sampling (with different degrees of error, leading to variable performance), and so an aggregate analysis is intuitively more informative than any single study. Venet et al.'s meta-analysis reveals that many of these signatures perform no better than randomly-generated signatures [7], suggesting that the composition of many published signatures are artifact-infested (presumably overladen with proliferation genes) (Supplementary Table 1). While it is standard practice to use cross-validation (at the minimum) in signature inference studies, it is clearly insufficient: given today's easy accessibility to data, it is inexcusable to perform signature inference as a single study without quantitative cross-references to other similar studies.

In Venet et al.'s example on breast cancer outcome, this creates an interesting opportunity: since the signatures vary widely in terms of gene composition and predictive performance, can a strong signature emerge based on the gene-composition intersection of the best-performing predictors (**see Supplementary Methods**) [7], and thereby isolating some factors for explaining (or confounding the explanation of) breast cancer outcome phenotypes?

A strongly predictive set of 83 genes does emerge, with clear additive power; i.e., the more genes from the set are used, the better the prediction performance (Super Proliferation Set, **SPS**; see Supplementary Data 1) (Figure 1A S1 to S20). About 20 SPS genes are required for a signature to be significantly associated with phenotype. In contrast, although proliferation genes are thought to be a source of confounding, they are not born equal: proliferation genes not part of SPS clearly lack additive power and significant association with phenotype (Figure 1A A1 to A20). This example illustrates the value of mining existing information and also lends insight as to which gene groups are more likely relevant, and therefore suitable for signature inclusion (i.e., use collective prior knowledge from meta-analysis to guide and refine future studies).

Systematic evaluation of confounders

Confounders are not homogeneous: although the vast majority of proliferation genes are non-causal correlates, a subset is likely phenotypically relevant (Figure 1A). To exemplify this point, SPS was compared with two proliferation gene sets (Prolif and meta-PCNA) (see **Supplementary**) revealing that almost all SPS genes are proliferation-associated (Figure 1B). But interestingly, only intersecting areas with SPS are strongly predictive, suggesting that incorporation of SPS genes are why these proliferation gene sets are powerful predictors in the first place.

Going beyond Venet et al.'s meta-analysis[7], the PAM50 is a commercialized signature assay with 15 genes shared with SPS [8]. The full PAM50 has a good $\log_{10}$ p-value of -3.48 on NKI; this drops significantly to -0.14 upon removing SPS genes. This means, at least where the NKI benchmark is concerned, SPS genes are a major contributor towards PAM50's predictive performance.

But what makes SPS special, and are there any distinguishing features between the two subsets SPS $\cap$Prolif$\cap$ meta-PCNA (the 43 genes common to all 3 signatures; Figure 1B) and SPS$\cap$Prolif $\setminus$ meta-PCNA (the 38 genes shared between SPS and Prolif, sans meta-PCNA; Figure 1B)? We first compared these against the core proliferation gene lists described by Whitfield et al. [9]. Both SPS$\cap$Prolif$\cap$ meta-PCNA and SPS$\cap$Prolif $\setminus$ meta-PCNA are closely associated with the core proliferation signatures, and include many classical markers of proliferation and breast cancer, including BRCA1 [9] (Supplementary Data 1). Therefore, there is strong relevance support for SPS. Network statistics based on a protein interaction network (See Supplementary) further reveal that SPS genes are hubs (highly connected network components), with SPS$\cap$Prolif$\cap$meta-PCNA being more highly connected than SPS$\cap$Prolif $\setminus$ meta-PCNA (Figure 1C) and with less variability in transitivity (also known as the clustering coefficient, measuring the degree of inter-connectivity amongst the first-degree neighbors) (Figure 1D).

There is clear advantage in systematically taking evidence from multiple sources---the genes in the intersection of Prolif (based on literature and annotation), meta-PCNA (based on correlation to PCNA expression) and SPS (based on taking conserved genes from the most powerful published signatures) exhibit specific additive effects (Figure 1A), have very strong predictive power (Figure 1B), and are super-hubs (highly connected; occupying important positions in the cellular networks) (Figure 1C). This body of evidence suggests that, despite belonging to the proliferation confounder, SPS genes might be important due to its phenotype relevance. Whether SPS genes should be considered confounders depends on the objective of the signature: if one is looking for a prognostic signature for breast cancer subtypes which are characterized by high proliferation (e.g. ER^-/$HER2^-$ and $HER2^{+}$), it might be appropriate to disregard these genes [10]. To generalize, some genes are associated with both confounding factors and useful signal; these need to be established via careful systematic evaluation.

Generalizability tests

Gene signature inference should not stop at one benchmark dataset as there is always the possibility the signature is over-fitted and therefore non-generalizable (i.e., the signature only works on one dataset, but not others). The minimum requirement should be at least one independent validation on a completely new dataset (cross-validation is not good enough [11-13]). Given wide availability of data, a good practice is to leverage on existing published data (which are not used for determining the signature) and evaluate against as many as possible to infer generalizability.

There are various flavors of generalizability tests: the simplest being to establish a baseline on the number of expected false positives and determine how the signature performs against it. In Venet et al. [7], about 54% of random signatures sampled are insignificant (i.e. nominal p-value > 0.05). Thus, we may postulate that a random signature has a 46% chance of being significant in a breast cancer dataset. Therefore, it has a $46\%^n$ chance of being significant across n independent breast cancer datasets. If n = 7, then there is a 0.4% (= $46\%^7$) chance of achieving significance across seven independent datasets. Having established this baseline, we may then go on to validate SPS on other published datasets. We downloaded 7 datasets from GEO for this purpose (see Supplementary). SPS performed well, with significant association with phenotype across all seven independent datasets (Figure 2). Given that there is only a 0.4% possibility of such occurrence, it is unlikely due to chance.

We may also model expected values based on p=46% as a binomial distribution. This is akin to a simulated coin flip where seven coins (with a chance of success of landing heads = 46%, and tails = 54%) are tossed simultaneously each time. For each toss, we count the number of heads. We repeat this 1,000 times to get the binomial distribution and compare this against that of observed values (Figure 2).

Given the binomial (theoretical) distribution, random signatures only have a 0.3% chance of being significant in all seven datasets. An “observed” distribution can also be produced empirically by producing 1000 randomly-generated signatures (equal in size to SPS), and testing each across the seven independent datasets. Note that the theoretical and observed distributions are quite different (chi-square test; p-value = 0.013). One possible explanation may be that the binomial distribution and/or the inferred probability value of 0.46 are unsuitable. But a more likely explanation is that while the breast cancer datasets are independent with regards to where they come from, they are nonetheless all breast cancer datasets and some common characteristics are expected. So, when a signature is significant in one dataset, there is an increased likelihood for it to be significant in another dataset (i.e. the assumption of independence is invalid). Another more likely explanation is that some sampled random signatures share some genes; i.e. the random signatures are not fully independent of each other. So when a signature is significant in one dataset, other signatures sharing genes with it are also likely significant in the same dataset. Regardless, both observed and theoretical distributions suggest getting significance in all seven datasets is highly unlikely, and therefore support the idea that SPS is generalizable despite not testing every breast cancer dataset possible.

But the result above is not sufficient as passing the above does not mean other signatures perform badly. In fact, it turns out that many signatures do beat expectation. In particular, approximately 80% of published signatures are generalizable (Figure 2). However, this is associated with SPS: the more SPS genes contained therein, the more likely a published signature is universal (Figure 2 Inset). More importantly and fortunately, although random signatures can beat any published signature on one dataset, they are hardly generalizable.

The presence of predictive power in a signature does not mean it is easily detectable, or not heavily confounded with other sources of heterogeneity. Combining principal components analysis (PCA) with generalizability tests is useful for checking this [14]. We generated 1,000 random signatures of size 83 (i.e. same size as SPS). And for each random signature, we tested the minimum p-value associated with principal components (PC) 1 to 10 induced by the 83 genes of this random signature on the seven datasets. We observed that the more SPS genes therein, the more significant this minimum p-value is. In this scenario, amongst PC 1 to 10, there is always at least one PC significantly correlated with survival or prognosis (Supplementary Table 2). And in these cases, SPS is correspondingly significantly enriched in the survival-associated PC. However, PC 1 to 3 (corresponding to the major components of variance) are not always the most differential with regards to survival (Supplementary Table 2).

As the datasets are not properly cleaned to deal with various sources of bias, it cannot be established *a priori* which PC is the correct one to use on which dataset (in practical usage, this is important if the intention is combine datasets for meta-analyses) [15,16]. But as a simple first pass, it is reasonable to consider using the PC achieving the highest significance among the top 10 PCs (see Figure 3A) and setting the score to the p-value of this PC (for determining correlation with phenotype of the corresponding dataset). This better reflects the practical-use scenario; as in the absence of perfect information, it is an intuitive choice to use the best PCs for prediction.

Relative to published signatures, SPS is not always the best performer (with the most significant p-values) but it does remain consistently significant throughout all seven datasets. A generalizable signature needs not always be the most significantly associated with phenotype (against other signatures) as p-values are unstable and its magnitude cannot be relied on as an objective gauge of the strength of phenotype association [6,17], but it should be reproducible: i.e., it should always pass the threshold for significance across any independent datasets (Figure 3B). To see the additive effects of low and high SPS enrichment more objectively, random sampling is always useful. Here, four sets of 1,000 random signatures (size 20) are generated, respectively drawing 0, 25, 50, and 100% of the 20 genes in the signature from SPS. These simulations are tested for the minimum p-value of PC 1 to 10 across all seven datasets. Again, it was observed that increased proportion of SPS genes clearly increases association with survival (Supplementary Figure 1).

Recommendations

Generally, it is good analytical practice to construct reasonable hypothesis statements, check the appropriateness of the summary statistics and reference distributions. But this does not exclude the existence of other sources of confounders. It is impracticable to exhaustively isolate and exclude all of these, especially since many will not be known *a priori*. Unfortunately, not addressing these would certainly have negative impact on gene signature inference; so something has to be done. Fortunately, robustness can be built into analysis without explicitly identifying and negating all sources of confounding.

The first recommendation is to build upon prior knowledge: meta-analysis of published signatures is useful for identifying recurring genes, which in turn, hints at biological relevance. Here, taking the intersection amongst best-performing published signatures facilitated inference of a powerful signature with generalizable properties.

The second recommendation is that when many random signatures are significant, it is likely that many confounders and real causes are present. Genes suspected to be associated with confounders can be informative. They should not be naively discarded without careful and systematic evaluation of their properties. In breast cancer, although many irrelevant signatures are confounded with proliferation-associated genes, an identifiable subset has robust properties such as strong correlation with phenotype with additive prediction effects. These properties are not observable in random subsets of other proliferation genes.

Finally, irrelevant signatures do not exhibit generalizability: when evaluating a signature, it is worthwhile to consider a wide spectrum of independent datasets. If the signature works well across all datasets, it is likely to be useful, and we should be less worried about its significance being due to chance or its being outperformed in a dataset by randomly generated signatures.

Conclusions

Inference of predictive signatures can be augmented with the use of prior knowledge (via meta-analysis); careful and systematic evaluation of gene sets, even if they overlap with known sources of confounding; and rigorously testing inferred signatures against as many published datasets as possible.

Funding

LW gratefully acknowledges support by a Kwan-Im-Thong-Hood-Cho-Temple chair professorship.

Author contributions

WWBG and LW co-designed the methodologies and co-wrote the manuscript.

Acknowledgements

WWBG and LW acknowledge Vincent de Tours, and his colleagues for codes and data obtained from their publication.

Competing interests

The authors declare no conflicting interests, financial or otherwise.

References

1 Goh, W.W.B. and Wong, L. (2018) Dealing with Confounders in Omics Analysis. *Trends Biotechnol* 36 (5), 488-498

2 Lutz, B. and Werner, M. (2012) The Anna Karenina principle: A way of thinking about success in science. *J. Am. Soc. Inf. Sci. Technol.* 63 (10), 2037-2051

3 Zaneveld, J.R. et al. (2017) Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nat Microbiol* 2, 17121

4 Begley, C.G. and Ioannidis, J.P. (2015) Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res* 116 (1), 116-126

5 Patil, P. et al. (2015) Test set bias affects reproducibility of gene signatures. *Bioinformatics* 31 (14), 2318-2323

6 Wang, W. et al. (2016) Feature selection in clinical proteomics: with great power comes great reproducibility. *Drug Discov Today* 22(6), 912-918

7 Venet, D. et al. (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 7 (10), e1002240

8 Dowsett, M. et al. (2013) Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J Clin Oncol* 31 (22), 2783-2790

9 Whitfield, M.L. et al. (2006) Common markers of proliferation. *Nat Rev Cancer* 6 (2), 99-106

10 Wirapati, P. et al. (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 10 (4), R65

11 Braga-Neto, U.M. and Dougherty, E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 20 (3), 374-380

12 Qin, L.X. et al. (2016) Cautionary Note on Using Cross-Validation for Molecular Classification. *J Clin Oncol*

13 Soneson, C. et al. (2014) Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One* 9 (6), e100335

14 Goh, W.W.B. et al. (2017) Can Peripheral Blood-Derived Gene Expressions Characterize Individuals at Ultra-high Risk for Psychosis? *Computational Psychiatry*, 1-16

15 Giuliani, A. (2017) The application of principal component analysis to drug discovery and biomedical data. *Drug Discov Today* 22(7), 1069-1076

16 Goh, W.W. et al. (2017) Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 35(6), 498-507

17 Halsey, L.G. et al. (2015) The fickle P value generates irreproducible results. *Nat Methods* 12 (3), 179-185

Figures

Figure 1 (A) Genes sampled from the super-proliferation set, or SPS, exhibits clear additive effect on significance as opposed to randomly selected proliferation genes. Y-axis: $\log_{10}$(p-value). X-axis: Genes sampled from SPS (S) and all proliferation genes (A). Sampling sizes range from 1, 5, 10 and 20. Inset values for A1 to A20 are the median $\log_{10}$(p-values). **(B) Overlaps between proliferation genes (Prolif), meta-PCNA (PCNA) and the SPS.** Intersecting genes with SPS have high predictive power for survival as indicated by the $\log_{10}$(p-values) (** and ***). **(C) SPS is enriched for high-degree nodes (hubs).** Y-axis: degree coefficient. **(D) SPS∩Prolif∩PCNA has reduce variability for transitivity (clustering-coefficient) compared to SPS∩Prolif\meta-PCNA and other genes in the global network.** Y-axis: Transitivity. (SPS∩Prolif∩PCNA is the intersection of the 3 gene sets; SPS∩Prolif\meta-PCNA is the intersection of SPS and Prolif, sans the component shared with meta-PCNA)

Figure 2 It is highly unlikely for random signatures to be universally significant across all 7 independent breast cancer datasets. Y-axis: Frequency distribution for signatures --- including 1,000 random signatures (blue), 1,000 counts from a binomial distribution based on an expected probability of success = 0.46 (red), and 48 published signatures (yellow). X-axis: The number of breast cancer datasets a signature is significant in. Inset: Generalizability of published signatures is associated with SPS enrichment.

Figure 3 Published signatures with more SPS genes are less likely to fail. (A) Signatures with less SPS genes have more tendency to fail (above the pink line marking p= 0.05). The higher the number of SPS genes in a published signature, the better it performs. Y-axis: min p-value PC1:10. X-axis: individual GEO datasets. **(B) Proportion of signatures that do better than SPS** (Top Table) **and the SPS min p-value PC1:10** (Bottom Table)**.**

Turning straw into gold: building robustness into gene signature inference

Wilson Wen Bin Goh [1*], Limsoon Wong [2,3*]

4. School of Biological Sciences, Nanyang Technological University, Singapore
5. Department of Computer Science, National University of Singapore, Singapore
6. Department of Pathology, National University of Singapore, Singapore

*Corresponding Author: Wilson Wen Bin Goh, wilsongoh@ntu.edu.sg; Limsoon Wong, wongls@comp.nus.edu.sg

Address for correspondence/proofs:

Wilson Wen Bin Goh, PhD

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

Email: wilsongoh@ntu.edu.sg, Tel: +65-65162902

Limsoon Wong, PhD

Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417

Email: wongls@comp.nus.edu.sg, Tel: +65-65162902

SUPPLEMENTARY

MATERIALS AND METHODS

Breast cancer microarray datasets

To investigate the impact of proliferation-gene removal from random signatures, and from the entire dataset altogether, the same breast cancer datasets (the NKI for evaluating breast cancer survival outcome) as in Venet et al. [1] are used here. Signatures are tested for correlation with outcome (survival status) where the lower the p-value, the higher the association (see below).

Additionally, 7 breast cancer datasets from GEO (GDS5027, GDS4051, GSE21653, GDS4083, GDS4114, GDS4766, and GDS4093) were downloaded, and used as further validation (see Generalizability test below).

Proliferation and meta-PCNA signature

There are two groups of proliferation signatures, the proliferation (Prolif) set comprising 1,003 genes, which is inaccurately called the "cell cycle" set by Venet et al. [2], and the meta-PCNA, which is a list of 129 genes most correlated to the PCNA gene.

Protein-protein interaction network and network analysis

A reference protein-protein interaction network is taken from Yong et al. [3]. Centrality analyses for degree and transitivity are performed using the R iGraph package [4].

Software

All codes for execution and graphics are written and executed in R. The scripts for breast cancer survival outcome were modified from the original codes of Venet et al. [1], except for the Venn diagrams, which are obtained using Venny (http://bioinfogp.cnb.csic.es/tools/venny/), and functional annotation, which was performed by supplying gene lists to DAVID (http://david.abcc.ncifcrf.gov/) [5].

Association with breast cancer outcome

Quantification of association with outcome first involves computing the first principal component (PC1) of the signature (using R's prcomp) and then splitting the cohort according to the median of PC1.

Given a binary stratification of the cohort, the hazard ratio (HR) and the related log-rank p-values are computed using the standard Cox procedure implemented in R's coxph.

Inference of the super-proliferation set and spiking

For each signature, two sets of p-values (inferred from Cox's analysis) can be calculated, P (inclusive of proliferation genes) and NP (excluding proliferation genes). The difference between these two p-values, delta(P–NP) measures the extent of dependency a signature's performance is on the proliferation signature.

We selected those genes supported by at least 2 signatures with delta(P – NP) below -3.5 (strong-prolif) (GLINSKY, DAI, RHODES, ABBA, WHITFIELD) (83 genes in total; 81 genes overlap with Prolif+meta-PCNA, 7.6% of all proliferation genes; c.f. Supplementary Table 1).

We spiked these genes into a neutral signature and evaluate influence on the p-value (by spiking, we mean to randomly pick SPS or non-SPS genes, add these to a neutral signature, and then evaluate changes on the survival p-value). For spiking, we selected SORLIE [6], which is a 15-gene signature, with no overlaps with known proliferation genes, and a nominal $\log_{10}$ p-value of -0.033 (highly non-significant).

Since we have many more proliferation genes in strong-prolif, we tested for additive effects relative to all proliferation genes (Prolif+meta-PCNA). We resampled subsets from sizes of 1 to 20 strong-prolif genes 1,000 times, added them to SORLIE, and tested for predictive power (c.f. Figure 1A S1 to S20). As a contrast, we also repeated the same experiment by resampling from all proliferation genes (c.f. Figure 1A A1 to A20).

Association with breast cancer outcome

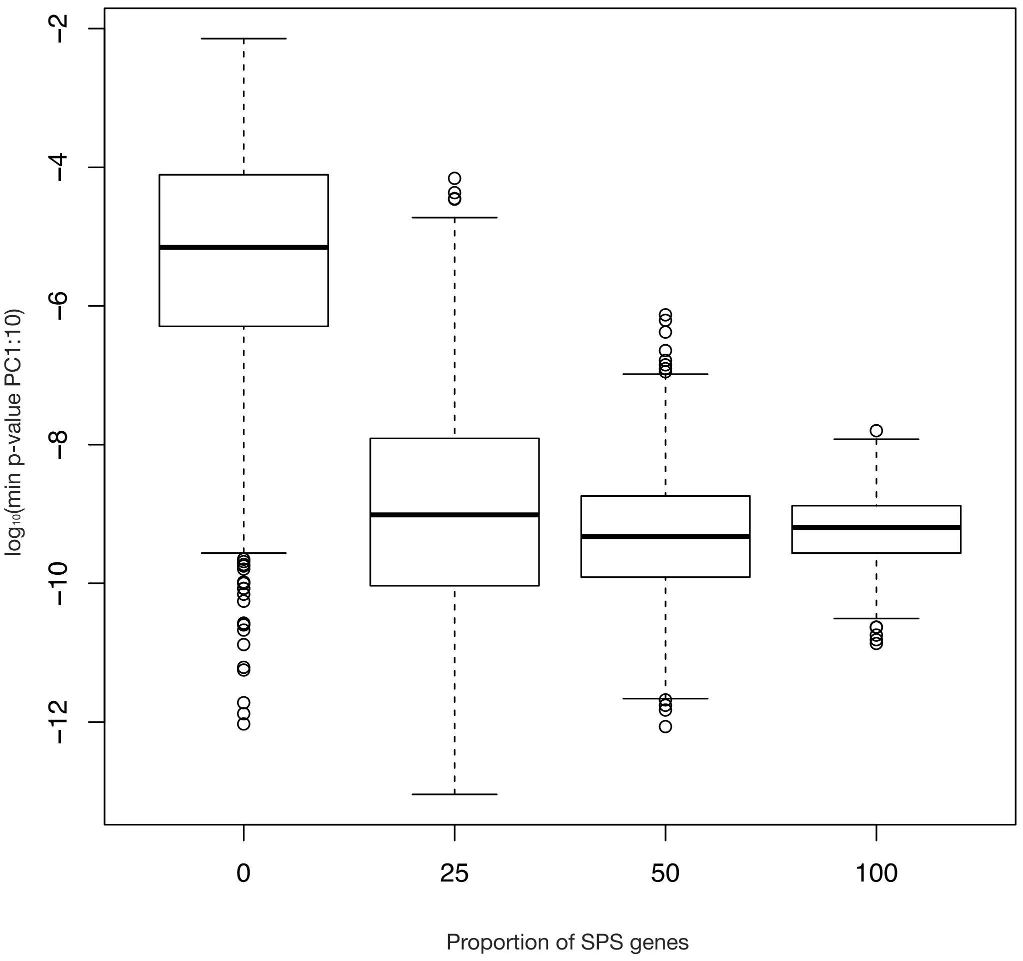
Quantification of association with outcome first involves computing the first principal component (PC1) of the signature (using R's prcomp) and then splitting the cohort according to the median of PC1. Given a binary stratification of the cohort, the hazard ratio (HR) and the related log-rank p-values are computed using the standard Cox procedure implemented in R's coxph. This is the same procedure used by Venet et al.

Generalizability test

SPS itself can also be considered a potential signature for breast cancer survival. And therefore, it must demonstrate generalizability, i.e., the ability to be predictively accurate across all other related datasets. We downloaded seven breast cancer datasets from GEO (https://www.ncbi.nlm.nih.gov/geo/), where data on survival or prognosis is present. We kept the original formatting and data processing on the GDS (GEO DataSet) files (i.e., no correction for potential technical/biological bias), extracted all probes that corresponded to SPS genes (no probe collapsing based on genes), and performed Principal Components Analysis (PCA) on the latter.

To assess generalizability of a random signature on these same seven datasets, the same procedure above was used with one modification: for a random signature, probes corresponding to genes in this random signature were extracted instead of SPS genes.

SUPPLEMENTARY FIGURES



Supplementary Figure 1 The higher the proportion of SPS genes in a random signature, the stronger the association with survival. Y-axis: min $\log_{10}$ p-value PC1:10. X-axis: Proportion of SPS genes (from 0 to 100%).

SUPPLEMENTARY TABLES

Supplementary Table 1 Statistics of 47 published breast cancer gene signatures + meta-PCNA (The first 24 are considered small and the remaining large signatures) (P: Cox analysis p-value inclusive of proliferation genes. NP: Cox analysis p-value exclusive of proliferation genes. Delta (P – NP) is the difference in p-value indicating the extent of dependency a signature's performance is on the proliferation signature.)

Signature	size	number of proliferation genes	Log$_{10}$(p_val) (P)	Log$_{10}$(p_val) (NP)	Delta (P - NP)
ADORNO	2	2	-0.495	0.000	-0.495
PEI	2	2	-0.094	0.000	-0.094
BUFFA	3	0	-2.161	-2.161	0.000
WELM	3	0	-1.545	-1.545	0.000
HE	6	0	-0.431	-0.431	0.000
TAVAZOIE	6	0	-0.180	-0.180	0.000
VALASTYAN	6	1	-0.315	-0.306	-0.009
GLINSKY	11	4	-4.092	-0.041	-4.051
HU	13	2	-0.725	-0.722	-0.003
YU	14	0	-1.667	-1.667	0.000
SORLIE	15	0	-0.033	-0.033	0.000
PAIK	16	6	-2.929	-2.577	-0.351
RAMASWAMY	16	3	-2.331	-1.567	-0.763
IVSHINA	17	14	-3.724	-2.501	-1.223
MILLER	18	4	-0.277	-0.482	0.205
KORKOLA	21	3	-0.599	-0.178	-0.421
BUESS	30	2	-0.665	-0.413	-0.252
MA	30	22	-5.251	-2.301	-2.950
DAI	35	29	-5.907	-2.576	-3.330
PAWITAN	46	19	-4.187	-2.653	-1.534
WONG-PROTEAS	46	7	-3.351	-2.063	-1.289
SHIPITSIN	56	5	-0.415	-0.036	-0.380
VANTVEER	60	14	-3.156	-2.155	-1.002
RHODES	67	43	-5.323	-1.553	-3.770
WANG-76	69	16	-3.371	-2.059	-1.313
CARTER	70	49	-5.127	-4.727	-0.400
HALLSTROM	78	24	-4.847	-2.032	-2.815
SOTIRIOU-GGI	90	63	-5.296	-5.063	-0.233
ABBA	111	79	-5.760	-2.123	-3.637
META-PCNA	129	71	-6.021	-0.598	-5.424
CHI	136	8	-0.994	-0.894	-0.100
MORI	156	7	-0.029	-0.050	0.021
SAAL	162	40	-4.884	-4.381	-0.503
LIU	167	12	-4.130	-4.067	-0.063
KOK	179	38	-3.280	-1.125	-2.156
WONG-MITOCHON	217	11	-5.316	-5.386	0.071
WANG-ALK5T204D	239	8	-0.692	-0.880	0.187
TAUBE	242	10	-0.599	-0.492	-0.107
WONG-ESC	335	112	-4.574	-5.162	0.588
SOTIRIOU-93	343	56	-4.091	-3.976	-0.115
CHANG	355	47	-6.226	-6.232	0.006
BEN-PORATH-EXP1	367	85	-4.892	-3.245	-1.647
CRAWFORD	377	153	-6.042	-3.453	-2.589
WEST	468	34	-1.560	-1.842	0.282
WHITFIELD	587	556	-6.545	-0.136	-6.409
BEN-PORATH-PRC2	631	9	-4.398	-3.599	-0.799
REUTER	714	63	-0.320	-0.343	0.023

HUA	1345	122	-3.683	-2.073	-1.610

Supplementary Table 2 Association of SPS genes (based on top 10 PCs) with breast cancer survival. The values in the table are the respective Kruskal-Wallis test p-values.

PC	GDS5027	GDS4051	GSE21653	GDS4083	GDS4114	GDS4766	GDS4093
1	0.544	0.000	0.000	0.015	0.109	0.329	0.026
2	0.734	0.419	0.044	0.019	0.631	0.047	0.966
3	0.000	0.488	0.163	0.349	0.037	0.193	0.470
4	0.200	0.525	0.054	0.349	0.631	0.014	0.186
5	0.001	0.817	0.921	0.190	0.522	0.664	0.231
6	0.020	0.729	0.039	0.574	0.873	0.539	0.194
7	0.005	0.862	0.746	0.851	0.522	0.138	0.162
8	0.344	0.862	0.055	0.708	0.337	0.942	0.246
9	0.310	0.488	0.987	0.925	0.873	0.247	0.499
10	0.108	0.908	0.647	0.708	0.749	0.914	0.389

SUPPLEMENTARY DATA

Supplementary Data 1 The SPS gene set and its corresponding overlaps with the proliferation signatures Prolif and (meta)-PCNA.

43	**Strong + Prolif + PCNA**	38	**Strong + Prolif**
ENTREZ_GENE_ID	Name	ENTREZ_GENE_ID	Name
9833	maternal embryonic leucine zipper kinase(MELK)	4605	MYB proto-oncogene like 2(MYBL2)
1033	cyclin dependent kinase inhibitor 3(CDKN3)	4751	NIMA related kinase 2(NEK2)
2305	forkhead box M1(FOXM1)	3161	hyaluronan mediated motility receptor(HMMR)
4001	lamin B1(LMNB1)	5347	polo like kinase 1(PLK1)
9768	KIAA0101(KIAA0101)	55839	centromere protein N(CENPN)

11004	kinesin family member 2C(KIF2C)	2621	growth arrest specific 6(GAS6)
9212	aurora kinase B(AURKB)	3833	kinesin family member C1(KIFC1)
1163	CDC28 protein kinase regulatory subunit 1B(CKS1B)	1846	dual specificity phosphatase 4(DUSP4)
891	cyclin B1(CCNB1)	3838	karyopherin subunit alpha 2(KPNA2)
79682	centromere protein U(CENPU)	9493	kinesin family member 23(KIF23)
890	cyclin A2(CCNA2)	672	BRCA1, DNA repair associated(BRCA1)
1164	CDC28 protein kinase regulatory subunit 2(CKS2)	9134	cyclin E2(CCNE2)
3148	high mobility group box 2(HMGB2)	2730	glutamate-cysteine ligase modifier subunit(GCLM)
9133	cyclin B2(CCNB2)	699	BUB1 mitotic checkpoint serine/threonine kinase(BUB1)
22974	TPX2, microtubule nucleation factor(TPX2)	29028	ATPase family, AAA domain containing 2(ATAD2)
701	BUB1 mitotic checkpoint serine/threonine kinase B(BUB1B)	7272	TTK protein kinase(TTK)
4085	MAD2 mitotic arrest deficient-like 1 (yeast)(MAD2L1)	993	cell division cycle 25A(CDC25A)
6241	ribonucleotide reductase regulatory subunit M2(RRM2)	79019	centromere protein M(CENPM)

11130	ZW10 interacting kinetochore protein(ZWINT)	4582	mucin 1, cell surface associated(MUC1)
332	baculoviral IAP repeat containing 5(BIRC5)	1894	epithelial cell transforming 2(ECT2)
6790	aurora kinase A(AURKA)	23397	non-SMC condensin I complex subunit H(NCAPH)
9918	non-SMC condensin I complex subunit D2(NCAPD2)	990	cell division cycle 6(CDC6)
9232	pituitary tumor-transforming 1(PTTG1)	3608	interleukin enhancer binding factor 2(ILF2)
11065	ubiquitin conjugating enzyme E2 C(UBE2C)	1736	dyskerin pseudouridine synthase 1(DKC1)
51203	nucleolar and spindle associated protein 1(NUSAP1)	55165	centrosomal protein 55(CEP55)
991	cell division cycle 20(CDC20)	9319	thyroid hormone receptor interactor 13(TRIP13)
2237	flap structure-specific endonuclease 1(FEN1)	9928	kinesin family member 14(KIF14)
1058	centromere protein A(CENPA)	1062	centromere protein E(CENPE)
4172	minichromosome maintenance complex component 3(MCM3)	3015	H2A histone family member Z(H2AFZ)
4175	minichromosome maintenance complex component 6(MCM6)	1063	centromere protein F(CENPF)

4288	marker of proliferation Ki-67(MKI67)	3014	H2A histone family member X(H2AFX)
983	cyclin dependent kinase 1(CDK1)	10615	sperm associated antigen 5(SPAG5)
9055	protein regulator of cytokinesis 1(PRC1)	2146	enhancer of zeste 2 polycomb repressive complex 2 subunit(EZH2)
5111	proliferating cell nuclear antigen(PCNA)	27338	ubiquitin conjugating enzyme E2 S(UBE2S)
4171	minichromosome maintenance complex component 2(MCM2)	1869	E2F transcription factor 1(E2F1)
55143	cell division cycle associated 8(CDCA8)	7033	trefoil factor 3(TFF3)
7298	thymidylate synthetase(TYMS)	10403	NDC80, kinetochore complex component(NDC80)
9700	extra spindle pole bodies like 1, separase(ESPL1)	5885	RAD21 cohesin complex component(RAD21)
51512	G2 and S-phase expressed 1(GTSE1)		
7153	topoisomerase (DNA) II alpha(TOP2A)		
5984	replication factor C subunit 4(RFC4)		
8318	cell division cycle 45(CDC45)		
83461	cell division cycle associated 3(CDCA3)		

REFERENCES

1 Venet, D. et al. (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* 7 (10), e1002240

2 Whitfield, M.L. et al. (2006) Common markers of proliferation. *Nat Rev Cancer* 6 (2), 99-106

3 Yong, C.H. et al. (2012) Supervised maximum-likelihood weighting of composite protein networks for complex prediction. *BMC Syst Biol* 6 Suppl 2, S13

4 Csardi, G. and Nepusz, T. (2006) The igraph Software Package for Complex Network Research. *InterJournal* Complex Systems, 1695

5 Dennis, G., Jr. et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4 (5), P3

6 Sorlie, T. et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 98 (19), 10869-10874