ELSEVIER

feature

Turning straw into gold: building robustness into gene signature inference

Wilson Wen Bin Goh[1], wilsongoh@ntu.edu.sg **and Limsoon Wong**[2,3], wongls@comp.nus.edu.sg

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, namely: leveraging existing data and knowledge; careful and systematic evaluation of gene sets, even if they overlap with known sources of confounding; and rigorous testing of inferred signatures against as many published data sets as possible. Here, using a re-examination of a breast cancer data set and 48 published signatures, we illustrate the value of adopting these GAPs.

Introduction

Statistical feature selection of 'omics' data is a practical means of deriving signatures for predictive purposes. Although 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].

Therefore, naïve reliance on basic statistics leads to a lack of signature reproducibility (getting a similar signature with a different data set) [4–6] and signature generalizability (able to correctly predict phenotype based on a different data set) [7]. Addressing confounders is 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 that the correct test statistics and reference distributions are used [1]. Broadly, these constitute GAPs in the context of general analysis. However, more robustness can be introduced for the purpose of signature inference. Using a re-examination of the data set of Venet *et al.* [7], we illustrate here the following GAPs: (i) the importance of meta-analysis; (ii) systematic evaluation of confounders; and (iii) generalizability tests.

The case study

In their study, Venet *et al.* evaluated 48 published breast cancer signatures using an independent data set [7]. A good signature is one that is associated significantly with outcome or phenotype. However, in this study, the authors found that most published signatures did not outperform randomly generated signatures, and even irrelevant signatures derived from other phenotypes did well; that is, statistical significance alone cannot prove relevance.

Suspected confounders include: (i) use of an 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%; (ii) 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; and (iii) an 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 is likely to

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

exhibit large fluctuations with 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 their study, Venet *et al*. evaluated 48 independently published breast cancer signatures against the NKI benchmark data set (see the Supplemental information online) [7], which revealed that these signatures were not only very different from each other, but also performed variably on the benchmark.

Each signature can be considered an independent sample (with different degrees of error, leading to variable performance); thus, an aggregate analysis is intuitively more informative than any single study. Venet *et al*.'s meta-analysis revealed that many of these signatures performed no better than randomly generated ones [7], suggesting that the composition of many published signatures is artifact infested (i.e., overladen with proliferation genes) (see Table S1 in the Supplemental information online). Although it is standard practice to use cross-validation (at the minimum) in signature inference studies, it is clearly insufficient: given current 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 of breast cancer outcome, this creates an interesting opportunity: given that 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 'Materials and Methods' in the Supplemental information online) [7], thereby isolating 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 Supplemental Data 1 in the Supplemental information online; Fig. 1a (S1–S20)]. Approximately 20 SPS genes are required for a signature to be significantly associated with phenotype. By 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 [Fig. 1a (A1–A20)]. This example illustrates the value of mining existing information and also lends insight into which gene groups are more likely relevant and, therefore, suitable for signature inclusion (i.e., use collective prior knowledge from a meta-analysis to guide and refine future studies).

Systematic evaluation of confounders

Confounders are not homogeneous: although most proliferation genes are noncausal correlates, a subset is likely phenotypically relevant (Fig. 1a). To exemplify this point, SPS was compared with two proliferation gene sets (Prolif and meta-PCNA; see the Supplemental information online), revealing that almost all SPS genes were proliferation associated (Fig. 1b). Interestingly, only intersecting areas with SPS were strongly predictive, suggesting that the incorporation of SPS genes was why these proliferation gene sets were 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 decreases 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 the predictive performance of PAM50.

However, what makes SPS special, and are there any distinguishing features between the two subsets SPS ∩Prolif∩meta-PCNA (the 43 genes common to all three signatures; Fig. 1b) and SPS∩Prolif\meta-PCNA (the 38 genes shared between SPS and Prolif, without meta-PCNA; Fig. 1b)? We first compared these against the core proliferation gene lists described by Whitfield *et al*. [9]. Both SPS∩Prolif∩meta-PCNA and SPS∩Prolif\meta-PCNA were closely associated with the core proliferation signatures, and included many classical markers of proliferation and breast cancer, including *BRCA1* [9] (Supplementary Data 1 in the Supplemental information online). Therefore, there is strong relevance support for SPS. Network statistics based on a protein interaction network (see the Supplemental information online) further revealed that SPS genes are hubs (highly connected network components), with SPS∩Prolif∩meta-PCNA being more highly connected than SPS∩Prolif\meta-PCNA (Fig. 1c) and with less variability in transitivity (also known as the clustering coefficient, measuring the degree of interconnectivity among the first-degree neighbors) (Fig. 1d).

There is a 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 (Fig. 1a), have very strong predictive power (Fig. 1b), and are superhubs (i.e., highly connected; occupying important positions in the cellular networks; Fig. 1c). This body of evidence suggests that, despite belonging to the proliferation confounder, SPS genes are important because of their 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 that 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 data set because there is always the possibility that the signature is overfitted and, therefore, nongeneralizable (i.e., the signature only works on one data set). The minimum requirement should be at least one independent validation on a completely new data set (cross-validation is not good enough [11–13]). Given the wide availability of data, a good practice is to leverage existing published data (which are not used for determining the signature) and evaluate against as many data sets as possible to infer generalizability.

There are various flavors of generalizability tests: the simplest being to establish a baseline of the number of expected false positives and determine how the signature performs against it. In the study by Venet *et al*. [7], ~54% of random signatures sampled were insignificant (i.e., nominal *P* value >0.05). Thus, we can postulate that a random signature has a 46% chance of being significant in a breast cancer data set. Therefore, it has a $46\%^N$ chance of being significant across *N* independent breast cancer data sets. If $N = 7$, then there is a 0.4% ($= 46\%^7$) chance of achieving significance across seven independent data sets. Having established this baseline, we can then go on to validate SPS on other published data sets. We downloaded seven data sets from GEO for this purpose (see the Supplemental information online). SPS performed well, with significant association with

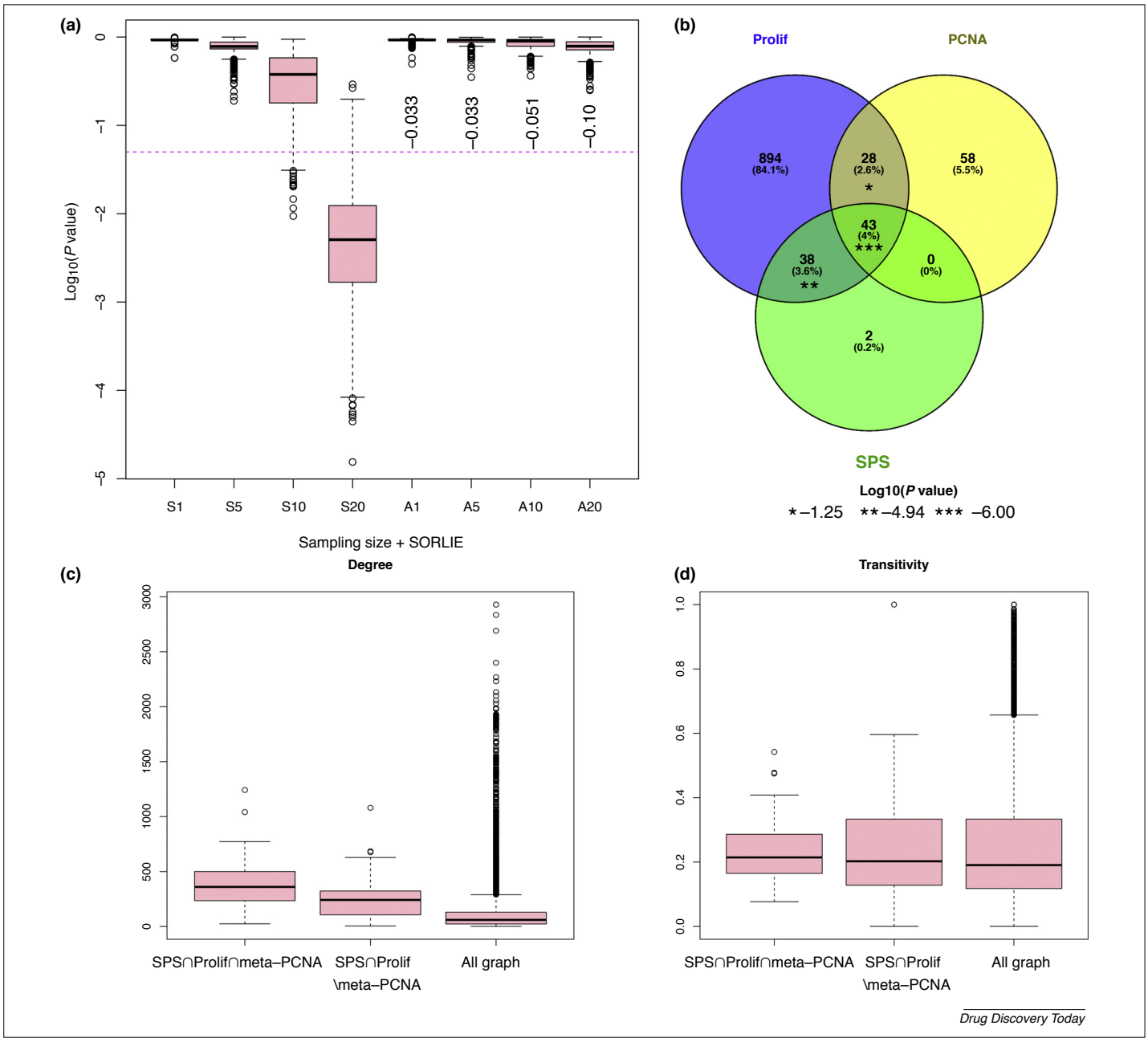


FIGURE 1

Not all confounders are born equal. (a) Genes sampled from the super-proliferation set (SPS) exhibit clear additive effects on significance (correlation with survival) compared with randomly selected proliferation genes. *Y*-axis: $\log_{10}$ (*P* value). *X*-axis: genes sampled from SPS (S) and all proliferation genes (A). Sampling sizes were 1, 5, 10, and 20. Inset values for A1–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 reduced variability for transitivity (clustering coefficient) compared with SPS∩Prolif\meta-PCNA and other genes in the global network. *Y*-axis: transitivity. (SPS∩Prolif∩PCNA is the intersection of the three gene sets; SPS∩Prolif\meta-PCNA is the intersection of SPS and Prolif, without the component shared with meta-PCNA.)

phenotype across all seven independent data sets (Fig. 2). Given that there is only a 0.4% possibility of such occurrence, it is unlikely to have resulted from chance.

We can 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 repeated this 1000 times to get the binomial distribution and compared this against that of observed values (Fig. 2).

Given the binomial (theoretical) distribution, random signatures only have a 0.3% chance of being significant in all seven data sets. 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 data sets. Note that the theoretical and observed distributions are different (chi-square test; *P* value = 0.013). One explanation is that the binomial distribution and/or inferred probability

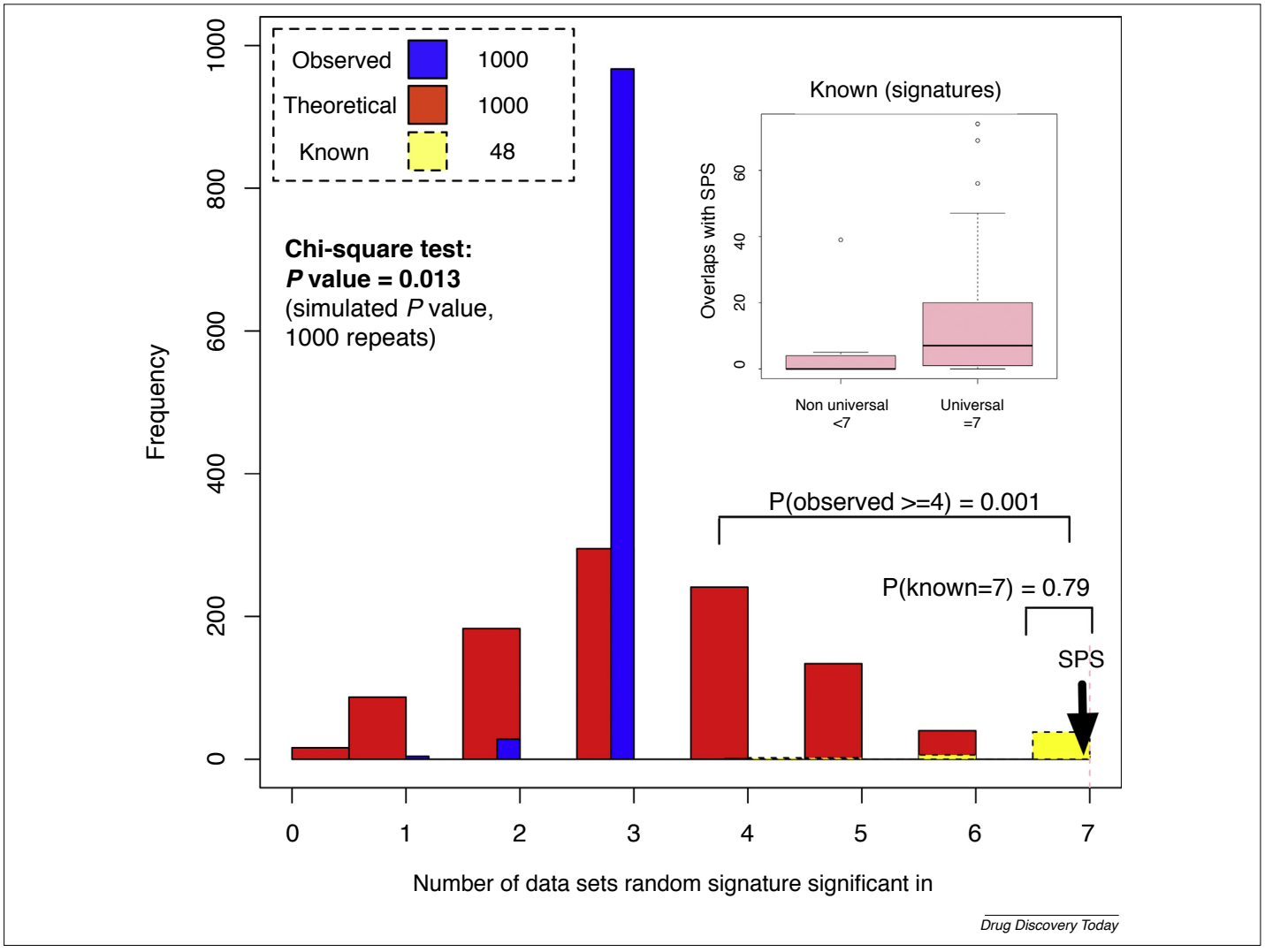


FIGURE 2

It is highly unlikely for random signatures to be universally significant across all seven independent breast cancer data sets. *Y*-axis: frequency distribution for signatures, including 1000 random signatures (blue), 1000 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 data sets a signature is significant in. Inset: generalizability of published signatures is associated with super-proliferation set (SPS) enrichment.

value of 0.46 are unsuitable. However, a more likely explanation is that, while the breast cancer data sets are independent with regards to where they come from, they are nonetheless all breast cancer data sets and some common characteristics are expected. Thus, when a signature is significant in one data set, there is an increased likelihood for it to be significant in another data set (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). Thus, when a signature is significant in one data set, other signatures sharing genes with it are also likely to be significant in the same data set. Regardless, both observed and theoretical distributions suggest that getting significance in all seven data sets is unlikely and, therefore, support the idea that SPS is generalizable despite not testing every breast cancer data set possible.

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

The presence of predictive power in a signature does not mean that 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 1000 random signatures of size 83 (i.e., same size as SPS). For each random signature, we tested the minimum *P* value associated with principal components (PC) 1–10 induced by the 83 genes of this random signature on the seven data sets. We observed that the more SPS genes therein, the more significant this minimum *P* value was. In this scenario, among PC 1–10, there was always at least one PC significantly correlated with survival or prognosis (see Table S2 in the Supplemental information online). In addition, in these cases, SPS was correspondingly significantly enriched in the survival-associated PC. However, PC 1–3 (corresponding to the major components of variance) were not always the

most differential with regards to survival (see Table S2 in the Supplemental information online).

Given that the data sets 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 data set (in practical usage, this is important if the intention is to combine data sets for meta-analyses) [15,16]. However, as a simple first pass, it is reasonable to consider using the PC achieving the highest significance among the top ten PCs (Fig. 3a) and setting the score to the *P* value of this PC (for determining the correlation with phenotype of the corresponding data set). 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.

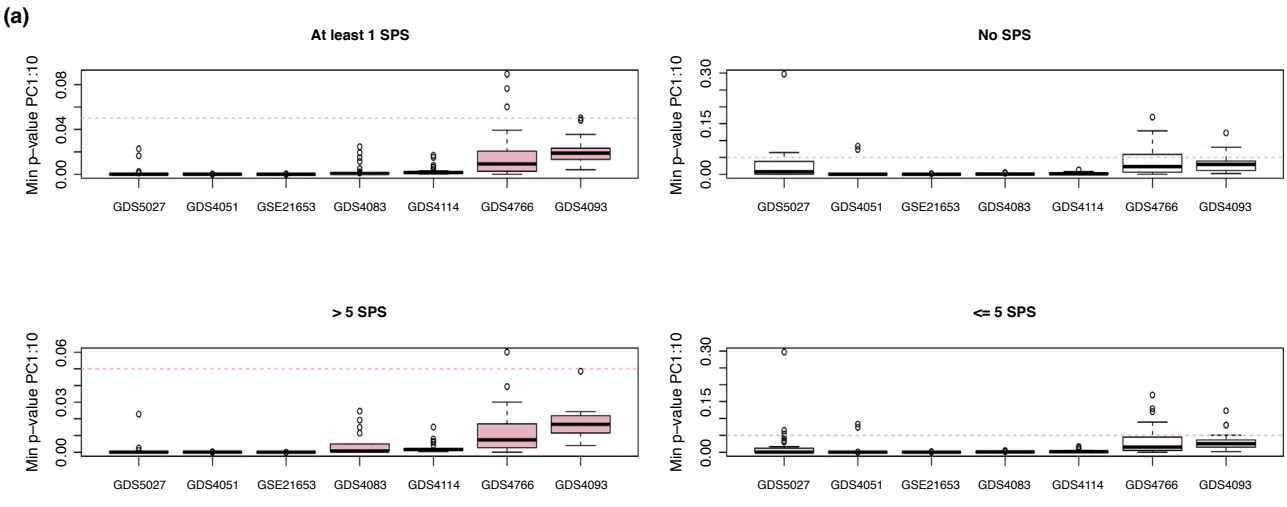
Relative to published signatures, SPS is not always the best performer (with the most significant *P* values) but it does remain consistently significant throughout all seven data sets. A generalizable signature need not always be the most significantly associated with phenotype (against other signatures) because *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 data sets (Fig. 3b)]. To see the additive effects of low and high SPS enrichment more objectively, random sampling is always useful. Here, four sets of 1000 random signatures (size 20) were generated, respectively drawing 0, 25, 50, and 100% of the 20 genes in the signature from SPS. These simulations were tested for the minimum *P* value of PC 1–10 across all seven data sets. Again, it was observed that an increased proportion of SPS genes clearly increased association with survival (see Fig. S1 in the Supplemental information online).

Recommendations

Generally, it is good analytical practice to construct reasonable hypothesis statements and to check the appropriateness of the summary statistics and reference distributions. However, this does not exclude the existence of other sources of confounders. It is impracticable to exhaustively isolate and exclude all of these, especially because many will not be known *a priori*. Unfortunately, not addressing these would negatively impact the gene signature inference; thus, 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 among 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 naïvely 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 a strong correlation with phenotype with additive prediction effects.



(b)

(i)

Lowest PC pval	GDS5027	GDS4051	GSE21653	GDS4083	GDS4114	GDS4766	GDS4093
At least 1 SPS	0.85	0.00	0.33	0.94	1.00	0.61	0.82
No SPS	0.33	0.00	0.07	1.00	1.00	0.47	0.47
High SPS (>5)	0.86	0.00	0.43	0.90	1.00	0.67	0.95
Low SPS (<=5)	0.56	0.00	0.11	1.00	1.00	0.48	0.52
Overall	0.69	0.00	0.25	0.96	1.00	0.56	0.71

(ii)

SPS min pval	0.00	0.00	0.00	0.01	0.04	0.01	0.03

Drug Discovery Today

FIGURE 3

Published signatures with more super-proliferation set (SPS) genes are less likely to fail. (a) Signatures with fewer 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 data sets. (b) Proportion of signatures that do better than SPS (i) and the SPS min *P* value PC1–10 (ii).

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 spectrum of independent data sets. If the signature works well across all data sets, it is likely to be useful, and we should be less worried about its significance resulting from chance or its being outperformed in a data set by randomly generated signatures.

Concluding remarks

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

Author contributions

W.W.B.G. and L.W. co-designed the methodologies and co-wrote the manuscript.

Acknowledgments

W.W.B.G. and L.W. acknowledge Vincent de Tours and his colleagues for codes and data obtained from their publication. L.W. gratefully acknowledges support by a Kwan-Im-Thong-Hood-Cho-Temple chair professorship.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2018.08.002.

References

1 Goh, W.W.B. and Wong, L. (2018) Dealing with confounders in omics analysis. *Trends Biotechnol.* 36, 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, 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, 116–126
5 Patil, P. *et al.* (2015) Test set bias affects reproducibility of gene signatures. *Bioinformatics* 31, 2318–2323
6 Wang, W. *et al.* (2016) Feature selection in clinical proteomics: with great power comes great reproducibility. *Drug Discov. Today* 22, 912–918
7 Venet, D. *et al.* (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* 7, 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, 2783–2790
9 Whitfield, M.L. *et al.* (2006) Common markers of proliferation. *Nat. Rev. Cancer* 6, 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, R65
11 Braga-Neto, U.M. and Dougherty, E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 20, 374–380
12 Qin, L.X. *et al.* (2016) Cautionary note on using cross-validation for molecular classification. *J. Clin. Oncol.* 34, 3931–3938
13 Soneson, C. *et al.* (2014) Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One* 9, e100335
14 Goh, W.W.B. *et al.* (2017) Can peripheral blood-derived gene expressions characterize individuals at ultra-high risk for psychosis? *Comput. Psychiatry* 1, 168–183
15 Giuliani, A. (2017) The application of principal component analysis to drug discovery and biomedical data. *Drug Discov. Today* 22, 1069–1076
16 Goh, W.W. *et al.* (2017) Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* 35, 498–507
17 Halsey, L.G. *et al.* (2015) The fickle P value generates irreproducible results. *Nat. Methods* 12, 179–185

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 authors.*