Why breast cancer signatures are no better than random signatures explained

Goh, Wilson Wen Bin; Wong, Limsoon

2018

Goh, W. W. B., & Wong, L. (2018). Why breast cancer signatures are no better than random signatures explained. Drug Discovery Today, 23(11), 1818-1823. doi:10.1016/j.drudis.2018.05.036

https://hdl.handle.net/10356/137544

https://doi.org/10.1016/j.drudis.2018.05.036

Breast Cancer Signatures Are No Better Than Random Signatures Explained

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 (34 words)

Random Signature Superiority (RSS) is the phenomenon where random gene signatures outperform known signatures in any given dataset. It is an under-explored problem, and understanding it is imperative for better design of diagnostic signatures.

ABSTRACT (113 words)

Random Signature Superiority (RSS) occurs when random gene signatures outperform published/known signatures. Unlike reproducibility and generalizability issues, it is relatively under-explored. Yet, understanding it is imperative for better analytical outcome. In breast cancer, RSS correlates strongly with enrichment for proliferation genes and signature size. Removal of proliferation genes from random signatures greatly reduces the predictive power of random signatures. It is noteworthy that almost all genes are somewhat correlated with the proliferation signature, making complete elimination of its confounding effects impossible. RSS goes beyond breast cancer, as it also exists in other diseases; it is especially strong in other cancers in a platform-independent manner, and less severe but present nonetheless in non-proliferative diseases.

KEYWORDS

Bioinformatics; Statistics; Feature selection; Biomarker; Diagnostics; Confounder; Cancer

INTRODUCTION

Statistical feature selection on high-throughput (-omics) data is important for biomarker research [1-3]. Existing methodologies are imperfect, resulting in irreproducible [4,5] and non-generalizable signatures [6]. These issues are well known, and a third but no less important problem has arisen from the work of Venet et al., where known breast cancer signatures are demonstrably no more predictive than random signatures [7]. We term this phenomenon Random Signature Superiority (**RSS**).

Genes do not function independently and are deeply correlated/entangled as a complex system (deep correlation) [8-10]. When normal and cancer samples are compared, they differ across a large multitude of genes, of which many are probably non-causal but correlated to phenotype (large effect size) [11,12]. Notably, the strongest difference is observed in the expression levels of genes that control cell proliferation. However, this difference also exists between any two tissues with different proliferation rates [13]. Proliferation is not a cause but, certainly, part of the pathogenetic mechanism in cancer, and it is also involved in many other processes aside from cancer. Statistical significance alone does not imply phenotypic relevance or confer directionality of association [5].

RSS is deeply confounded with proliferation genes including cell-cycle, cell death, contact-based growth inhibition, etc. While proliferation is a strong predictor for survival outcome, there are thousands of genes qualified (based on cross-tissue correlation) to be included into the proliferation signature. In fact, most genes are to an extent, correlated with proliferation (deep correlation), and it is difficult to fully eliminate this confounding factor (at least 50% of the breast cancer transcriptome is correlated with meta-PCNA, a known proliferation signature) [13]. In published cancer signatures, removal of proliferation genes impacts their predictive power [7], explicit removal of proliferation genes from random signatures should likewise impact RSS. It is likely that RSS is driven strongly by proliferation, although the influence of other confounding factors cannot be discounted.

It is fathomable RSS exists in other diseases. Therefore, it is important to know if this is true and if RSS affect other prediction goals, e.g. class-label assignment. This has wide impact on clinical applications including gene-signature inference and biomarker development.

Factors influencing RSS

There are several factors influencing RSS: the size of the signature under evaluation, its overall prediction power, its association with confounder (it is known that direct removal of proliferation from known signatures can obliterate the predictive power of some signatures [7]), and

the composition of the random signatures (i.e., whether or not it contains high proportion of confounders).

On the signature itself, the higher the proportion of proliferation genes in a known signature, the more significant its association with survival regardless of signature size (Figure 1A). Removal of proliferation from these signatures can have drastic effects on predictive power (Figure 1B).

Removal of proliferation genes do not always produce same effects. It is possible that they are not all born equal. Using a signature with little predictive power (SORLIE; 15 genes; no overlaps with proliferation and a very insignificant predictive $\log_{10}$ p-value of -0.033), additive effects of proliferation genes in general (Figure 1C A1 to A20), and a select subset of proliferation genes based on Venet et al.'s top three signatures (Figure 1C S1 to S20) may be evaluated. It is apparent that amongst the latter set of proliferation genes, additive effects exist. This additive effect is significantly muted in the general set.

Unsurprisingly, removal of proliferation genes from random signatures diminishes RSS. (Figures 2A and 2B). The dramatic decrement of random signature p-values shows that the presence of proliferation genes in random signatures is a major confounder. It is also more pronounced in large signatures (size of random signatures equals size of signature under consideration), although the proportion is around 8-10% irrespective of signature size, the median increases corresponding to signature size (Figure 2C and 2D).

Splitting all random signatures into those that do not contain proliferation genes (No Proliferation, or NP), and those that do (Proliferation, or P), those containing proliferation genes generally are more significant (Figure 3A). A chi-square test with Yate's continuity correction on a frequency-based contingency table (splitting the rows by a statistical threshold of p-value < 0.05, and the columns by whether proliferation genes are observed or not) produces similar results (p-value < 2.2e-16) (Figure 3B). Additionally, the odds ratio reveals that NPs are 1.44X more likely to be non-significant and ~2X less likely to be significant than Ps. However, there is a size-dependency, this differential effect is strongly observed in large but not small random signatures (Figure 3C).

RSS is more likely when random signature size is large. But this is simply due to increased likelihood of incorporating proliferation genes. A small random signature has a less than 50% chance

to be significant (the median lines are on the right or close to the $\log_{10}(0.05)$ mark) whereas for a large random signature, we are guaranteed that more than 50% of the time, it will be significant (Figure 2A). Even after adjustment for proliferation (Figure 2B), the outcome that a random signature is significant is close to 50%.

Powerful non-cancer signatures

Venet et al. stated that irrelevant signatures (e.g. for predicting post-prandial laughter on peripheral blood mononuclear cells, skin fibroblast localization and social defeat in mice) [7] can also predict breast cancer survival outcome. A re-test suggests this is likely due to signature size and presence of proliferation signature (same factors as RSS). Interestingly, all three signatures do not contain proliferation genes, and the signature sizes are 26, 163 and 259. The smallest signature is likely an anomaly (due to sheer chance or incorporation of unknown confounders) while the last two signatures are large, it seems a viable explanation why these large irrelevant signatures do well is due to sheer size.

However, evidence suggesting irrelevant signatures are powerful appears possibly anomalous. Two moderate-sized non-cancer signatures (Schizophrenia studies of Schwarz et al. [14] and Hess et al. [15]), with almost no overlap with proliferation, clearly lacks predictive power ($\log_{10}P$= -0.345 and 0.352 respectively). It is thus unlikely signatures from other contexts are powerful, unless they have strong overlaps with confounders.

Random signature superiority is a universal problem

Since RSS was demonstrated in the context of breast cancer gene expression, it is interesting to know how pervasive the problem is (Figure 3D). We considered renal cancer (RCC) [16], analyzed based on next-generation sequencing (RNA-Seq) and DIA-SWATH-based proteomics; and two microarray datasets on Duchenne's Muscular Dystrophy (DMD), which unlike cancer, has a relatively less complex etiology [17,18]. RSS exists in RCC (another cancer), and although first described in microarray-based gene expression, it occurs also in proteomics (Over 85% of random signatures containing 25 genes are as accurate as the corresponding 25-gene observed signature for SWATH-RCC). In contrast, RSS is only mildly present in DMD (10% of random signatures containing 25 genes are as accurate as the corresponding 25-gene observed signature for Haslett-DMD). This suggests RSS

is a pervasive phenomenon---it is strongly present in diseases with a clear proliferative basis, and mildly present in others where proliferation and complex genotype upheavals are absent.

As observed earlier, signature size influences RSS. Increasing signature size (from 25 to 100), or relaxing the statistical threshold (from 0.01 to 0.05) allows random signatures to become more powerful, resulting in larger RSS p-values (i.e., the proportion of random signatures having better nominal p-value than the observed signature in question) (Figure 3E). However, in these two instances, removal of proliferation genes will not have any effect. In breast cancer, the proliferation signature is associated with survival, but here, we are demonstrating the presence of RSS in class-label prediction (and the class labels are not about survival), of which the proliferation signature is only a part of perhaps many other differential genes useful for the respective class-label prediction tasks.

Recommendations

The Cox survival analysis has its null hypothesis as H_0 = "The survival curves of the patient classes induced by the observed signature is no different from each other". Venet et al. [7] suggest to also ensure that an observed signature is no worse than 5% of random signatures of its size. This corresponds to an actual null hypothesis H_0' = "The difference between the survival curves of the patient classes induced by the observed signature is no different from the difference between the survival curves of the patient classes induced by random signatures". H_0' admits all random signatures as null samples. However, two kinds of null samples are potentially confounded and need to be excluded as null samples. The first we already discussed: random signatures containing proliferation genes are confounded (Figure 2A/B). A second scenario is that random signatures with a large overlap with the observed signature are also likely to be as significant as the observed signature. A null hypothesis that takes this into account is H_0'' = "The difference between the survival curves of the patient classes induced by the observed signature is no different from the difference between the survival curves of the patient classes induced by random signatures which have no overlap with observed signature".

As recommendation, the effect of proliferation genes can be studied on top of such reduced random signatures as well, and to see whether our conclusions are affected. For example, if (i) the observed signature S has no proliferation genes, (ii) has significant Cox p-value, and (iii) more than 5% random signatures not overlapping S are also significant, (iv) but few (i.e. less than 5%) of random

signatures not overlapping S and not containing proliferation genes are significant, then one can conclude that the significance of random signatures in (iii) is due to confounding by proliferation genes, and thus can be safely ignored. Then one has good reason to believe that S is truly significant. To demonstrate this effect, we excluded inclusion of signatures genes in randomly generated non-proliferation signatures (Supplementary Figure 1), the median bars have all shifted further to the left. This indicates that an even larger proportion of random signatures sans proliferation- and signature-genes have significant p-values. Moreover, the actual/real signatures (i.e. red dots) are now mostly not among the top 5% most significant random signatures. This strongly further confirms the RSS effect.

For diagnostic signatures, it is unnecessary to deliberately remove proliferation genes. Diagnostic signatures just need to be consistent, sensitive, and specific. A possible exception is perhaps for early diagnostic, where in this case we do want signals from the early stage (or even pre-disease stage). However, for signatures intended for enhancing mechanistic understanding, especially causal factors, removing genes contributing to RSS (e.g. proliferation genes for cancer) is important. In this case, the development of data-processing methods able to reduce the contributions of such genes may be useful.RSS is particularly pronounced in diseases associated with proliferation. It is milder in diseases (even complex ones) that are not. Nonetheless, the RSS should not be ignored even in the mild case, as some 10% of random signatures containing 25 or more genes are performing as well as or better than observed/reported signatures of similar size.

Rejection of the null hypothesis (significant p-value) is insufficient. During statistical analysis, the null hypothesis assumes a background of non-association with phenotype. Since a random signature is more likely correlated with outcome, it implies that some assumption of the null hypothesis is incorrect, and we cannot rely on this approach to identify truly powerful signatures. Thus, during comparative analysis, if we find a signature with significant p-value, it does not mean that genes in the signature are necessarily causative or relevant to the disease/phenotype. This is a manifestation of the Anna Karenina Effect where association with a variety of confounders can easily lead to rejection of the null hypothesis[19-22].

CONCLUSIONS

RSS is an under-examined area, and yet, understanding how and why it happens, is imperative for better approaches towards the development of diagnostic signature. In breast cancer,

RSS is confounded with signature size and the presence of the proliferation signature, for which a large part of the breast cancer transcriptome is correlated with. RSS also exists in other diseases; it is present strongly in other cancers but less pronounced in non-proliferative diseases. Finally, while non-cancer signatures can be predictive, this is simply due to confounding or chance.

FUNDING

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

AUTHOR CONTRIBUTION

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

ACKNOWLEDGEMENT

WWBG and LW acknowledge Vincent de Tours, and his colleagues for codes and data obtained from their publication, and Professor Teo Tian Seng for his insights on proliferation genes.

COMPETING INTERESTS

The authors declare no conflicting interests, financial or otherwise.

REFERENCES

1 Goh, W.W. and Wong, L. (2016) Design principles for clinical network-based proteomics. *Drug Discov Today* 21 (7), 1130-1138

2 Goh, W.W. and Wong, L. (2016) Integrating Networks and Proteomics: Moving Forward. *Trends in Biotechnology* 34 (12), 951--959

3 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

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

6 Goh, W.W.B. and Wong, L. (2016) Evaluating feature-selection stability in next-generation proteomics. *Journal of Bioinformatics and Computational Biology* 14 (5), 1650029

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 Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science* 286 (5439), 509-512

9 Barabasi, A.L. et al. (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet* 12 (1), 56-68

10 Bensimon, A. et al. (2012) Mass spectrometry-based proteomics and network biology. *Annu Rev Biochem* 81, 379-405

11 Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell* 100 (1), 57-70

12 Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell* 144 (5), 646-674

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

14 Schwarz, E. et al. (2012) Identification of a biological signature for schizophrenia in serum. *Mol Psychiatry* 17 (5), 494-502

15 Hess, J.L. et al. (2016) Transcriptome-wide mega-analyses reveal joint dysregulation of immunologic genes and transcription regulators in brain and blood in schizophrenia. *Schizophr Res* 176 (2-3), 114-124

16 Guo, T. et al. (2015) Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nature Medicine* 21 (4), 407-413

17 Haslett, J.N. et al. (2002) Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle. *Proc Natl Acad Sci U S A* 99 (23), 15000-15005

18 Pescatori, M. et al. (2007) Gene expression profiling in the early phases of DMD: a constant molecular signature characterizes DMD muscle from early postnatal life throughout disease progression. *FASEB J* 21 (4), 1210-1226

19 Bornmann, L. and Marx, W. (2011) *The Anna Karenina principle: A mechanism for the explanation of success in science*

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

21 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

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

FIGURE LEGENDS

Figure 1 (A) Regardless of signature size, if a known signature contains more proliferation genes, it is likely to be stronger. Y-axis: $\log_{10}$(p-value) associated with survival outcome. X-axis: Proportion of proliferation genes in known signature. Inset: correlation (cor) and its associated p-value. (B) The presence of proliferation genes in actual signatures have a very strong determination on its performance. Many signatures with a very high proportion of proliferation genes lose their power when the proliferation genes are removed. This trend is observable in both small and large signatures (c.f. Supplementary Figure 1). Y-axis: delta $\log_{10}$(p-value) associated with survival outcome. X-axis: Proportion of proliferation genes in known signature. Inset: correlation (cor) and its associated p-value. (C) Proliferation Genes sampled from the strongest signatures, 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)

Figure 2 Removal of proliferation genes from random signatures results in dramatic shift towards non-significance, with a clear size dependency. (A/B) Distributions of random signatures before and after removal of proliferation genes. Y-axis: Signatures ordered by increasing size. X-axis: nominal $\log_{10}$(p-value) of association with survival. Red dots are known breast cancer signatures, yellow is the distribution of 1,000 random signature p-values, with its median indicated by a vertical black line. Lower 5% quantile is shaded green to the left of the corresponding vertical short black bar. Since actual signatures retain their original composition (with proliferation genes), they remain unchanged. (C) Distribution of proliferation genes in random signatures. Y-axis: Proportion of proliferation genes. X-axis: Signatures are arranged from smallest (PEI) to largest (HUA)

as in panels A/B. (D) Median of proliferation genes in random signatures. Y-axis: Median number of proliferation genes in random signatures of a given size of proliferation genes. X-axis: Signatures are arranged from smallest (PEI) to largest (HUA) as in panels A/B.

Figure 3 (A) Significant random signatures are likely to contain at least one proliferation gene (c.f. Supplementary Figure 2C/D). (B) Contingency table showing the number of significant and non-significant random signatures with and without at least one proliferation gene. Odds ratios suggest that signatures without proliferation genes are less likely to be significant. (C) Stratification by large and small signatures reveals that signature size is a strong determinant of whether a random signature will be significant. (D) RSS has a size dependency, and the larger the random signature size, the more likely random signatures will beat the observed signature. Y-axis: RSS p-value, which is the proportion of random signatures having better nominal p-value than the observed signature in question. X-axis: random signature size. (E) RSS based on two statistical thresholds, nominal p-value < 0.01 and 0.05. Y-axis: RSS p-value. X-axis: statistical threshold used for determining the random signature size.

Breast Cancer Signatures Are No Better Than Random Signatures Explained

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. [4] are used here. Signatures are tested for correlation with outcome (survival status) where the lower the p-value, the higher the association (see below).

Renal cancer transcriptomics and proteomics datasets

To demonstrate that Random Signature Superiority (RSS) also occurs in other cancers, in a platform-independent manner, two published datasets on clear cell renal carcinoma (ccRCC)---a transcriptomic dataset from the TCGA [16] and a proteomics dataset from the study of Guo et al. [17]---are included here.

Duchenne muscular dystrophy microarray datasets

To determine whether RSS also happens in non-proliferative diseases, two published microarray datasets on DMD (muscle tissues) based on the studies of Haslett et al. [18] and Pescatori et al. [19] are used here.

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. [11], and the meta-PCNA, which is a list of 129 genes most correlated to the PCNA gene.

We consider the union of Prolif and meta-PCNA (Prolif+ meta-PCNA, 1,061 genes) as well as separately, to determine enrichment in known and random signatures.

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. [4], 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/) [23].

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.

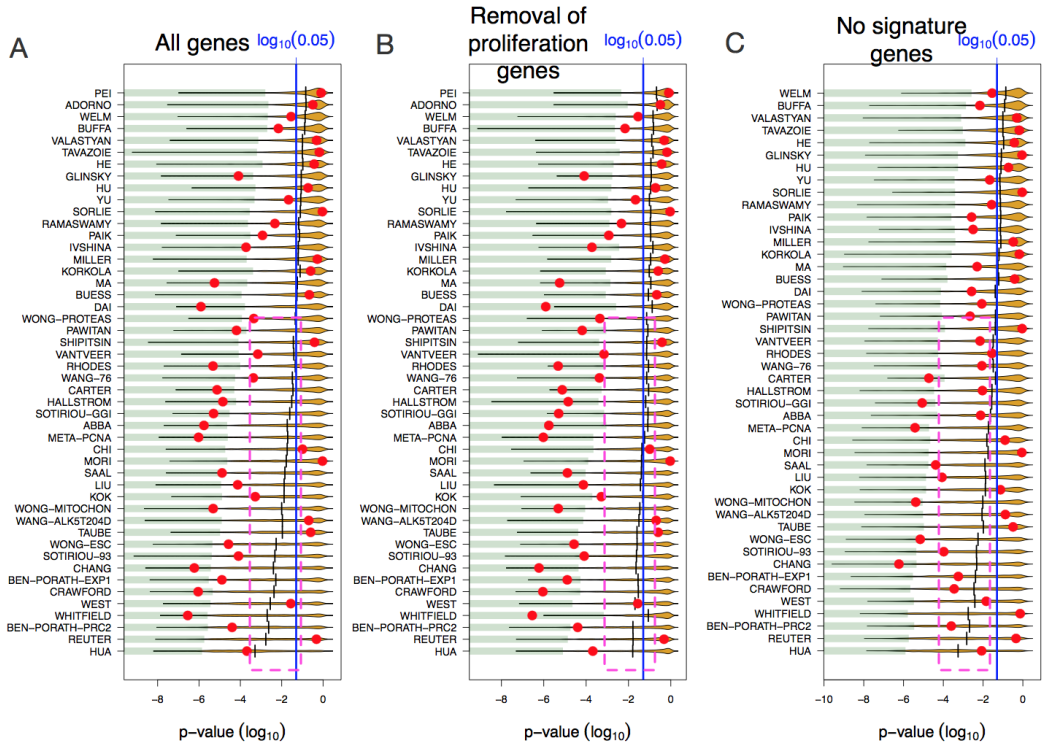
Cross-validation based benchmarking

For non-breast cancer datasets, we do not have survival information or extensive signature sets from prior studies. Instead, we check for RSS based on class-label prediction.

Given a fixed random split of the data into training and validation sets, two scenarios are considered: random signatures of fixed sizes 25, 50 and 100, and fixed statistical thresholds at nominal p-value ≤ 0.01 and 0.05. Random signatures are used for training a Naïve Bayes classifier on the class labels of the training set. The cross-validation accuracy is the proportion of correctly predicted class labels on the validation set.

The accuracies based on random signatures are used to form an empirical null distribution. The actual accuracy is the cross-validation outcome based on the inferred signature in the training set (based on the ranked t-statistic for fixed sizes, or nominal p-value cutoff for statistical thresholds). The RSS p-value is defined as the proportion of random signatures that beats or is equal to the observed accuracy.

RESULTS



Supplementary Figure 1 Removal of confounders from random signatures results in dramatic shift towards non-significance, with a clear size dependency. **(A)** Distributions of random signatures before removal of proliferation genes. **(B)** Distributions of random signatures after removal of proliferation genes. **(C)** Distributions of random signatures when signature genes are also explicitly excluded. (**A-C**) Y-Axis: Signatures ordered by increasing size. X-axis: nominal $\log_{10}$(p-value) of association with survival. Red dots are known breast cancer signatures, yellow is the distribution of 1,000 random signature p-values, with its median indicated by a vertical black line. Lower 5% quantile is shaded green to the left of the corresponding vertical short black bar. Since actual signatures retain their original composition, they remain unchanged.

Here, we observed reported signature S against random signatures that (i) are of the same size as S and (ii) have no overlap with S. If lots (i.e. more than 5%) of such random signatures are significant, then the observed signature does not contain more meaningful information than random ones (and

the significance of these random ones cannot be attributed to their containing similar information as the observed signature.) On the other hand, if few (i.e. less than 5%) of such random signatures are significant, then the observed signature may indeed contain more meaningful information than random ones that do not contain similar information as the observed signature.