Breast Cancer RESEARCH

VIEWPOINT

Prognostic signatures in breast cancer: correlation does not imply causation

Charlotte Ng[†1], Britta Weigelt[†2], Anita Grigoriadis[3] and Jorge S Reis-Filho[*1]

Abstract

Testing the statistical associations between microarray-based gene expression signatures and patient outcome has become a popular approach to infer biological and clinical significance of laboratory observations. Venet and colleagues recently demonstrated that the majority of randomly generated gene signatures are significantly associated with outcome of breast cancer patients, and that this association stems from the fact that a large proportion of the transcriptome is significantly correlated with proliferation, a strong predictor of outcome in breast cancer patients. These findings demonstrate that a statistical association between a gene signature and disease outcome does not necessarily imply biological significance.

Breast cancer encompasses a plethora of distinct diseases characterised by different biological features and clinical outcomes [1-3]. Microarray-based gene expression profiling studies have played a pivotal role in unravelling the molecular and clinical diversity of the disease (for a review see [3]). These studies led to the development of a molecular classification of breast cancer [4], where the different molecular subtypes identified were found to be associated with distinct clinical outcomes [5,6], and to the development of numerous multigene predictors (that is, gene signatures) of outcome, which were initially reported to outperform the current clinicopathological algorithms to define the prognosis of breast cancer patients [7,8] (reviewed in [3,9]).

Microarrays have also played a pivotal role in addressing one of the major bottlenecks in translational research: ascribing relevance in the human disease context of results obtained from *in vitro* studies and animal models. The availability of multiple gene expression datasets with patient follow-up in the public domain allowed the investigation of whether a microarray-based signature derived from a set of laboratory experiments would have biological significance. For instance, a signature derived from tumour-initiating breast cancer cells was shown to be of prognostic significance in a publicly available microarray dataset, and this was used as the basis to suggest that the tumourigenic breast cancer cell signature 'may detect transcriptional profiles associated with mutations that arrest cells in an immature state of differentiation and function as markers of more aggressive tumors' [10].

In their recent paper [11], Venet and colleagues made the intriguing observation that gene signatures developed to identify phenomena completely unrelated to cancer – such as the effect of postprandial laughter on peripheral blood mononuclear cells, the localisation of skin fibroblasts or social defeat obtained from mice brains – were significantly associated with outcome in a cohort of 295 breast cancer patients of the Netherlands Cancer Institute (NKI-295) [8]. In addition, it was also shown that, out of 1,890 gene signatures deposited in the Molecular Signatures Database, 67% were associated with breast cancer outcome at P <0.05, and 23% were associated at P <10^{-5}. The large number of signatures significantly associated with outcome may be due to the enrichment of the Molecular Signatures Database with cancer-related signatures; hence the authors generated for each Molecular Signatures Database signature a signature of identical size but composed of randomly selected genes. Strikingly, out of these randomly derived signatures, 77% were associated with outcome at P <0.05 and 30% were associated at P <10^{-5}. Furthermore, the authors went on to show that only 18 of the 47 published prognostic signatures that were either derived for the purpose of finding better prognostic tools or, in most cases, were used to suggest biological relevance of laboratory findings performed statistically better than the best 5% of random gene signatures of the same size [11].

A critically relevant set of observations made by Venet and colleagues include the fact that >90% of randomly

[†]These authors contributed equally
*Correspondence: jorge.reis-filho@icr.ac.uk
[1]The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK
Full list of author information is available at the end of the article

generated signatures containing >100 genes were shown to be associated with outcome of breast cancer patients [11]. Further, up to 26% of all probes within the microarray platform used for the analysis of the samples from the NKI-295 dataset were significantly associated with outcome on univariate analysis. Even when more stringent parameters (that is, the q value) to account for the false discovery stemming from multiple comparisons were used, 17% of all probe sets were shown to be significantly associated with outcome [11]. What are the statistical and/or biological reasons for these observations?

Given that previous studies had revealed that proliferation is the main and shared determinant of the prognostic accuracy of multigene predictors of outcome in breast cancer patients [3,12-14], the authors developed a proliferation metagene called meta-PCNA. This metagene was composed of the top 1% of genes whose expression was most positively correlated with the expression of the proliferating cell nuclear antigen (PCNA) across 36 normal tissues. Venet and colleagues confirmed that proliferation is a major prognostic determinant of outcome in unstratified breast cancer patients [11]. meta-PCNA was then used to adjust the expression data of breast cancer gene signatures, which resulted in a dramatic reduction in the association between most published and random signatures and outcome.

So why do random gene signatures with >100 genes correlate with breast cancer patient outcome? The crux of the problem appears to be the large number of proliferation-related genes in the breast cancer transcriptome itself, given that the authors found that 58% of the microarray probes used for the analysis of the NKI-295 dataset were correlated with meta-PCNA [11]. Virtually any large collection of genes will therefore inevitably be enriched for proliferation-related genes. Moreover, given that there are many genes whose expression levels correlate with cell cycle and/or proliferation but whose main biological functions/gene ontology may not be related to these phenomena, any attempt to remove known proliferation-related genes as defined by gene ontology are likely to be futile [11]. While this does not imply that the published signatures do not have prognostic value, the underlying unifying feature among them is the effect of proliferation and the signal of additional biological relevance beyond this is minimal.

Arguably, one of the major contributions of Venet and colleagues was to bring to the attention of the breast cancer research community the limitations of an approach ever so familiar in this day and age: using microarrays to suggest that a mechanism is relevant to human breast cancer from the finding that a gene expression marker for this mechanism predicts outcome of breast cancer patients [11]. Their study has also reminded us of the old maxim that 'correlation does not imply causation'. The assessment of the expression levels of a gene or gene signature may be clinically useful without yielding interesting biological or mechanistic insights. On the other hand, an association between a gene signature derived from laboratory experiments and the prognosis of breast cancer patients does not necessarily imply that the genes which compose a given signature are of biological significance to the disease.

Abbreviations

NKI-295, The Netherlands Cancer Institute cohort of 295 breast cancer patients; PCNA, proliferating cell nuclear antigen.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors' work is supported by Breakthrough Breast Cancer. BW is funded by a Cancer Research UK postdoctoral fellowship. The authors acknowledge NHS funding for the NIHR Biomedical Research Centre.

Author details

[1]The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London SW3 6JB, UK. [2]Signal Transduction Laboratory, Cancer Research UK London Research Institute, London WC2A 3LY, UK. [3]Breakthrough Research Unit, Bermondsey Wing, Guy's Hospital, London SE1 9RT, UK.

Published: 19 June 2012

References

1. Colombo PE, Milanezi F, Weigelt B, Reis-Filho JS: **Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction.** *Breast Cancer Res* 2011, **13:**212.
2. Weigelt B, Pusztai L, Ashworth A, Reis-Filho JS: **Challenges translating breast cancer gene signatures into the clinic.** *Nat Rev Clin Oncol* 2012, **9:**58-64.
3. Reis-Filho JS, Pusztai L: **Gene expression profiling in breast cancer: classification, prognostication, and prediction.** *Lancet* 2011, **378:**1812-1823.
4. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
5. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci U S A* 2001, **98:**10869-10874.
6. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, Livasy C, Carey LA, Reynolds E, Dressler L, Nobel A, Parker J, Ewend MG, Sawyer LR, Wu J, Liu Y, Nanda R, Tretiakova M, Ruiz Orrico A, Dreher D, Palazzo JP, Perreard L, Nelson E, Mone M, Hansen H, Mullins M, Quackenbush JF, Ellis MJ, Olopade OI, Bernard PS, *et al.*: **The molecular portraits of breast tumors are conserved across microarray platforms.** *BMC Genomics* 2006, **7:**96.
7. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536.
8. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347:**1999-2009.
9. Weigelt B, Baehner FL, Reis-Filho JS: **The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade.** *J Pathol* 2010, **220:**263-280.
10. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke MF: **The prognostic role of a gene signature from tumorigenic breast-cancer cells.** *N Engl J Med* 2007, **356:**217-226.
11. Venet D, Dumont JE, Detours V: **Most random gene expression signatures**

are significantly associated with breast cancer outcome. *PLoS Comput Biol* 2011, **7:**e1002240.

12. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schutz F, Goldstein DR, Piccart M, Delorenzi M: Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008, **10:**R65.
13. Reyal F, van Vliet MH, Armstrong NJ, Horlings HM, de Visser KE, Kok M, Teschendorff AE, Mook S, van 't Veer L, Caldas C, Salmon RJ, van de Vijver MJ, Wessels LF: A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Res* 2008, **10:**R93.
14. Mosley JD, Keri RA: Cell cycle correlated genes dictate the prognostic power of breast cancer gene lists. *BMC Med Genomics* 2008, **1:**11.

doi:10.1186/bcr3173
Cite this article as: Ng C, *et al.*: Prognostic signatures in breast cancer: correlation does not imply causation. *Breast Cancer Research* 2012, **14:**313.