


TECHNICAL ADVANCE

Open Access



Generation of an algorithm based on minimal gene sets to clinically subtype triple negative breast cancer patients

Brian Z. Ring¹, David R. Hout², Stephan W. Morris², Kasey Lawrence², Brock L. Schweitzer², Daniel B. Bailey², Brian D. Lehmann³, Jennifer A. Pietenpol³ and Robert S. Seitz^{2*} 

Abstract

Background: Recently, a gene expression algorithm, TNBCtype, was developed that can divide triple-negative breast cancer (TNBC) into molecularly-defined subtypes. The algorithm has potential to provide predictive value for TNBC subtype-specific response to various treatments. TNBCtype used in a retrospective analysis of neoadjuvant clinical trial data of TNBC patients demonstrated that TNBC subtype and pathological complete response to neoadjuvant chemotherapy were significantly associated. Herein we describe an expression algorithm reduced to 101 genes with the power to subtype TNBC tumors similar to the original 2188-gene expression algorithm and predict patient outcomes.

Methods: The new classification model was built using the same expression data sets used for the original TNBCtype algorithm. Gene set enrichment followed by shrunken centroid analysis were used for feature reduction, then elastic-net regularized linear modeling was used to identify genes for a centroid model classifying all subtypes, comprised of 101 genes. The predictive capability of both this new "lean" algorithm and the original 2188-gene model were applied to an independent clinical trial cohort of 139 TNBC patients treated initially with neoadjuvant doxorubicin/cyclophosphamide and then randomized to receive either paclitaxel or ixabepilone to determine association of pathologic complete response within the subtypes.

Results: The new 101-gene expression model reproduced the classification provided by the 2188-gene algorithm and was highly concordant in the same set of seven TNBC cohorts used to generate the TNBCtype algorithm (87 %), as well as in the independent clinical trial cohort (88 %), when cases with significant correlations to multiple subtypes were excluded.

Clinical responses to both neoadjuvant treatment arms, found BL2 to be significantly associated with poor response (Odds Ratio (OR) = 0.12, $p = 0.03$ for the 2188-gene model; OR = 0.23, $p < 0.03$ for the 101-gene model). Additionally, while the BL1 subtype trended towards significance in the 2188-gene model (OR = 1.91, $p = 0.14$), the 101-gene model demonstrated significant association with improved response in patients with the BL1 subtype (OR = 3.59, $p = 0.02$).

Conclusions: These results demonstrate that a model using small gene sets can recapitulate the TNBC subtypes identified by the original 2188-gene model and in the case of standard chemotherapy, the ability to predict therapeutic response.

Keywords: Translational oncology, Breast cancer, Biomarkers, Statistical analysis

* Correspondence: rseitz@insightgenetics.com

²Insight Genetics Incorporated, Nashville, Tennessee, USA

Full list of author information is available at the end of the article



Introduction

TNBC has a higher rate of 5-year distant recurrence than other breast cancers, and despite adjuvant chemotherapy as standard of care for this cancer, 5-year recurrence rates are around 30 % [1]. Those patients that achieve a pathological complete response (pCR) to neoadjuvant chemotherapy have significantly better overall survival [1, 2]. Furthermore, the correlation between pCR and distant recurrence is considerably stronger within TNBC patients compared to ER+ patients [3] leading the Food and Drug Administration to allow pCR as a clinical endpoint for TNBC while strongly recommending against it in ER+ patients [4]. Many studies have established that breast tumors are heterogeneous, both in histology and clinical outcome, and these differences can serve as the basis for clinical classification and prognostication [5]. Additionally, molecular classification of cancer subtypes is becoming an increasingly important tool in devising treatment plans. For example, mutation analysis of KRAS in colorectal cancer [6], and EGFR mutation and ALK rearrangement detection in non-small cell lung cancer [7, 8] are now standard of care.

There currently exist no clinically applied molecular subclassification tools for TNBC. The intrinsic breast cancer classification system [9], which has proven useful in assigning biological information to breast cancer subclasses, categorizes the majority of TNBC cases within the basal subclass [10]. However, significant heterogeneity – both clinically and pathologically – exists in TNBC, and better subclassification tools are needed for clinical decision-making. To this end, Lehmann et al. used 21 breast cancer data sets containing 587 TNBC cases and employed cluster and gene expression analysis to identify a set of 2188 genes for the classification of TNBC into six subtypes (two basal-like (BL1 and BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL), and luminal androgen receptor (LAR) subtypes) [11]. The goal is of this study to translate the knowledge of biologically distinct subtypes into rational design of pre-clinical studies for TNBC clinical trials and to facilitate the identification of novel predictive markers. A previous study has shown the promise of clinical utility by retrospectively subtyping 130 TNBC patients who had received standard neoadjuvant chemotherapy comprised of anthracycline, cyclophosphamide and a taxane. This study showed that patients with basal-like BL1 tumor subtypes had an improved response (52 % exhibiting pCR) whereas basal-like BL2 tumor subtypes showed a worse response to standard chemotherapy (0 % pCR) [12].

In the derivation of the TNBCtype subclassification tool, the final group of 2188 classifying genes was identified from an initial set of approximately 13,000 genes by selection of those genes with expression significantly distinct from the median gene expression among all the

cluster-defined subclasses. Although a seminal advance for the TNBC field, this large classification panel could best be applied for clinical use only after further refinement. Such optimization would serve three purposes. First, a more limited set of genes would speed translation of the classifier into a cost-effective clinical tool. Second, although not necessarily the case [13], the genes most predictive of a subtype may include those most relevant to the regulation and function of that subtype; therefore, a smaller set of genes may increase the likelihood of correlating biological meaning to the panel members and allowing easier comparison of TNBC subtype molecular profiles to other similarly well-defined molecular prognostic and predictive tools. Third, and most importantly, a small set of classifying genes could improve the reproducibility of the TNBC subtyping panel.

Initial gene expression analysis often has the problem of inclusion of genes with little signal contribution. This problem arises from having tens of thousands of genes produced in a typical assay but a considerably smaller number of measured samples within which to assess these potential classifiers. This statistical problem, coupled with the inherent noise of microarray platforms, creates a challenge to the derivation of reproducible classification panels. One study estimated that to achieve similar (i.e., overlapping) gene panels in multiple cohorts, the number of analyzed tumor samples would need to be at least several thousand [14]. It is well established that overfitting occurs with large-scale gene expression data when poor or no feature or dimensional reduction is attempted [15, 16]. Careful reduction of the genes included in a TNBC classifier would potentially lead to a robust clinically applicable tool for subtype identification.

Here we describe the development and validation of a new TNBC classification tool using only 101 genes, less than 5 % of the size of the original 2188 gene model of TNBCtype and yet able to reproduce its classification. The association of the BL1 and BL2 subtypes with pathologic response was also reproduced in an independent patient cohort using this new model.

Methods

Gene expression data set processing

Twenty-one expression data sets representing 13,060 unique genes that were previously used to develop (14 data sets; $N = 386$ patients) and validate (7 data sets; $N = 201$ patients) the 2188-gene TNBC classification model were used as prepared for the published analysis (Robust Multi-array Average (RMA) normalized, log transformed) [11]. Expression data of an additional breast cancer data set (TNBC: GSE41998) was downloaded along with clinical metrics including age, menopausal status, and outcome as measured by pCR. Patient datasets were previously made publicly available under the ethical policies of the

National Institutes of Health's GEO database (<http://www.ncbi.nlm.nih.gov/geo/info/submission.html>). No additional ethics review was required for the *in silico* analysis of these data sets.

As with the Lehmann et al. analysis, when multiple probes for a gene were present, the probe with the highest inter-quartile range was selected. Triple-negative status in the GSE41998 breast cancer samples was determined by the given pathological diagnosis ($N = 139$ patients), with an additional seven cases being excluded because bimodal modeling of ER, PR, and ERBB2 expression gave posterior probabilities greater than 0.5 that these genes were expressed. The original 2188-gene centroid classifier, five individual subtype classification models, and a 101-gene centroid classifier were applied to this patient set.

Statistical analysis and model building

The TNBC subtype signatures from the original 2188-gene model were used to identify gene sets that distinguished the classes via gene set enrichment analysis (GSEA) using the C2 curated gene sets of canonical pathways [17]. Linear regression, targeted maximum likelihood estimation [18], random forest [19], and elastic-net regularized linear models [20] were employed to create subclassifying models, with each subclass (i.e., BL1, BL2, LAR, M, MSL, and IM) being defined by an individual model. Strength of association with outcome variables was assessed using logistic regression and the Fisher exact test. Classification error was estimated using a bootstrap analysis, and the elastic net models showed the least error (average disagreement of 6 % for all five models). The genes that contribute to the five individual subtype models were combined to create a 101-gene centroid model. All model coefficients and cutoffs were determined using the 14 discovery data sets, as in the original Lehmann et al. analysis, and were not altered afterwards. Pathway analysis of the 258 shrunken centroid defined genes was performed with Cytoscape using the ClueGO tools [21, 22]. All p -values are two-sided.

Results

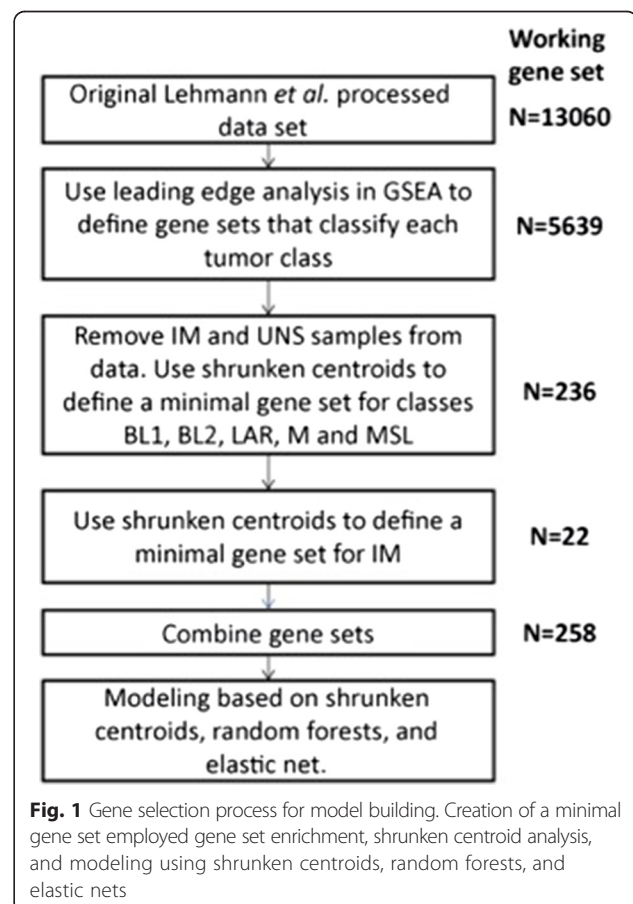
Building limited gene models of TNBC subtypes

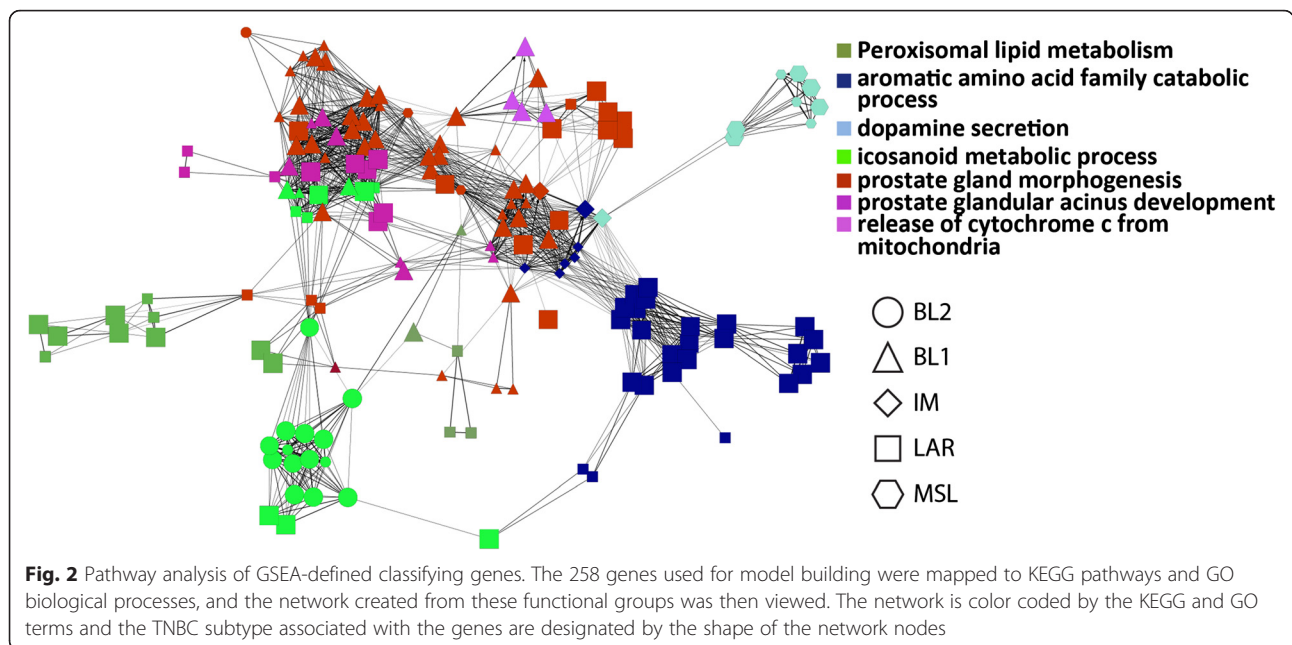
The expression data sets used to generate the original 2188-gene model were assigned TNBC subclass identities based on the Lehmann et al. results. Gene set enrichment analysis [17] was performed on the 14 training gene sets and 5639 genes were identified as belonging to pre-defined gene sets that associate with the TNBC subclasses. Given previous observations that tumor infiltrating lymphocytes (TILs) correlate with increased expression of genes involved in immune response [23], the 'Immunomodulatory' (IM) subtype likely reflects the presence of gene expression contributed by immune infiltrates with the tumor cells

having the signature of a different subtype. Therefore we performed principal component analysis (PCA) to identify and remove the IM component. The presence of an IM component almost completely defined the IM class (data not shown), and its significant association with other classes caused a significant loss of information. Therefore, cases assigned an IM identity were excluded and analyzed separately.

Additionally, cases not classified by the original TNBC-type were also excluded, as well as cases that a Z-test showed to have non-significant differences between the most highly correlated centroids. Shrunken centroid analysis [24] was used for further feature reduction. Using all non-IM cases, 236 genes were identified as likely classifiers. Analyzing the IM cases compared to all other combined cases identified a further 22 gene classifiers, resulting in 258 genes in total used for subsequent model building (Fig. 1).

Pathway analysis of the shrunken centroid-defined list of 258 genes used for model building and their associated GO and KEGG terms showed biological processes consistent with their putative classification role, which lent confidence to this limited gene list (Fig. 2). Different gene sets and algorithms were used for the initial gene





set enrichment and this pathway analysis, and no supervision was employed over pathways used to define subtypes. As an example, most of the genes associated with the BL1 subclass correlated with the expression of genes previously observed in basal cells [25]. Additionally, genes associated with the LAR subclass mapped to clusters of peroxisomal lipid metabolism and aromatic acid metabolism and catabolism, which matches the functions previously mapped to this subtype [10].

Linear regression, targeted maximum likelihood estimation [18], random forest [19], and elastic-net regularized linear models [20] were employed to create subclassification models, with the latter approach giving the best fit to the TNBCtype-designated subclasses with the least number of required genes. Six elastic net models were created to identify each subtype individually, or an expression-based centroid defined by the genes used in all the elastic net models—101 genes in total. While the 101 genes were selected independently of the original 2188 genes, 96 genes were in common between the two models. These models were then applied to the set of seven, validation cohorts employed in the TNBCtype analysis. The elastic net-defined models showed a predicted misclassification rate of 2–9 % in the discovery set of cohorts in a bootstrap analysis, and 6–17 % in the validation cohorts (Table 1). The 101-gene centroid had a 7 % discordance in the discovery cohort, while in the validation cohort the discordance was 25 % (Table 2A). The centroid model allows a tumor to be assigned to only one subclass, in contrast to the individual models, though some cases show characteristics of multiple subtypes. When cases were excluded that showed a significant correlation to multiple subtypes and

an insignificant (via a Z-test) difference between correlation to these subclasses, the discordance in the validation cohort decreased to 13 % (Table 2B). As the 101 gene centroid model did not require the use of fixed cutoff to classify samples and was thus platform independent, this model was used for the subsequent analyses of this investigation.

Comparison of the original and newly developed subtyping models in an independent cohort of patients treated with doxorubicin/cyclophosphamide followed mitotic inhibitors

A previous study had demonstrated that TNBC molecular subtypes differed in response to neoadjuvant chemotherapy [12]. To determine if subtype classifications provided by TNBCtype and the newly developed limited gene models were concordant, we evaluated an independent cohort of 278 early-stage breast cancer patients, of which 139 were TNBC patients, treated neoadjuvantly with doxorubicin/cyclophosphamide (AC) followed by ixabepilone or paclitaxel [26]. Agreement between the 2188-gene and the 101-gene centroid classifiers was 83 % among cases that had a

Table 1 Misclassification rate

	Lehmann et al. discovery set (N = 386)	Lehmann et al. validation set (N = 201)
BL1	0.07	0.14
BL2	0.06	0.06
LAR	0.02	0.12
M	0.09	0.17
MSL	0.04	0.10

Misclassification rate as estimated by bootstrap analysis of elastic net models in the Lehmann et al. [11] discovery and validation cohorts

Table 2 Comparison of 2188- and 101-gene centroid classifiers in the Lehmann et al. [11] validation cohorts

			TNBCtype (2188 gene centroid model)						
			BL1	BL2	LAR	M	MSL	IM	Unclassified
101 gene centroid model	(A) All significantly classed cases	BL1	25			1	3	16	4
		BL2	1	15	1	2	1	7	2
		LAR		1	14	3	1		
		M	6	1	1	26	2		2
		MSL	2	1	2	4	17	5	2
		Unclassified	2	5		6	1	16	3
	(B) Ambiguous cases unclassified	BL1	19				2	14	3
		BL2	1	12				7	2
		LAR			13	1	1		
		M	2		1	21	1		1
		MSL	1			2	14	4	2
		Unclassified	13	11	4	18	7	19	5

In panel A, all cases are used. In Panel B, ambiguous cases—cases that showed an insignificant difference (via a Z-test) between subclasses—are placed in the unclassified group

significant correlation with at least one subtype. When the comparison was limited to those samples that had a significant difference in the correlations between the highest and second highest subtype, the agreement increased to 86 % (Table 3).

BL1 and BL2 TNBC subtypes differ in pathologic response to mitotic inhibitors

The clinical response of TNBC patients in the GSE41998 cohort to neoadjuvant AC was strongly predictive of pCR after subsequent additional treatment with ixabepilone or paclitaxel ($p < 0.0001$, Table 4). Age was also a very strong predictor of outcome, as was menopausal status (Table 5). Stage was not included in the provided data set and could

therefore not be assessed as a factor. *Also not provided was BRCA1/2 status, a factor previously shown to be predictive of response to certain chemotherapeutic agents* [27]. In previous cohorts, patients with BL1 subtype tumors had better response to chemotherapy and those with BL2 had lower rates of response [12], and similar results were found in this cohort. Patients with BL2 subclass tumors, as defined by either the 2188- or 101-gene models (with confounding calls removed for both cases as described above), had the least response and higher incidence of more than minimal residual disease after therapy (OR = 0.12, $p = 0.03$; OR = 0.23, $p = 0.02$ for the 2188- and 101-gene models, respectively, via a Fisher exact test) (Table 6). The BL1 subclass as defined by the 101-gene model exhibited the best

Table 3 Comparison of 2188- and 101-gene centroid classifiers in the GSE41998 TNBC cohort [26]

			TNBCtype (2188 gene centroid model)					
			BL1	BL2	LAR	M	MSL	Unclassified
101-gene centroid model	(A) All significantly classed cases	BL1	36			1	1	
		BL2	5	12		2	2	
		LAR		1	10		1	1
		M	3	1		25	2	
		MSL				3	21	
		Unclassified	1		1	1	1	1
	(B) Ambiguous cases unclassified	BL1	26					
		BL2	5	11		1	2	
		LAR		1	9			1
		M		1		15	1	
		MSL				1	11	
		Unclassified	14	1	2	15	14	1

In panel A, all cases are used. In Panel B, ambiguous cases—cases that showed an insignificant difference (via a Z-test) between subclasses—are placed in the unclassified group

Table 4 Comparison of clinical response

		pCR/mRCB		pCR	
		No	Yes	No	Yes
Clinical Response to AC	complete response	5	18	8	15
	partial response	36	31	39	28
	stable disease	20	2	20	2
	progressive disease	2	0	2	0

Clinical response to neoadjuvant AC (doxorubicin/cyclophosphamide) and pCR/minimal RCB after subsequent neoadjuvant ixabepilone or paclitaxel in the GSE41998 TNBC cohort [26]

response (OR = 3.59, $p = 0.02$). The direction of association of BL1 with pCR or minimal residual cancer burden (RCB) in the 2188-gene model was similar (OR = 1.91) but did not reach significance ($p = 0.14$). When all cases were examined with the 101-gene model (i.e., including cases that mapped to multiple subtype but were assigned to the subtype with the highest correlation), the BL2 subtype retained a significant association with poor response to therapy while BL1 cases lost significance. When the 2188-gene model was analyzed in this manner, neither BL1 nor BL2 had a significant association with response to therapy, though the effect sizes were similar (data not shown *Whether this is due to superior performance by the 101-gene algorithm or a due to the small size of the cohort cannot be determined without further study.*

Discussion

TNBC comprises up to 20 % of all breast cancers (as many as 40,000 women newly diagnosed in the US each year), and occurs more frequently in young and African-American women [1]. TNBC has higher rates of metastatic recurrence and poorer prognosis than other breast cancers, with a 5-year survival of only ~70 % after treatment with the most aggressive conventional cytotoxic chemotherapies. This current state is due in large part to the heterogeneity of TNBC and the still limited knowledge regarding therapeutic targets and biomarkers that can predict the responsiveness of these cancers to either standard-of-care or investigational therapies. Despite overall poor outcomes, approximately 30 % of TNBC patients respond to standard chemotherapy [1]. Thus, there is a

Table 6 Association of centroid model-determined subtype and pCR

		pCR/mRCB			Odds Ratio	p value
		Yes	No	Percentage		
2188 gene centroid	BL1	20	14	59 %	1.91	0.14
	BL2	1	8	11 %	0.12	0.03
	LAR	3	6	33 %	0.5	0.49
	M	7	12	37 %	0.56	0.31
	MSL	15	9	63 %	2.13	0.16
	Unc	7	15	32 %		
101 gene centroid	BL1	16	7	70 %	3.59	0.02
	BL2	4	14	22 %	0.23	0.02
	LAR	3	6	33 %	0.5	0.48
	M	7	7	50 %	1.1	0.99
	MSL	6	5	55 %	1.35	0.75
	Unc	17	25	40 %	NA	NA

Cases with significant association with more than one subclass were excluded. P values determined by Fisher Exact test

critical unmet need to develop focused diagnostics to identify patients that would benefit from standard chemotherapy and better align new therapeutic regimens with actionable targets expressed in TNBC patients. The TNBC subtype algorithm represents a major advance toward addressing the heterogeneity and therapeutic sensitivities of TNBC [11]. However, certain features of this original algorithm, such as the large number of genes that comprise it (2188 in total), are not optimal for its routine clinical application. The refinements described herein represent a portion of the optimization steps being performed to ultimately offer TNBC subtyping as a test with clinical utility.

Bioinformatics refinement of the original, academic research-based TNBCtype algorithm allowed minimization of the expression signature representative of all of the TNBC subtypes from 2188 to only 101 genes. Importantly, there was excellent agreement between the originally proposed 2188-gene subclassification model and the new “lean” 101-gene classifier in both a set of discovery and validation TNBC cohorts as well as in an independent TNBC

Table 5 Clinical variables with association to outcome

		AC response		pCR/RCB	
		T score	P value	score	P value
Univariate analysis	age	-2.71	0.007	-2.1	0.036
	tumor size	-0.29	0.768	-1.08	0.28
	menopausal status	-3.41	0.001	-1.52	0.127
Multivariate analysis	age	-0.32	0.749	-2.29	0.022
	tumor size	-0.97	0.331	-2.15	0.032
	menopausal status	-2.29	0.022	0.7	0.485

Association of clinical variables with outcome as measured by logistic regression in the GSE41998 TNBC cohort

clinical trial cohort treated neoadjuvantly with AC followed by the mitotic inhibitors [26]. The gene set enrichment analysis that allowed the pruning of the original model of 2188 genes into only 101 genes showed comparable classification and predictive utility. The data suggest that in the 101-gene model, the genes that define each subclass have similar biological function (Fig. 2). Further, from a practical standpoint, the reduction of the classifier to 101 genes with definition of the individual TNBC subtypes by only 8 to 15 genes will allow placement on assay platforms that would be technically challenging or impossible for the 2188-gene signature.

Preliminary evidence suggestive of the clinical utility of TNBC subtyping has already been demonstrated for both the original 2188-gene and the optimized 101-gene models. In the clinical trial cohort [26] analyzed herein using both models, the BL2 subtype was demonstrated to significantly associate with lack of tumor response to standard chemotherapy, whereas the BL1 subtype significantly associated pCR. Age was a significant predictor of pathological responses in this cohort, but the BL1 and BL2 subtypes (as defined by the 101-gene model) were independent of this factor. To put these findings in context and emphasize their potential relevance to clinical management, it is important to note that historical data show only approximately 25 % of TNBC patients will respond with pCRs to the conventional anthracycline/cyclophosphamide/mitotic inhibitor combination chemotherapy used as neoadjuvant treatment in the test cohort [28, 29]. By subclassifying a TNBC population with the 101-gene model, we found that 70 % of patients with tumors classified as BL1 experienced pCR, in contrast to only 22 % of those with BL2 tumors. Our findings corroborate the independent study published by Masuda et al., who employed the 2188-gene model on a cohort of patients from the MD Anderson Cancer Center treated with neoadjuvant chemotherapy containing sequential taxane and anthracycline-based regimens and likewise found BL1 TNBC patients to have a high rate of pCR (52 %) and BL2 patients to have the lowest (0 %) pCR rate of all subtypes [12]. Collectively, these data are supportive not only of the ability of the gene expression models to classify TNBC into stable homogenous subtypes, but also of the likely predictive utility of these subtypes to assess therapeutic sensitivities.

In the original identification of the TNBC subtypes by Lehmann and co-workers, it was noted that the BL1 subtype was typified by high expression of cell cycle and DNA damage response genes [11]. Additionally, TNBC cell lines that shared expression patterns with this subtype preferentially responded to cisplatin and it was hypothesized that patients with BL1 would have higher response rates to platinum compounds and PARP inhibitors [27] compared to the other subtypes [11]. The 101-gene model is being further refined to an even more limited gene sets

to individually classify each subtype. Thereafter, clinical utility studies will follow to assess the ability of subtyping to guide therapeutic decisions regarding the use of platinum agents, PARP inhibitors, as well as other agents believed to have efficacy in subsets of TNBC patients (e.g., checkpoint blockade inhibitors, androgen receptor antagonists and anti-angiogenics such as bevacizumab, etc.). Previous attempts with targeted therapies in unselected TNBC have largely been unsuccessful as has been the case with VEGFR and EGFR inhibitors [30, 31]. However, alignment of targeted therapies with select subsets of TNBC that display biologies dependent on a given target may accelerate development of new therapeutics that are more efficacious for patients with TNBC.

Conclusion

Our results demonstrate that a model using small gene sets can recapitulate the TNBC subtypes identified by the original 2188-gene model and in the case of standard chemotherapy, the ability to predict therapeutic response. *Additional studies are planned comparing both models on randomized clinical trial samples to fully explore the utility of models to identify responsive patient populations.*

Abbreviations

TNBC: Triple-negative breast cancer; OR: Odds ratio; pCR: Pathologic complete response; ER: Estrogen receptor; KRAS: Kirsten rat sarcoma viral oncogene homolog; EGFR: Epidermal growth factor receptor; ALK: Anaplastic lymphoma tyrosine kinase receptor; BL1: Basal like 1; BL2: Basal like 2; IM: Immunomodulatory; M: Mesenchymal; MSL: Mesenchymal stem like; LAR: Luminal androgen receptor; RMA: Robust multi-array average; PR: Progesterone receptor; ERBB2: erb-b2 receptor tyrosine kinase; GSEA: Gene set enrichment analysis; TILs: Tumor-infiltrating lymphocytes; PCA: Principal component analysis; AC: Anthracycline/cyclophosphamide; RCB: Residual cancer burden.

Competing interests

Robert S. Seitz, Kasey Lawrence, Daniel B. Bailey, Stephan W. Morris, David R. Hout, Brock L. Schweitzer are employees of and hold stock in Insight Genetics, Inc. Brian Z. Ring is compensated as a consultant for Insight Genetics, Inc. Jennifer A. Pietsenpol and Brian D. Lehmann are inventors of intellectual property for TNBCtype licensed by Insight Genetics Inc.

Authors' contributions

Conception and design: BZR, BDL, JAP, SWM and RSS. Development of methodology: BZR and RSS. Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): BZR, BDL, SWM, JAP and RSS. Writing, review, and/or revision of the manuscript: DRH, RSS, DBB, BLS, KL, BDL, and JAP. Study supervision: DRH and RSS. All authors have read and approved the manuscript.

Acknowledgments

This research was supported by National Cancer Institute Grants CA098131 and CA68485; and Komen for the Cure Foundation (SAC110030 to JAP and CCR13262005 to BDL).

Author details

¹Institute of Personalized and Genomic Medicine, College of Life Science, Huazhong University of Science and Technology, Wuhan, China. ²Insight Genetics Incorporated, Nashville, Tennessee, USA. ³Department of Biochemistry, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA.

Received: 31 July 2015 Accepted: 17 February 2016
Published online: 23 February 2016

References

- Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, Lickley LA, Rawlinson E, Sun P, Narod SA. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin Cancer Res*. 2007;13(15 Pt 1):4429–34.
- Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, Symmans WF, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol*. 2008;26(8):1275–81.
- Wang-Lopez Q, Chalabia N, Abriala C, Radosevic-Robina N, Durandoa X, Mouret-Reyniera MA, et al. Can pathologic complete response (pCR) be used as a surrogate marker of survival after neoadjuvant therapy for breast cancer? *Crit Rev Oncol Hematol*. 2015;95:88–104.
- Guidance for industry: pathological complete response in neoadjuvant treatment of high-risk early-stage breast cancer: use as an endpoint to support accelerated approval. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm305501.pdf>.
- Gyorffy B, Hatzis C, Sanft T, Hofstadter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Res*. 2015;17(1):514.
- Van Cutsem E, Lenz HJ, Kohne CH, Heinemann V, Tejpar S, Melezinec I, Beier F, Stroh C, Rougier P, van Krieken JH et al. Fluorouracil, leucovorin, and irinotecan plus cetuximab treatment and RAS mutations in colorectal cancer. *J Clin Oncol*. 2015;33(7):692–700.
- Rosell R, Carcereny E, Gervais R, Vergnenegre A, Massuti B, Felip E, Palmero R, Garcia-Gomez R, Pallares C, Sanchez JM et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol*. 2012;13(3):239–46.
- Solomon BJ, Mok T, Kim DW, Wu YL, Nakagawa K, Mekhail T, Felip E, Cappuzzo F, Paolini J, Usari T et al. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *N Engl J Med*. 2014;371(23):2167–77.
- Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747–52.
- Bertucci F, Finetti P, Cervera N, Esterni B, Hermitte F, Viens P, Birnbaum D. How basal are triple-negative breast cancers? *Int J Cancer*. 2008;123(1):236–40.
- Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest*. 2011;121(7):2750–67.
- Masuda H, Baggerly KA, Wang Y, Zhang Y, Gonzalez-Angulo AM, Meric-Bernstam F, Valero V, Lehmann BD, Pietenpol JA, Hortobagyi GN et al. Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes. *Clin Cancer Res*. 2013;19(19):5533–40.
- Weigelt B, Peterse JL, van 't Veer LJ. Breast cancer metastasis: markers and models. *Nat Rev Cancer*. 2005;5(8):591–602.
- Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A*. 2006;103(15):5923–8.
- Pochet N, De Smet F, Suykens JA, De Moor BL. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*. 2004;20(17):3185–95.
- Lee G, Rodriguez C, Madabhushi A. Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE/ACM Trans Comput Biol Bioinform*. 2008;5(3):368–84.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50.
- Gruber S, Van der Laan M. tmle: an R package for targeted maximum likelihood estimation. *J Stat Softw*. 2012;51(12):35.
- Liaw A, Wiener M. Classification and regression by random forest. *R News*. 2002;2(3):18–22.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
- Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, et al. ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25(8):1091–3. doi:10.1093/bioinformatics/btp101.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.
- Denkert C, von Minckwitz G, Brase JC, Sinn BV, Gade S, Kronenwett R, Pfitzner BM, Salat C, Loi S, Schmitt WD, et al. Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2–positive and triple-negative primary breast cancers. *J Clin Oncol*. 2014;33:983–91.
- Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*. 2002;99(10):6567–72.
- Goldstein AS, Huang J, Guo C, Garraway IP, Witte ON. Identification of a cell of origin for human prostate cancer. *Science*. 2010;329(5991):568–71.
- Horak CE, Pusztai L, Xing G, Trifan OC, Saura C, Tseng LM, Chan S, Welcher R, Liu D. Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or paclitaxel in early-stage breast cancer. *Clin Cancer Res*. 2013;19(6):1587–95.
- Morales JC, Li L, Fattah FJ, Dong Y, Bey EA, Patel M, Gao J, Boothman DA. Review of Poly (ADP-ribose) Polymerase (PARP) mechanisms of action and rationale for targeting in cancer and other diseases. *Crit Rev Eukaryot Gene Expr*. 2014;24(1):15–28.
- Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, Ollila DW, Sartor CI, Graham ML, Perou CM. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clin Cancer Res*. 2007;13(8):2329–34.
- Nahleh Z. Neoadjuvant chemotherapy for “triple negative” breast cancer: a review of current practice and future outlook. *Med Oncol*. 2010;27(2):531–9. doi:10.1007/s12032-009-9244-6. Epub 2009 Jun 10. Review.
- Cameron D, Brown J, Dent R, Jackisch C, Mackey J, Pivot X, Steger GG, Suter TM, et al. Adjuvant bevacizumab-containing therapy in triple-negative breast cancer (BEATRICE): primary results of a randomised, phase 3 trial. *Lancet Oncol*. 2013;14(10):933–42.
- Carey LA, Rugo HS, Marcom PK, Mayer EL, Esteva FJ, Ma CX, Liu MC, Storniolo AM, et al. TBCRC 001: randomized phase II study of cetuximab in combination with carboplatin in stage IV triple-negative breast cancer. *J Clin Oncol*. 2012;30(21):2615–23.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

