Luminal progenitor and fetal mammary stem cell expression features predict breast tumor response to neoadjuvant chemotherapy

Adam D. Pfefferle[1,2], Benjamin T. Spike[3], Geoff M. Wahl[3], and Charles M. Perou[1,2,4,#]

[1]Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC 27599, USA
[2]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA
[3]Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92130, USA
[4]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA

Adam D. Pfefferle: adamp@email.unc.edu; Benjamin T. Spike: bspike@salk.edu; Geoff M. Wahl: wahl@salk.edu; Charles M. Perou: cperou@med.unc.edu

[#]Corresponding Author
Charles M. Perou
Lineberger Comprehensive Cancer Center
University of North Carolina at Chapel Hill
450 West Drive, CB#7264
Chapel Hill, NC 27599
Phone: 919-843-5740
Fax: 919-843-5718
E-mail: cperou@med.unc.edu

Methods

Mammary cell subpopulation gene signatures

Gene expression measurements from fluorescence-activated cell sorting (FACS) enriched mammary cell subpopulations were obtained from three human and two murine published studies: GSE16997 [1], GSE19446 [2], GSE27027 [3], GSE35399 [4], and GSE50470 [5]. The human and murine datasets were separately combined using distance weighted discrimination (DWD) normalization to adjust for systemic microarray data biases between studies [6]. FACS subpopulation gene signatures were then derived within the human and murine dataset separately using a common approach. First, genes highly expressed within each FACS subpopulation were identified using a two-class (subpopulation X versus all others) Significance Analysis of Microarrays (SAM) analysis [7,8], with genes highly expressed and with a false discovery rate (FDR) of <5% being considered significant. Next, the intersection of each study's subpopulation gene signature was identified (e.g. aMaSC-Lim09 ∩ aMaSC-Shehata ∩ aMaSC-Prat). The intersecting gene set for each cell type was then further limited to genes uniquely found in the subpopulation of interest by removing genes found in any other subpopulation's gene set (e.g. removing members of aStr-Lim09 ∪ aStr-Shehata ∪ aStr-Prat ∪ LumProg-Lim09 ∪ LumProg-Shehata ∪ LumProg-Prat ∪ MatureLum-Lim09 ∪ MatureLum-Shehata ∪ MatureLum-Prat from the aMaSC intersecting gene set) and by removing genes associated with the myoepithelial subpopulation using a published myoepithelial gene signature produced using the same approach as those derived here [9]. Through this process, a consensus gene signature was produced for each mammary cell FACS subpopulation, for each species, which we designated as 'enriched' (e.g. aMaSC-HsEnriched).

Each FACS 'enriched' signature was further refined by supervised clustering using the human UNC308 breast tumor dataset to identify subpopulation 'features' [8]. The purpose of this process was to identify clusters of genes highly correlated across a diverse human tumor dataset, as these gene features are more likely regulated by similar factors and therefore, may by more clinically useful than the entire enriched signature. These refined features (e.g. fMaSC-feature1 for example) were defined as having at least ten genes with a Pearson correlation greater than 0.5 across all tumors in the UNC308 dataset [10]. Expression scores for both the 'enriched' and 'feature' gene signatures were determined by calculating the mean expression of the signature within each tumor; all gene signature lists are provided in Supplemental Table 1. Signatures were separately standardized to have an average expression value of zero and a standard deviation of one (N(0,1)) to allow for across signature comparisons.

Comparison of human and murine normal mammary populations

To identify possible commonalities between human and mouse normal mammary FACS populations, we used the gene set analysis (GSA) R package v1.03 [11] and R v2.12.2. Murine populations were analyzed for significant overlap with each HsEnriched gene signature. Significant overlap was defined as having $p \leq 0.05$ and $FDR \leq 0.1$ to control for multiple comparisons [11].

Mammary cell subpopulation centroids

Human mammary cell subpopulation centroids were created using the union of the 'enriched' epithelial gene signatures (aMaSC-HsEnriched ∪ LumProg-HsEnriched ∪ MatureLum-HsEnriched). The DWD single sample predictor (SSP) function [6] was used to calculate the shortest Euclidean distance between each tumor and each epithelial cell enriched centroid using three human datasets comprising over 3000 patients: UNC308 [8], Combined855 [12], and Metabric2136 [13]. To gauge the strength of each mammary subpopulation association, the silhouette width was calculated for each sample using R v3.0.1 and the 'cluster' package. Samples with a positive silhouette width were considered to have strong association. Similarly, this process was repeated using the murine cell subpopulation dataset to calculate Euclidean distances for a murine expression dataset comprising 27 models of mammary carcinoma and normal mammary tissue [14].

Chemotherapy response

Logistic regression analysis was used to determine if gene signatures derived from normal cell populations were capable of predicting pathological complete response (pCR) in breast cancer patients treated with neoadjuvant anthracycline and taxane chemotherapy regimens. For this purpose, a combined breast cancer gene expression dataset was created from three public datasets (GSE25066 [15], GSE32646 [16], and GSE41998 [17]). Only neoadjuvant anthracycline and taxane treated patients with complete clinical data (Age, ER status, PR status, HER2 status, tumor stage and pCR) were considered in the analysis, resulting in a dataset of 702 patients. The three datasets were combined using DWD normalization to adjust for systemic microarray data biases between studies [6], with the clinical characteristics found in Supplemental Table 2. The significance of each mammary subpopulation gene signature and several published predictors of pCR was determined using a series of stepwise tests. First, the ability for each signature to predict pCR was determined with a univariate analysis (UVA) using R v3.0.1 (Supplemental Table 3). Those signatures that were significant ($p<0.05$) were then considered in a multivariate analysis (MVA) with several clinical variables (Age, ER status, PR status, HER2 status, tumor stage, PAM50 subtype [18], and PAM50 proliferation score [18]) to determine if each mammary subpopulation gene signature added new information for predicting pCR (Supplemental Table 4).

References

1. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, Feleppa F, Huschtscha LI, Thorne HJ, Fox SB, Yan M, French JD, Brown MA, Smyth GK, Visvader JE, Lindeman GJ (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med 15: 907-913.
2. Lim E, Wu D, Pal B, Bouras T, Asselin-Labat ML, Vaillant F, Yagita H, Lindeman GJ, Smyth GK, Visvader JE (2010) Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. Breast Cancer Res 12: R21.
3. Spike BT, Engle DD, Lin JC, Cheung SK, La J, Wahl GM (2012) A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. Cell Stem Cell 10: 183-197.
4. Shehata M, Teschendorff A, Sharp G, Novcic N, Russell A, Avril S, Prater M, Eirew P, Caldas C, Watson CJ, Stingl J (2012) Phenotypic and functional characterization of the luminal cell hierarchy of the mammary gland. Breast Cancer Res 14: R134.
5. Prat A, Karginova O, Parker JS, Fan C, He X, Bixby L, Harrell JC, Roman E, Adamo B, Troester M, Perou CM (2013) Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes. Breast Cancer Res Treat 142: 237-255.
6. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS (2004) Adjustment of systematic microarray data biases. Bioinformatics 20: 105-114.
7. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98: 5116-5121.
8. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM (2010) Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res 12: R68.
9. Keller PJ, Arendt LM, Skibinski A, Logvinenko T, Klebba I, Dong S, Smith AE, Prat A, Perou CM, Gilmore H, Schnitt S, Naber SP, Garlick JA, Kuperwasser C (2012) Defining the cellular precursors to human breast cancer. PNAS 109: 2772-2777.
10. Hoadley KA, Weigman VJ, Fan C, Sawyer LR, He X, Troester MA, Sartor CI, Rieger-House T, Bernard PS, Carey LA, Perou CM (2007) EGFR associated expression profiles vary with breast tumor subtype. BMC Genomics 8: 258.
11. Efron B, Tibshirani R (2007) On testing the significance of sets of genes. Annals of Applied Statistics 1: 107-129.
12. Harrell JC, Prat A, Parker JS, Fan C, He X, Carey L, Anders C, Ewend M, Perou CM (2012) Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse. Breast Cancer Res Treat 132: 523-535.
13. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Group M, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486: 346-352.
14. Pfefferle AD, Herschkowitz JI, Usary J, Harrell JC, Spike BT, Adams JR, Torres-Arzayus MI, Brown M, Egan SE, Wahl GM, Rosen JM, Perou CM (2013) Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. Genome Biol 14: R125.
15. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, Martin M, Cotrina J, Gomez H, Hubbard R, Chacon JI, Ferrer-Lozano J, Dyer R, Buxton M, Gong

Y, Wu Y, Ibrahim N, Andreopoulou E, Ueno NT, Hunt K, Yang W, Nazario A, DeMichele A, O'Shaughnessy J, Hortobagyi GN, Symmans WF (2011) A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA 305: 1873-1881.

16. Miyake T, Nakayama T, Naoi Y, Yamamoto N, Otani Y, Kim SJ, Shimazu K, Shimomura A, Maruyama N, Tamaki Y, Noguchi S (2012) GSTP1 expression predicts poor pathological complete response to neoadjuvant chemotherapy in ER-negative breast cancer. Cancer Sci 103: 913-920.
17. Horak CE, Pusztai L, Xing G, Trifan OC, Saura C, Tseng LM, Chan S, Welcher R, Liu D (2013) Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. Clin Cancer Res 19: 1587-1595.
18. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 27: 1160-1167.