A single nucleotide variant on chromosome 5 (5q14) residing within ARSB distinguishes patients with basal-like human breast cancer.

Shahan Mamoor, MS
shahanmamoor@gmail.com
East Islip, NY 11730

Patients diagnosed with basal-like (also known as basal) subtype breast cancer face a more aggressive disease course and more dismal prognosis than patients diagnosed with luminal A and luminal B breast cancer molecular subtypes (1-4). We mined published microarray data (5, 6) to understand in an unbiased fashion the most distinguishing genetic and transcriptional features of tumors from patients with basal or basal-like subtype breast cancer. We identified a single nucleotide polymorphism, rs180049, residing within ARSB at chromosome 5q14.1, as among the most significant genetic differences in the tumors of patients with basal-like breast cancer. In a separate cohort of patients with basal subtype breast cancer, we observed transcriptome-wide differential expression of an ARSB transcript. We contend that an overwhelming enrichment of basal subtype-associated SNPs at 5q14 (7-20) might allude to structural features of the genome proximal to 5q14 - like that which might permit integration of a tumor virus genome (21) - rather than reflecting transcriptional changes resulting from polymorphism of any single gene.

Keywords: basal subtype, basal-like breast cancer, ARSB, systems biology of breast cancer, targeted therapeutics in breast cancer, breast cancer, molecular subtype.

Classification of disease types and disease subtypes enables development of enhanced diagnostic capabilities, prognostic information, as well as a nuanced understanding of human disease that can support superior medical, surgical and radiological treatment of complex diseases such as cancer. Molecular subtypes in human breast cancer include luminal A, luminal B, basal or basal-like, and HER2+. Systems-level analyses, such as comparative tumor transcriptome analyses, can inform such approaches. Importantly, differences in genetic background that can influence disease frequency or progression in basal subtype (1, 2) or triple negative breast cancer (which shares significant overlap with basal subtype breast cancer (3, 4)) are difficult to discern without whole-genome sequencing data. However, in the case of whole genome sequencing, data indicating that changes at the level of the genome at single nucleotide resolution are accompanied by differences in gene expression, between molecular subtypes and at the level of the transcriptome, are still lacking, but could be ascertained based on combined analysis of microarray data that encompasses both transcript and SNP-level data. We mined published microarray data (5, 6) to identify the most significant genetic and transcriptional differences when comparing human breast cancer molecular subtypes. Integrative microarray dataset analysis has provided data illuminating genes whose sequence and expression defines basal subtype breast cancers, and we present evidence here.

Methods

We utilized datasets GSE10893 (5) and GSE87049 (6) for this global differential gene expression analysis of basal subtype human breast cancer in conjunction with GEO2R. GSE10893 was generated using Illumina Human-1 Genotyping BeadChip technology with $n=37$ basal-like breast tumors and $n=132$ primary tumors of the breast from patients diagnosed with breast cancer of different, non-basal molecular subtypes (luminal A, luminal B, HER2+, claudin, claudin-low, normal-like); analysis was performed using platform GPL8864. GSE87049 was generated using Affymetrix Human Gene 1.0 ST Array [transcript (gene) version] technology with $n=14$ basal-like breast tumors and $n=109$ breast tumors of other molecular subtypes (luminal A, luminal B, HER2-enriched, and normal-like); analysis was performed using platform GPL6244. The Benjamini and Hochberg method of p -value adjustment was used for ranking of differential expression but raw p -values were used to assess statistical significance of global differential expression. Log-transformation of data was auto-detected, and the NCBI generated category of platform annotation was used. A statistical test was performed to evaluate whether ARSB gene expression was significantly different between the tumors of patients with basal and non-basal molecular subtype breast cancer using a two-tailed t-test with Welch's correction. For Kaplan-Meier survival analysis, we used the Kaplan-Meier plotter tool (22) for correlation of ARSB mRNA expression levels with recurrence-free survival in $n=4929$ breast cancer patients, $n=953$ basal subtype breast cancer patients, $n=1809$ luminal A subtype breast cancer patients, $n=1353$ luminal B subtype breast cancer patients, $n=695$ HER2+ subtype breast cancer patients, and $n=119$ normal subtype breast cancer patients.

Results

SNP array genotyping identifies rs180049 as a defining sequence variation in human basal breast cancer.

We used published microarray data (5) to compare the whole tumor genomes of 37 patients with basal or basal-like subtype breast cancer, with that of 132 patients with disease of other molecular subtype, including luminal A, luminal B, HER2+, claudin, claudin-low and normal-like, to identify single nucleotide polymorphisms that describe the most significant variations in the tumor genomes of patients with basal-like subtype breast cancer (BLBC).

SNP genotyping arrays utilize large differences in signal, after hybridization to DNA rather than RNA, to identify polymorphic sites in the genome at single nucleotide resolution. We identified a single nucleotide polymorphism (rs180049) residing within the ARSB gene, on chromosome 5 (5q14.1), as

among the most significant genetic differences in the tumors of patients with basal subtype breast cancer as compared to that of other molecular subtypes. Polymorphism at rs180049 was the 147th most significant difference, genome-wide, when comparing the tumor genomes of basal subtype patients with that of tumors from patients with other molecular subtypes (Rank: 147 out of 108250 total SNPs, equating to 99.9% (p -value: $3.43\text{E-}06$)).

ARSB is differentially expressed in basal-like subtype human breast cancer.

Analysis of a second published microarray dataset, comparing tumor transcriptome data from a separate cohort of Mexican patients with basal subtype breast cancer (14 tumors from patients with basal-like breast cancer and 109 tumors from patients with breast cancer of other molecular subtype [luminal A, luminal B, HER2-enriched, and normal-like]) (6) revealed transcriptome-wide differential expression of a ARSB transcript in patients with basal subtype human breast cancer (Table 2). When sorting each of the transcripts whose expression was measured globally, ARSB expression ranked 5976 out of 33297 total transcripts, equating to 82.1% differential expression; this was statistically significant (p -value: $6.80\text{E-}03$).

ARSB expression correlates with patient survival in luminal A and luminal B subtype breast cancer patients.

We asked whether primary tumor expression of ARSB was correlated with patient survival outcomes in patients with breast cancer when analyzing outcomes based on PAM50 molecular subtype. We observed a correlation between primary tumor expression of ARSB and recurrence-free survival (RFS) in patients with luminal A subtype human breast cancer (Figure 2; p -value: 0.082; hazard ratio: 0.83 (0.68-1.02)). ARSB primary tumor expression was a positive prognostic indicator in patients with luminal A subtype cancer. Luminal A patients with low tumor expression of ARSB possessed median RFS of 109.51 months, while luminal A patients with high tumor expression of ARSB possessed median RFS of 129.6 months.

We also observed a correlation between primary tumor expression of ARSB and recurrence-free survival (RFS) in patients with luminal B subtype human breast cancer (Figure 2; p -value: 0.025; hazard ratio: 0.82 (0.69-0.98)). ARSB primary tumor expression was a positive prognostic indicator in patients with luminal B subtype cancer. Luminal B patients with low tumor expression of ARSB possessed median RFS of 125.65 months, while luminal B patients with high tumor expression of ARSB possessed median RFS of 164.3 months.

We also observed a correlation between primary tumor expression of ARSB and recurrence-free survival (RFS) in all patients (irrespective of molecular subtype) (Figure 2; p -value: 0.001; hazard ratio: 0.84 (0.76-0.93)). ARSB primary tumor expression was a positive prognostic indicator in patients with breast cancer. Patients with low tumor expression of ARSB possessed median RFS of 43.93 months, while patients with high tumor expression of ARSB possessed median RFS of 56.58 months.

ARSB primary tumor expression was not correlated with recurrence-free survival in patients with HER2+ subtype breast cancer (Figure 2; p -value: 0.89; hazard ratio: 0.98 (0.77-1.25)) and in patients with normal subtype breast cancer (Figure 2; p -value: 0.3; hazard ratio: 0.7 (0.35-1.38)).

Discussion

We found through analysis of published microarray datasets (5, 6) that a genomic position marked by SNP rs180049 was among the most significant sequence variations in the tumor genomes of patients

1 with basal subtype human breast cancer when compared with that of patients with breast cancer of other
2 molecular subtypes. rs180049 resides within the ARSB gene. Similar comparative tumor transcriptome
3 study based on molecular subtype in a separate patient dataset revealed differential expression of an ARSB
4 transcript when comparing tumors of other molecular subtypes against that of tumors from patients of the
5 basal subtype. Kaplan-Meier analysis of patient survival data revealed that ARSB expression was
6 positively correlated with recurrence-free survival in patients with luminal A and luminal B breast cancer.
7 Changes in ARSB expression may be important events to understand as part of the myriad processes that
8 likely govern transformation and tumor maintenance in basal or basal-like human breast cancer.

9 As we map the polymorphic landscape of the basal-like breast cancer genome in humans, we have
10 identified and described over a dozen alleles enriched in the tumor genomes of patients with basal-like
11 breast cancer, each of which reside on chromosome 5, centered on positions located at 5q14 (7-20).
12 Differences in DNA sequence associating with the basal subtype, residing at one genomic location but not
13 specific to any single gene, could reflect a location at which DNA tumor virus genome integration occurs
14 in humans. An integration site for high-risk human papillomavirus type 16 (HPV16), found in a head and
15 neck squamous cell carcinoma, has been described to reside at 5q14 (22). An integration site for Merkel
16 cell polyomavirus has also been described at 5q14.2 (23). Thus, it is possible that these SNPs reflect sites
17 in the human genome that cooperate or function together structurally to permit or favor a viral integration
18 event at that location, genome rearrangement that might follow (are secondary to) a viral integration event,
19 or a third possibility, genome rearrangements that are specific to the basal subtype but do not involve a
20 tumor virus. Another explanation to explain the fact that multiple SNPs that associate with the basal-like
21 subtype in human breast cancer reside at 5q14 might involve additive effects of alternative alleles that
22 result in transcriptional changes that together contribute to transformation or tumor progression.
23
24
25
26
27
28

References

1. Rakha, E.A., Reis-Filho, J.S. and Ellis, I.O., 2008. Basal-like breast cancer: a critical review. *Journal of clinical oncology*, 26(15), pp.2568-2581.
2. Milioli, H.H., Tishchenko, I., Riveros, C., Berretta, R. and Moscato, P., 2017. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC medical genomics*, 10(1), pp.1-17.
3. Foulkes, W.D., Smith, I.E. and Reis-Filho, J.S., 2010. Triple-negative breast cancer. *New England journal of medicine*, 363(20), pp.1938-1948.
4. Dietze, E.C., Sistrunk, C., Miranda-Carboni, G., O'regan, R. and Seewaldt, V.L., 2015. Triple-negative breast cancer in African-American women: disparities versus biology. *Nature Reviews Cancer*, 15(4), pp.248-254.
5. Weigman, V.J., Chao, H.H., Shabalin, A.A., He, X., Parker, J.S., Nordgard, S.H., Grushko, T., Huo, D., Nwachukwu, C., Nobel, A. and Kristensen, V.N., 2012. Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast cancer research and treatment*, 133(3), pp.865-880.
6. Romero-Cordoba, S.L., Salido-Guadarrama, I., Rebollar-Vega, R., Bautista-Piña, V., Dominguez-Reyes, C., Tenorio-Torres, A., Villegas-Carlos, F., Fernández López, J.C., Uribe-Figueroa, L., Alfaro-Ruiz, L. and Hidalgo-Miranda, A., 2021. Comprehensive omic characterization of breast cancer in Mexican-Hispanic women. *Nature communications*, 12(1), pp.1-19.
7. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within VCAN distinguishes patients with basal-like human breast cancer.
8. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within NDUFAF2 distinguishes patients with basal-like human breast cancer.
9. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within SCAMP1 distinguishes patients with basal-like human breast cancer.
10. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within PLK2 distinguishes patients with basal-like human breast cancer.

11. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within ADAMTS6 distinguishes patients with basal-like human breast cancer.
12. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within MAST4 distinguishes patients with basal-like human breast cancer.
13. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within EDIL3 distinguishes patients with basal-like human breast cancer.
14. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within IQGAP2 distinguishes patients with basal-like human breast cancer.
15. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within ERBIN distinguishes patients with basal-like human breast cancer.
16. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within MSH3 distinguishes patients with basal-like human breast cancer.
17. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within SSBP2 distinguishes patients with basal-like human breast cancer.
18. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within RNF180 distinguishes patients with basal-like human breast cancer.
19. Mamoor, S., 2022. A single nucleotide variant on chromosome 5 residing within BTF3 distinguishes patients with basal-like human breast cancer.
20. Mamoor, S., 2022. Single nucleotide variants on chromosome 5 residing within MEF2C and MEF2C-AS1 distinguish patients with basal-like human breast cancer.
21. Györfy, B., Lanczky, A., Eklund, A.C., Denkert, C., Budczies, J., Li, Q. and Szallasi, Z., 2010. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast cancer research and treatment*, 123(3), pp.725-731.
22. Wentzensen, N., Vinokurova, S. and Doeberitz, M.V.K., 2004. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer research*, 64(11), pp.3878-3884.
23. Martel-Jantin, C., Filippone, C., Cassar, O., Peter, M., Tomasic, G., Vielh, P., Brière, J., Petrella, T., Aubriot-Lorton, M.H., Mortier, L. and Jouvion, G., 2012. Genetic variability and integration of Merkel cell polyomavirus in Merkel cell carcinoma. *Virology*, 426(2), pp.134-142.

Rank: 147
SNP: rs180049
p-value: 3.43E-06
t: -4.79716824
B: 4.184992

Table 1: A genomic position marked by rs180049 (residing within the ARSB gene) is among the most significant genetic differences in the tumor genomes of basal subtype human breast cancer when compared with that of other molecular subtypes.

The rank of global differential expression, SNP ID, the *p*-value with respect to differential expression transcriptome-wide, *t*, a moderated *t*-statistic, and *B*, the log-odds of differential expression between the groups compared are listed in this chart.

Figure 1: rs180049 is significantly depleted in the primary tumors of patients with basal subtype human breast cancer.

A quantitative measure of ARSB polymorphism at rs180049 in the tumors of patients with basal subtype breast cancer (left) and in the tumors of patients with breast cancer of non-basal molecular subtype (right) is graphically depicted; the result of a statistical test evaluating significance of difference in this signal between the tumors of basal subtype patients and the tumors of patients with cancer of other molecular subtype is $p < 0.0001$. Depletion here indicates that in the basal subtype, probe signal was significantly decreased as compared to other molecular subtypes considered together, demonstrating a difference in allelic frequency in basal subtype patients at this genomic site.

Rank: 5976
probe: 8112807
p-value: 6.80E-03
t: 2.7522014
B: -3.06759
Gene: ARSB
Gene name: arylsulfatase B

Table 2: ARSB is among the genes most differentially expressed, transcriptome-wide, in a separate cohort of patients with basal subtype breast cancer.

The rank of global differential expression, probe/transcript ID, the *p*-value with respect to differential expression transcriptome-wide, *t*, a moderated *t*-statistic, *B*, the log-odds of differential expression between the groups compared, the gene and gene name are listed in this chart.

Figure 2: ARSB primary tumor expression correlates with recurrence-free survival in luminal A and luminal B subtype breast cancer patients.

Depicted in this Kaplan-Meier plot is the probability of recurrence-free survival for $n=4929$ breast cancer patients (above, left), $n=953$ basal subtype breast cancer patients (above, middle), $n=695$ HER2+ subtype breast cancer patients (above, right), $n=1809$ luminal A subtype breast cancer patients (below, left), $n=1353$ luminal B subtype breast cancer patients (below, middle), and $n=119$ normal subtype breast cancer patients (below, right) stratified into two groups, based on low or high expression of ARSB in patient primary tumors. The log rank p -value denoting statistical significance of difference in recurrence-free survival when comparing the two groups, as well as hazard ratio for this comparison is listed above. Listed below is the number of patients at risk (number of patients alive) per interval, after stratification based on ARSB expression; in the first interval, number at risk is number of patients alive; in each subsequent interval, number at risk is the number at risk less those who have expired or are censored.

Luminal A

Low ARSB expression: 109.51 months

High ARSB expression: 129.6 months

Luminal B

Low ARSB expression: 125.65 months

High ARSB expression: 164.3 months

All patients

Low ARSB expression: 43.93 months

High ARSB expression: 56.58 months

Median RFS is superior in luminal A and luminal B subtype breast cancer patients with high primary tumor expression of ARSB.

The median RFS (recurrence-free survival) of $n=1809$ luminal A subtype breast cancer patients, $n=1353$ luminal B subtype breast cancer patients, and $n=4929$ patients with breast cancer, with low primary tumor expression of ARSB and high primary tumor expression of ARSB is listed in this chart.