Published OnlineFirst January 12, 2018; DOI: 10.1158/1541-7786.MCR-17-0227

Survival analysis

The vital status and overall survival time were obtained from clinical data in BCR XML format. We then extracted days to follow up, vital status, days to birth, pathologic stage, and days to death from the XMLs. To calculate the overall survival, days to death was used for expired patients, and days to last follow up was used for surviving patients. The XML data contained many versions of follow up information, and the latest version featuring non-missing value was used. The AA and EA patients were filtered and separately divided into three groups: high, intermediate, and low, according to three quantiles of IR expression level for each identified gene. The low group consisted of patients with IR expression in the first one-third of the quantile, the intermediate group consisted of patients with IR expression in the first and second third of the quantile, and the high group consisted of patients with IR expression that exceeded the second third of the quantile. We performed Kaplan–Meier overall survival analysis between different combinations of AA and EA groups: AA high group versus EA high group, AA low group versus EA low group, AA high group versus EA low group, and AA low versus EA high group. Moreover, we applied Cox regression on the aforementioned groups, adjusting for age and cancer stage.

Results

Our working hypothesis is illustrated in Fig. 1, and an overview of our study is depicted in Supplementary Fig. S1.

Differential expression of retained introns in population groups

We used TCGA classification to define the genetic ancestry of the breast cancer patients. Although the admixture of AA and EA ancestry likely shades the results of our analysis, we did verify the population stratification of TCGA cases based on the genomic data. Using the recently published approach (26), which successfully captured genomic differences between population and subtypes in the TCGA data, we confirmed that only two TCGA EA cases were genetically different; one was AA and the other one was Asian (Supplementary Table S1).

We identified 5,280 retained intron events (i.e., 3,134 unique genes) in the set of the 124 AA and 703 EA, and we calculated the expression levels (i.e., PSI) using rMATS (see Materials and Methods). Among the 5,280 events, 44 retained introns (i.e., intron retention [IR]), comprising 41 genes, were statistically significant at $P < 0.05$, with FDR < 0.3 (Fig. 2A and B; Supplementary Table S2). These IRs and genes were further investigated. An absolute value difference of 0.1 in PSI values between AA and EA (i.e., |PSI of the retained intron in EA cases − PSI of the retained intron in AA cases|) was used as the criterion to determine whether or not a retained intron was differentially or commonly expressed in the AA and EA cases. Of the 44 statistically significant IRs, 10 IRs (comprising 10 genes) and four IRs (comprising four genes) were found to be differentially expressed according to our 10% difference criterion. The 10 IRs more frequently occurred in AA cases, and the four IRs in EA cases. Thirty IR events in 29 genes were common to both AA and EA cases. To further computationally validate the random chance of the 44 IRs being the true signal, we calculated the empirical *P* values (see Materials and Methods), and we found that all 44 IRs occur at a probability higher than that expected by chance ($P < 0.01$).

Population-dependent expression of IR in breast cancer subtypes

Breast cancer is classified into four major molecular subtypes based on the status of three genes: ER, PR, and HER2. As each of

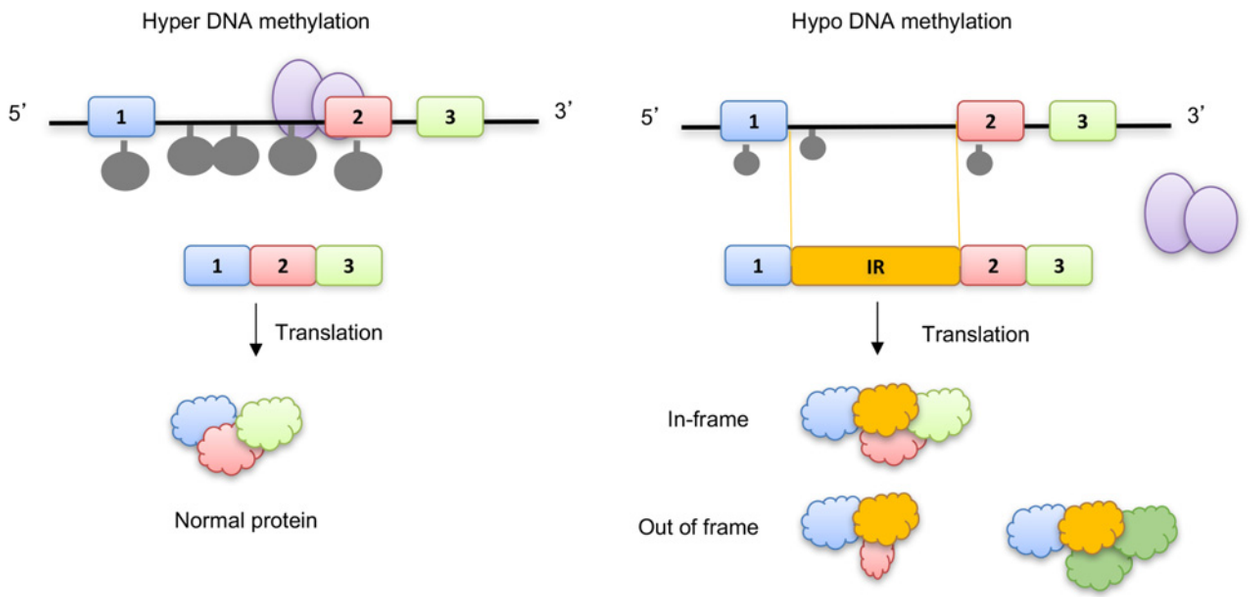

Figure 1.

Illustration of working hypothesis and biological impact of retained introns. We hypothesized that hypo-DNA methylation is correlated with intron retention events. Out-of-frame intron retention potentially causes a frame-shift of all of the downstream exons, which eventually produces mRNA that leads to a severe loss of function, such as a truncated protein or nonsense-mediated mRNA decay (NMD).

Published OnlineFirst January 12, 2018; DOI: 10.1158/1541-7786.MCR-17-0227

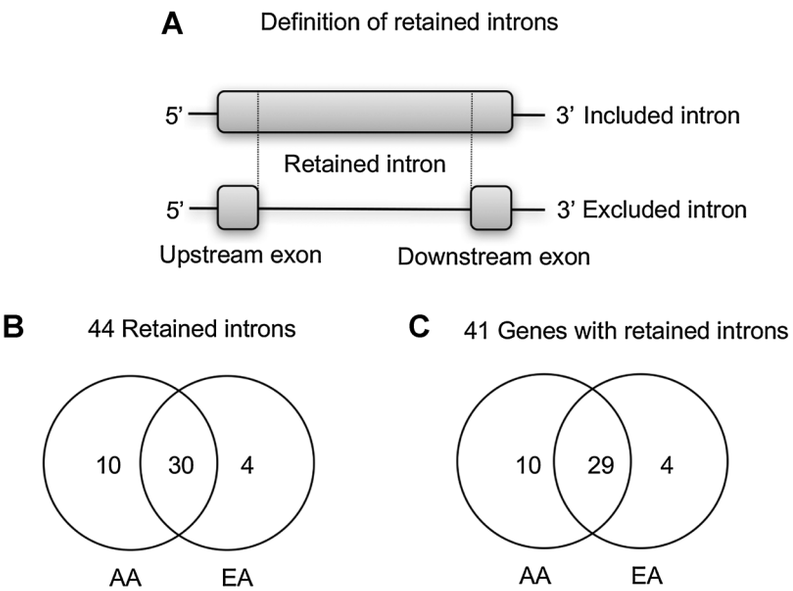

Figure 2.
Retained introns in AA and EA cases. **A,** The illustration on the left indicates the retained exons: The intron flanked by two exons (i.e., upstream and downstream exons) can be included (i.e., retained intron [IR] or intron retention) in the mature mRNA, when the splice site of these flanking exons is not correctly recognized by the spliceosome. We hypothesize that under-methylated introns have a higher probability to be retained. Numbers of included and excluded introns of expressed genes that were identified by rMATS program are shown in Supplementary Table S2. **B,** We found 44 statistically significant IRs (Supplementary Table S2). Thirty retained introns are commonly expressed in both AA and EA cases. Ten retained introns and four retained introns are more expressed in AA and in EA, respectively. **C,** The numbers represent the count of genes corresponding to the number of retained introns in the Venn diagram shown in **B**. Two genes have multiple retained intron events.

these molecular subtypes of breast cancer has a clinical implication, we examined whether or not these subtypes specify the differential expression of retained introns vis-à-vis AA and EA cases. For each classified group of four molecular subtypes, which comprises 507 cases (see Materials and Methods), we counted how many IRs are differentially expressed between AA and EA data (Supplementary Fig. S2). Among the 44 statistically significant IRs, 29 IRs (65%) of the Basal-like subtype were differentially expressed between AA and EA cases, four (9%) of HER2+, 40 (90%) of Luminal A, and 22 (50%) of Luminal B (Fig. 3A).

Although 30 IRs were common to both AA and EA cases according to the PSI value criterion (i.e., less than 10% difference defined as commonly expressed), all 30 IRs showed a population-dependent expression in specific subtype ($P < 0.05$, *t*-test), 28 in Luminal A, 19 in Basal-like, 17 in Luminal B, and two in HER2+. For example, the retained intron in the *ARL6IP4* gene did not show differences in PSI value between AA and EA cases when subtypes were not considered. However, in each subtype, this retained intron demonstrated higher expression in the AA cases compared to EA cases (Fig. 3B). This suggests that certain retained introns can be expressed not only population-dependently, but also subtype-dependently. As we expected, the highest proportion (i.e., 90%) of IRs with population-dependent expressions was Luminal A, which is the most prevalence breast cancer subtype. Moreover, we found that Basal-like showed the second highest proportion of IRs with population-dependent expressions. The prevalence of Basal-like is the triple negative breast cancer (TNBC) subtype, accounting for 15% to 20% of breast cancer subtype. Because of the sample size as a rate of each subtype are different in populations, the Basal-like and Luminal A subtypes were considered for the correlation test of IR expression with methylation as described below.

Implication of length and weak splice sites in retained introns

The length of introns and weak splice sites of exons flanking an intron may be key features that predict IR (27, 28). We first investigated the correlation of the length of IRs with their differential expression levels (PSI; Supplementary Table S2). After removing one outlier, which was 1,621 nt in length from the retained intron in the *TAMM41* gene, 43 IRs were 348 ± 193 nt (mean ± SD, $n = 43$) in length, and the length of retained introns in humans (i.e., 259 ± 290 nt, $n = 1{,}515$) was used as a control measurement. We also performed a linear regression (i.e., lm function in R) for the length and expression of the 43 IRs. The length of retained introns did not correlate with differential expression ($P = 0.93$). This suggests that the length of retained introns may not be a feature that leads to intron retention, which is consistent with previous studies (28). To further investigate whether or not the cases of retained introns that we identified were biased toward the shorter length of introns, we measured the length of all human introns. The length of the retained introns (i.e., 259 ± 290) were much shorter than the length of all human introns (i.e., 61,323 ± 19,750 nt, $n = 1{,}091{,}486$). Moreover, although short introns generally are retained (28), the longer retained introns may not be detectable in this study because of (1) the TCGA RNA-Seq, which only includes mRNA with the poly(A) tail or (2) the degradation of the introns by nonsense-mediated mRNA decay (NMD; ref. 29).

The functional interpretation of the length of IR is important because retained introns potentially may cause a frame-shift of all of the downstream exons, which eventually produces mRNA that leads to a severe loss of function, such as a truncated protein or NMD (right panel of Fig. 1). We determined both out-of-frame and in-frame for the 44 IRs using the remainder of the length divided by the divisor 3 (Supplementary Table S2). For example, the IR with a remainder of 1 or 2 indicates out-of-frame, and a remainder of 0 indicates in-frame. Moreover, 78% (i.e., 11 of 14) of differentially expressed IRs in AA and EA cases were out-of-frame, and 60% (i.e., 18 of 30) of commonly expressed IRs were out-of-frame.

When upstream and downstream exons that flank an intron have a weak splice site (i.e., a non-consensus splice site sequence), the flanked introns tend to be retained (27, 28). Therefore, we investigated whether or not the splice sites of the 44 statistically

Published OnlineFirst January 12, 2018; DOI: 10.1158/1541-7786.MCR-17-0227

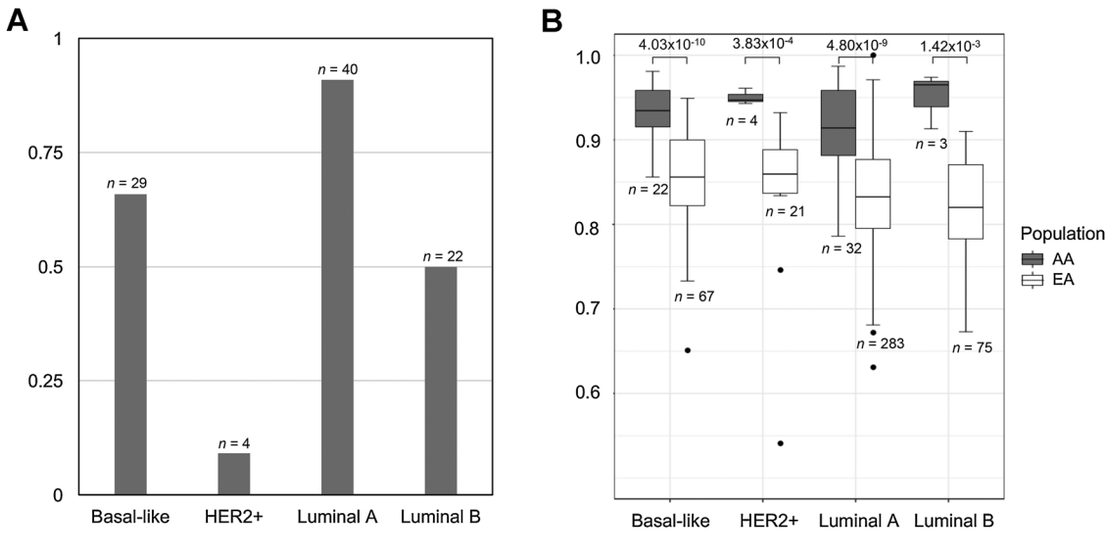

Figure 3.

Specification of breast cancer subtypes in differentially expressed IRs between AA and EA cases. **A,** Each bar represents a proportion of IR with population-dependent expression among 44 IR events for each subtype ($P < 0.05$, *t*-test, two-tailed). Basal-like subtype is represented by 89 cases (i.e., 22 AA and 67 EA), HER2+ by 25 cases (i.e., four AA and 21 EA), Luminal A by 315 cases (i.e., 32 AA and 283 EA), and Luminal B by 78 cases (i.e., three AA and 75 EA). **B,** For example, higher expression of a retained intron in the *ARL6IP4* gene was observed in AA cases compared to EA cases for each subtype (i.e., $P = 4.03 \times 10^{-10}$ for basal-like, $P = 3.83 \times 10^{-4}$ for HER2+, $P = 4.80 \times 10^{-9}$ for Luminal A, and $P = 1.42 \times 10^{-3}$ in Luminal B, *t* test, two-tailed).

significant IRs displayed consensus splice site (i.e., GT at the 5′ end of the intron and AG at the 3′ end) by examining the first and last two bases of the retained intron. These four bases were extracted from the human reference genome sequence (GRCh37.75) using the twoBitToFa tool (i.e., see https://genome.ucsc.edu/golden Path/help/twoBit.html). None of the 44 IRs were weak splice sites, which suggests that other features (i.e., methylation and splicing regulatory elements) can be implicated in the IRs rather than the status of their splice site sequences.

Methylation patterns in population groups and breast cancer subtypes

In addition, we investigated whether or not methylation levels differ between AA and EA breast cancer cases. Following the assumption that all methylation loci are independent of one another, for each methylation locus, we compared β-values for all genomic regions [i.e., promoter regions, exons, and introns ($n = 392{,}591$), see Materials and Methods] between AA and EA cases with a *t*-test for normally distributed data and a Mann–Whitney–Wilcox test for non-normally distributed data. The observed *P* value from the test (i.e., $-\log P$) was plotted against the expected value (i.e., a null distribution) as a quantile–quantile (Q–Q) plot. The broad differences observed in methylation are likely due to not only the population differences (Fig. 4A for all types), but also subtype differences (Fig. 4B for Luminal A only, and Fig. 4C for Basal-like only). Similar results for the methylation loci in only gene body [i.e., exons and introns ($n = 285{,}642$)] were observed (Fig. 4D–F).

Correlation of IRs with methylation in population groups and breast cancer subtypes

Because (1) DNA methylation is implicated in splicing regulation (8–12) and (2) we also found a specific pattern of methylation status in population groups and breast cancer subtypes (Fig. 4), we examined whether or not methylation is involved in IR expression. To do this, we first searched for the methylation loci for an association test, which are located in the upstream exon, intron, and downstream exon of a given retained intron. Among 44 IRs, 40 unique methylation sites (CpG sites) were found in 20 IRs associated with 18 genes. Twelve IRs had multiple methylation loci, and 44 distinct events were tested in this study. For these 44 events, we performed a regression analysis using regression models (see Materials and Methods) to identify a correlation of the IR expression with methylation by population groups and subtypes. The dependent variable was IR expression, and the independent variables were methylation and population for cases of all subtypes, Basal-like, and Luminal A subtypes. We identified 15 events that displayed a statistically significant correlation ($P < 0.05$) of IR expression with methylation (Supplementary Table S3). Ten events were statistically significant for cases of all subtypes, nine for Luminal A (i.e., seven overlapped in 10 of all subtypes), and two for the Basal-like subtype. Seven events showed a negative correlation, which suggests that under-methylated introns may have a higher probability to be retained; however, eight IR events did not reveal a negative correlation with methylation. Among 15 events that demonstrated statistically significant correlation, six methylation loci (i.e., cg18811158, cg11570697, cg01423722, cg26343154, cg00502028, and cg10063637) demonstrated population-dependent status in all subtype cases, two loci (i.e., cg26343154 and cg10063637) demonstrated the status in Luminal A cases, and six other loci (i.e., cg11570697, cg01423722, cg00502028, cg08042487, cg00036328, and cg13596773) demonstrated the status in Basal-like subtype cases (Supplementary Table S3).

As described above, both weak splice sites of exons flanking an intron and the existence of splicing regulatory element are both important factors contributing to IR. Because none of 15 events were weak splice sites of IRs, we further examined whether or not

Published OnlineFirst January 12, 2018; DOI: 10.1158/1541-7786.MCR-17-0227

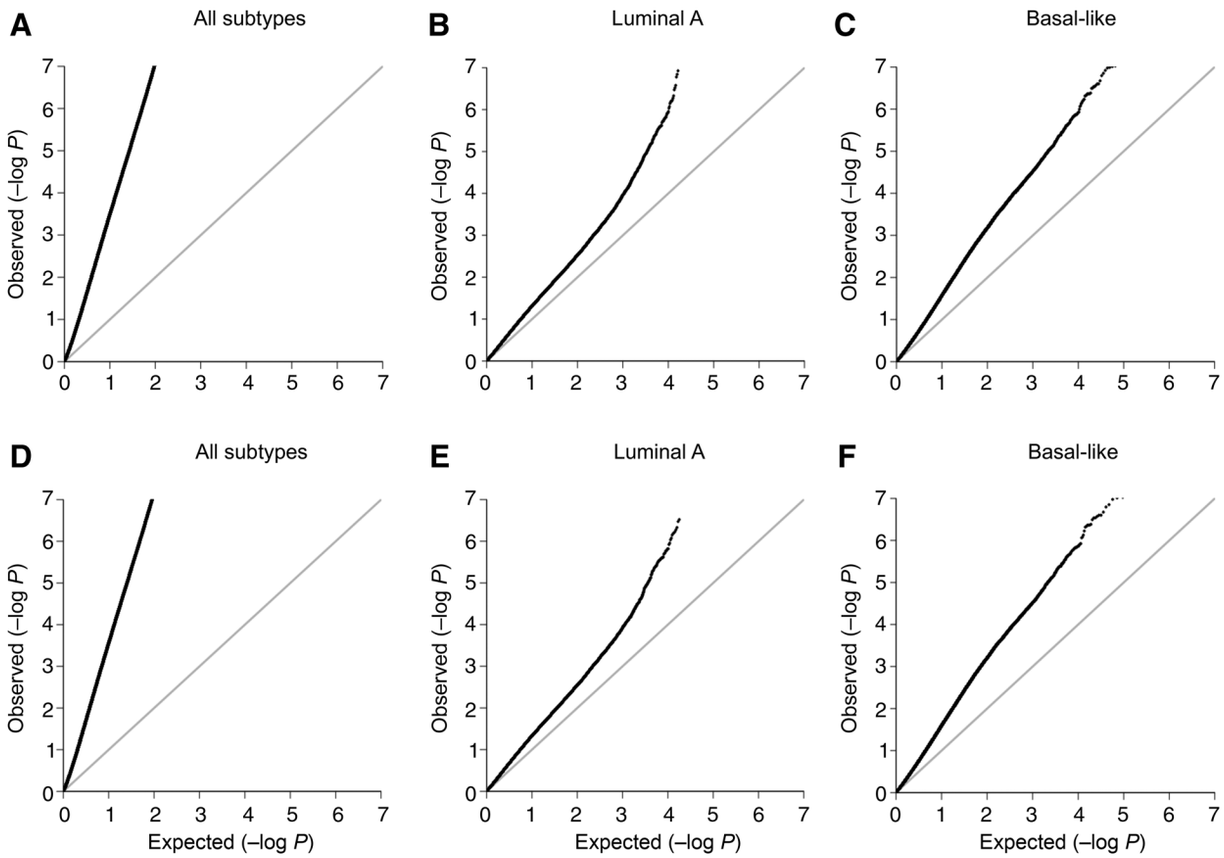

Figure 4.

Differential methylation between AA and EA cases among breast cancer subtypes. For all methylation loci in all genomic regions ($n = 392,591$; selected probes when a case count of "NA" for β value was less than 10), β-values of two population groups were compared, and the observed P values were calculated using a t-test (t. test in R) or Mann–Whitney–Wilcox test (wilcox.test in R) based on the data distribution for each methylation site. The normality of β value distributions in EAs and AAs for each methylation site was tested with the Shapiro–Wilk test (shipiro.test in R). The x-axis represents the expected P value: a null distribution. Q–Q plots are shown for the P values of β-value distribution between (**A**) 101 AA and 520 EA cases for all subtypes, (**B**) 26 AA and 214 EA cases for the Luminal A subtype, and (**C**) 18 AA and46 EA cases for the Basal-like subtype. The Q–Q plots depicted in **D**–**F** represent the same statistical results for the methylation loci in only intragenic regions (i.e., exon and intron, $n = 285,642$) for all subtypes, Luminal A, and Basal-like, respectively.

methylation loci are involved in the splicing regulatory elements to understand the splicing regulation of IR.

Two methylation loci of 15 events (i.e., cg02725362 in the downstream exon of the IR of *RHOT2* and cg19723402 in the downstream exon of the IR of *TUBGCP6*) were the CpG sites located in the exonic splicing enhancer, which is a hexameric motif, *GAACAC*(Met)g and *TCCACC*(Met)g, respectively (30). Their methylation levels did not differ between AA and EA cases; nonetheless, they displayed a statistically significant correlation with IR expression for all subtype cases.

We investigated whether or not germline variants may affect the methylation sites due to the different frequencies of the alleles in the population. We determined which methylation sites are known sites of germline variation. Among 40 methylation sites, we identified 25 that feature the SNPs (dbSNP147 version), using the UCSC genome browser. However, 21 SNPs displayed a minor allele frequency (MAF) of under 0.01, three SNPs did not have MAF, and only one SNP (i.e., rs149371042) was a common variation (i.e., MAF = 0.0218), which indicates a deletion/insertion variation (Supplementary Table S4).

Biological interpretation using a protein interaction network

To investigate whether or not the genes with retained introns are functionally similar, we performed gene ontology (GO) enrichment using DAVID (31) to annotate the gene functions of 44 IRs. We did not observe a statistically significant (i.e., $P < 0.01$) incidence of the GO category as described in Hochberg, which suggests that at least these retained introns may be functionally diverse.

Proteins that are closer to each other in a network are more likely to have related functions (32). Proteins directly connected (i.e., sharing a single edge) to a given protein (i.e., sharing a single edge) are called "first interactors." First interactors of a known disease or phenotype protein tend to be involved in the same disease or biological process (33, 34), and mutations in first interactors of disease proteins can lead to similar disease phenotypes (35). Therefore, first interactors can be used to estimate functional linkage(s) between candidate proteins and established disease proteins (36). We compiled a list of established breast cancer genes ($n = 280$) with two resources: (1) breast cancer gene signatures identified by Paik *et al.* (37); and (2) genes reported

Published OnlineFirst January 12, 2018; DOI: 10.1158/1541-7786.MCR-17-0227

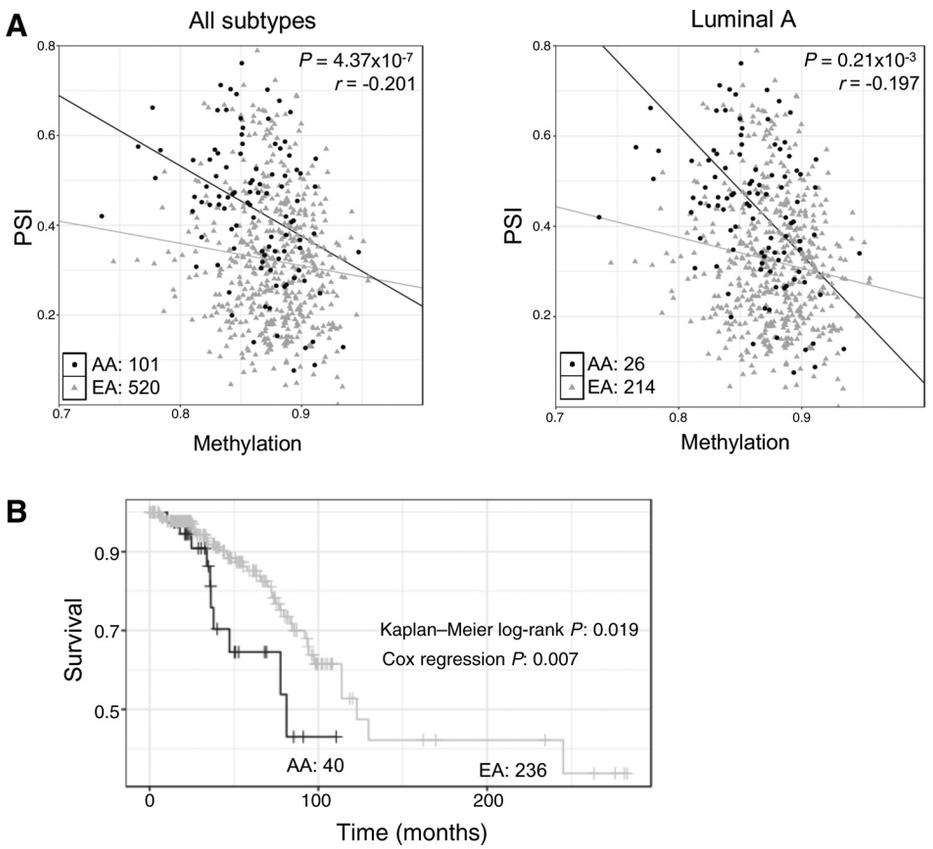

Figure 5.

Case study of the retained intron (chr1: 12089339-12089821) and methylation (cg26343154) in the MIIP gene. **A,** Multivariate linear regression results for statistical correlation of a high IR expression with a low methylation in AA and EA cases are plotted in the left panel of for all subtypes and the right panel for the Luminal A subtype. **B,** Overall survival curve by two groups for all subtypes, a group of 40 AA cases with a high IR expression (third quantile), and a group of 236 EA cases with a low IR expression (first quantile) in the MIIP gene.

under the "breast" tumor type in the Sanger Cancer Genome Census (38). To define the first interaction between our 18 genes and the established breast cancer genes, we used the STRING databases (highest confidence score ≥ 0.9), which is a collection of physical and functional associations (39). As shown in Supplementary Fig. S3A, five genes (i.e., *MIIP*, *PFKL*, *TUBGCP6*, *RHOT2*, and *RPL10A*, marked with a red asterisk) directly interact with eight known breast cancer genes. The sub-network (Supplementary Fig. S3B), comprising four known breast genes and *PFKL*, which has a known and direct association with *AKT1* and *HIF1A*, is considerably enriched in the HIF-1 signaling pathway (i.e., hsa04066; FDR = 0.0012). The HIF pathway is known to be deregulated among different cancer types, which results in an increasing hypoxic tumor cell resistance to chemotherapy (40). This resistance stems from Tubulin Gamma Complex Associated Protein 6 (TUBGCP6), which is the first interactor of two known breast cancer genes (i.e., *BRCA1* and *NEK2*); moreover, TUBGCP6 and BRCA1 are annotated in the γ-tubulin ring complex (i.e., GO:0008274; FDR = 0.004), which is the core functional unit of the centrosome. An altered γ-tubulin ring complex has been observed in breast cancer cells, and the delocalization from the centrosome can confer the resistance to microtubule-targeted chemotherapies (41).

Clinical interpretation and case study of the MIIP gene

For nine genes corresponding to 15 events, as shown in Supplementary Table S3, we performed a Kaplan–Meier overall survival analysis (see Materials and Methods). Five genes (i.e., *TM7SF2*, *RHOT2*, *MIIP*, *ARL6IP4*, and *NMRAL1*) exhibited a clinically, and statistically, significant implication. Moreover, the migration and invasion inhibitory protein (*MIIP*) gene, a recently recognized oncogene, is highly expressed in the advanced stages of cancer, such as breast cancer (42–46). Although the IR identified in *MIIP* was upregulated more in AA (mean of PSI = 0.41) than EA (mean of PSI = 0.32), we observed high-methylation (i.e., cg26343154) in EA cases (median β-value = 0.877) compared to AA cases (median β-value = 0.865; $P = 2.92 \times 10^{-5}$, Wilcox test) for all types and Luminal A type ($P = 6.0 \times 10^{-3}$; 0.881 for EA cases; 0.858 for AA cases). Furthermore, we found that that the IR expression was inversely correlated with the methylation status in the population groups (i.e., cg26343154, $r = -0.201$, $P = 4.37 \times 10^{-7}$, multivariate linear regression, left panel of Fig. 5A). As shown in Fig. 5B, a group of EA cases with a low IR expression level in *MIIP* exhibited an expected longer survival time than a group of AA cases with a high expression level of IR.

Discussion

The main limitation of this study is the inability to perform a replication analysis due to the nature of the extant accessible data. Most publicly available and independent breast cancer patient cohorts are all part of the TCGA database. Furthermore, microarray-based gene expression data are not suitable for either AS or mRNA analysis, and RNA-Seq is currently available only for EA patients; the database lacked AA patients, and studies of smaller sample sizes are not comparable with the TCGA data. However, we confirmed statistically that the 44 IRs among populations are more likely to be a true signal (empirical $P < 0.01$); moreover, we

Published OnlineFirst January 12, 2018; DOI: 10.1158/1541-7786.MCR-17-0227

performed the validation analysis with an independent cohort of breast cancer patients. Because the number of AA patients in the sample was relatively small (47), and they lacked RNA-Seq data, we used only the EA independent cohort set ($n = 17$, GSE69240) (48) and not the independent AA set ($n = 3$). For the independent EA cases ($n = 17$) and TCGA AA cases ($n = 127$), we performed the same analysis by inputting the two group populations into rMATS and identified retained introns. Among 44 IRs, six were confirmed by this analysis. Although we demonstrated evidence of differences in intron retention between the two populations, further experimental validation will be required to confirm these differences. Moreover, a larger independent cohort of AA breast cancer patients will be required for further validation.

Other limitations arise from the underlying biology. For one, the effect of a single methylation locus may not be individually predictive of intron retention. It is likely that multiple germline variants and methylation loci have additive effects in determining splicing outcomes. In the event of very low correlations between individual methylation loci and intron retention, it will be important to consider the aggregate correlation of all regulatory elements on a single splicing event. Furthermore, certain variants may have higher frequencies in breast cancer cases and thus are predisposed to the retained introns in a population- and subtype-dependent manner regardless of whether they are linked to methylation.

We found that seven hypomethylations were correlated with higher levels of intron expression in mRNA. In other words, the methylation level of an intron is inversely correlated with its retention in mRNA transcripts from the gene in which it is located, supporting that methylation regulates IR via directly affecting the rate of transcription elongation (49). Furthermore, we observed a statistically significant population-dependent difference in the methylation level of retained introns: in samples from AA donors, retained introns were less methylated and more highly expressed compared to those of samples from EA donors (i.e., *MIIP* in Supplementary Table S3 and Fig. 5). A differential biological response to change in methylation status could underlie some of the known yet unexplained disparities between certain population groups that have been observed among breast cancer patients (50).

Our findings have translational implications for improving diagnosis, prognosis, and treatment for breast cancer. Understanding not only epigenetic differences between population groups, but also their correlation with breast cancer is an important step toward personalized cancer care (51). Moreover, IR is not limited to breast cancer; transcriptomes from many different types of cancer show a higher incidence of IR compared to those from healthy controls (14, 15). Furthermore, our research and findings contribute to the understanding of how epigenetic markers in the gene body communicate with transcriptional machinery to control mRNA diversity.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Authors' Contributions

Conception and design: Y. Lee
Development of methodology: D. Kim, Y. Lee
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M. Shivakumar, O.I. Olopade, D. Kim, Y. Lee
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): D. Kim, M. Shivakumar, S. Han, Y.-J. Lee, Y. Zheng, O.I. Olopade, D. Kim, Y. Lee
Writing, review, and/or revision of the manuscript: D. Kim, M. Shivakumar, M.S. Sinclair, Y.-J. Lee, O.I. Olopade, D. Kim, Y. Lee
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Y. Lee
Study supervision: D. Kim, Y. Lee

Acknowledgments

This work was partially funded by NIGMS grant P50GM115318 and a grant with the Pennsylvania Department of Health (#SAP 4100070267) for Dokyoon Kim. Michael S. Sinclair was supported by National Library of Medicine Training grant T15LM007124. The support and resources from the Center for High Performance Computing and Vice President's Clinical and Translational Research Scholar Program at the University of Utah are gratefully acknowledged. The authors wish to thank Brian Greene, EdD, Scientific Editor and Writer at the Center for Research and Evaluation at the University of Pittsburgh, School of Nursing, for his editorial assistance in the preparation of this manuscript for submission. In addition, we gratefully acknowledge the TCGA Consortium and all its members for the TCGA Project initiative, for providing a sample, tissues, data processing and making data and results available. The results published here are in whole or part based upon data generated by TCGA pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at the following website: http://cancergenome.nih.gov.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 27, 2017; revised September 15, 2017; accepted November 27, 2017; published OnlineFirst January 12, 2018.

References

1. Butte AJ. Translational bioinformatics: coming of age. JAMIA 2008;15:709–14.
2. Altman RB. Translational bioinformatics: linking the molecular world to the clinical world. Clin Pharmacol Ther 2012;91:994–1000.
3. Kumar D, Bansal G, Narang A, Basak T, Abbas T, Dash D. Integrating transcriptome and proteome profiling: strategies and applications. Proteomics 2016;16:2533–44.
4. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793–5.
5. Yan H, Tian S, Slager SL, Sun Z, Ordog T. Genome-wide epigenetic studies in human disease: a primer on -omic technologies. Am J Epidemiol 2016;183:96–109.
6. Grant SG, Chapman VM. Mechanisms of X-chromosome regulation. Ann Rev Genet 1988;22:199–233.
7. Heard E, Martienssen RA. Transgenerational epigenetic inheritance: myths and mechanisms. Cell 2014;157:95–109.
8. Naftelberg S, Schor IE, Ast G, Kornblihtt AR. Regulation of alternative splicing through coupling with transcription and chromatin structure. Annu Rev Biochem 2015;84:165–98.
9. Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. Trends Genet 2015;31:274–80.
10. Gelfman S, Cohen N, Yearim A, Ast G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. Genome Res 2013;23:789–99.

Published OnlineFirst January 12, 2018; DOI: 10.1158/1541-7786.MCR-17-0227

11. Maunakea AK, Chepelev I, Cui K, Zhao K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Res 2013;23:1256–69.
12. Anastasiadou C, Malousi A, Maglaveras N, Kouidou S. Human epigenome data reveal increased CpG methylation in alternatively spliced sites and putative exonic splicing enhancers. DNA Cell Biol 2011;30:267–75.
13. Sammeth M, Foissac S, Guigo R. A general definition and nomenclature for alternative splicing events. PLoS Comput Biol 2008;4:e1000147.
14. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. Genome Med 2015;7:45.
15. Wong JJ, Au AY, Ritchie W, Rasko JE. Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. Bioessays 2016;38:41–9.
16. Wong Justin JL, Ritchie W, Ebner Olivia A, Selbach M, Wong Jason WH, Huang Y, et al. Orchestrated intron retention regulates normal granulocyte differentiation. Cell 2013;154:583–95.
17. Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, et al. Translational control of intron splicing in eukaryotes. Nature 2008; 451:359–62.
18. Lejeune F, Maquat LE. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. Curr Opin Cell Biol 2005;17:309–15.
19. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.
20. Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, et al. RNA sequencing of cancer reveals novel splicing alterations. Sci Rep 2013;3:1689.
21. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;14:R36.
22. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27:2987–93.
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009; 25:2078–9.
24. Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A 2014;111:E5593–601.
25. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 2010;7:1009–15.
26. Huo D, Hu H, Rhie SK, Gamazon ER, Cherniack AD, Liu J, et al. Comparison of breast cancer molecular features and survival by African and European ancestry in the cancer genome atlas. JAMA Oncol 2017; 3:1654–62.
27. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res 2014;24: 1774–86.
28. Sakabe NJ, de Souza SJ. Sequence features responsible for intron retention in human. BMC Genomics 2007;8:59.
29. Wong JJ, Ritchie W, Ebner OA, Selbach M, Wong JW, Huang Y, et al. Orchestrated intron retention regulates normal granulocyte differentiation. Cell 2013;154:583–95.
30. Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB. Inference of splicing regulatory activities by sequence neighborhood analysis. PLoS Genet 2006;2:e191.
31. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
32. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol 2007;3:88.
33. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999;402:C47–52.
34. Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. J Med Genet 2006;43:691–8.
35. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011;12:56–68.
36. Lee Y, Li H, Li J, Rebman E, Achour I, Regan KE, et al. Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. J Am Med Inform Assoc 2013;20:619–29.
37. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–26.
38. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
39. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017; 45:D362–D8.
40. Poon E, Harris AL, Ashcroft M. Targeting the hypoxia-inducible factor (HIF) pathway in cancer. Expert Rev Mol Med 2009;11:e26.
41. Cho EH, Whipple RA, Matrone MA, Balzer EM, Martin SS. Delocalization of gamma-tubulin due to increased solubility in human breast cancer cell lines. Cancer Biol Ther 2010;9:66–76.
42. Kiyama S, Morrison K, Zellweger T, Akbari M, Cox M, Yu D, et al. Castration-induced increases in insulin-like growth factor-binding protein 2 promotes proliferation of androgen-independent human prostate LNCaP tumors. Cancer Res 2003;63:3575–84.
43. Song F, Zhang L, Ji P, Zheng H, Zhao Y, Zhang W, et al. Altered expression and loss of heterozygosity of the migration and invasion inhibitory protein (MIIP) gene in breast cancer. Oncol Rep 2015;33:2771–8.
44. Wang H, Arun BK, Wang H, Fuller GN, Zhang W, Middleton LP, et al. IGFBP2 and IGFBP5 overexpression correlates with the lymph node metastasis in T1 breast carcinomas. Breast J 2008;14:261–7.
45. Wang H, Rosen DG, Wang H, Fuller GN, Zhang W, Liu J. Insulin-like growth factor-binding protein 2 and 5 are differentially regulated in ovarian cancer of different histologic types. Mod Pathol 2006;19:1149–56.
46. Wang H, Wang H, Shen W, Huang H, Hu L, Ramdas L, et al. Insulin-like growth factor binding protein 2 enhances glioblastoma invasion by activating invasion-enhancing genes. Cancer Res 2003;63:4315–21.
47. Spratt DE, Chan T, Waldron L, Speers C, Feng FY, Ogunwobi OO, et al. Racial/ethnic disparities in genomic sequencing. JAMA Oncol 2016;2: 1070–4.
48. Abba MC, Gong T, Lu Y, Lee J, Zhong Y, Lacunza E, et al. A molecular portrait of high-grade ductal carcinoma in situ. Cancer Res 2015;75: 3980–90.
49. Wong JJ, Gao D, Nguyen TV, Kwok CT, van Geldermalsen M, Middleton R, et al. Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. Nat Commun 2017;8:15134.
50. Chlebowski RT, Chen Z, Anderson GL, Rohan T, Aragaki A, Lane D, et al. Ethnicity and breast cancer: factors influencing differences in incidence and outcome. J Nat Cancer Inst 2005;97:439–48.
51. Mohammed SI, Springfield S, Das R. Role of epigenetics in cancer health disparities. Methods Mol Biol 2012;863:395–410.

AACR American Association for Cancer Research

Molecular Cancer Research

Population-dependent Intron Retention and DNA Methylation in Breast Cancer

Dongwook Kim, Manu Shivakumar, Seonggyun Han, et al.

Mol Cancer Res 2018;16:461-469. Published OnlineFirst January 12, 2018.