## **Genome-wide identification of m6A-associated functional SNPs as potential functional susceptibility variants for breast cancer**

## Shicheng Guo1, xxx2, xxx, xxx, xxx, Minghua Wang

1. Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI, USA

Corresponding:

Department of Medical Genetics

School of Medicine and Public Health

University of Wisconsin-Madison, Madison

Tel: 281-685-5882

Email: [Guo.shicheng@marshfieldresearch.org](mailto:Guo.shicheng@marshfieldresearch.org)

Figure counts: 3

Table counts: 1

Run title: Breast cancer associated m6A functional SNPs

## **Abstract**

In the past decades, linkage and association study have identified more than 320 breast cancer susceptibility genes. However, majority of them located in intergenic, intron or UTR regions and functional of these susceptibility SNPs are lack of full understanding. Since m6A RNA-methylation significantly enriched in intron and UTR regions, m6A provided a promising interface to discover the function changes caused by m6A related SNPs.

In this study, we investigate the m6A-SNPs in breast cancer GWAS signals and try to identify m6A associated breast cancer significant GWAS-SNP. We examined the association of m6A-SNPs with femoral neck (FN) and lumbar spine (LS) BMD in 32,961 individuals and quantitative heel ultrasounds (eBMD) in 142,487 individuals. Furthermore, we performed expression quantitative trait locus (eQTL) analyses for the m6A-SNPs using whole genome data of about 10.5 million SNPs and 21,323 mRNAs from 43 Chinese individuals, as well as public available data. Differential expression analyses were also performed to support the identified genes. We found 138, 125, and 993 m6A-SNPs which were associated with FN-BMD, LS-BMD, and eBMD (*P* < 0.05), respectively. The associations of rs11614913 (*P* = 8.92 × 10−10) in *MIR196A2* and rs1110720 (*P* = 2.05 × 10−10) in *ESPL1* with LS-BMD reached the genome-wide significance level. In addition, a total of 24 m6A-SNPs were significantly associated with eBMD (*P* < 5.0 × 10−8). Further eQTL analyses showed that 47 of these BMD-associated m6A-SNPs were associated with expressions of the 46 corresponding local genes. Moreover, the expressions of 26 of these genes were associated with BMD. The present study represents the first effort of investigating the associations and the mechanisms underlying the link between m6A-SNPs and BMD. The results suggested that m6A-SNP may play important roles in the pathology of osteoporosis.

Key words: GWAS, SNP, Epigenetics, m6A, breast cancer

## **Background**

Breast cancer is the most frequently cancer in the female population and more than 12 % of women will be diagnosed with breast cancer in their lifetime. In the past decades, linkage analysis, genome-wide association study (GWAS) and candidate gene based case-control studies have identified more than xx breast cancer susceptibility genes including BRCA1, BRCA2, xx, xxx. However, majority of polymorphism in susceptibility genes did not received well function annotation since 80% of these SNPs do not located in exome regions, but intron, intergenic or untranslated region (UTR). In the post-GWAS era, one of most important task is to make full understanding etiological or pathological mechanism of genetic variations in corresponding diseases. Although, several interesting researches have tried to mapping the genetic variants to specific cell types or tissue types and also tried to connect with the SNPs with enhancer, promoter elements and micro-RNA binding regions(1, 2), majority of these GWAS significant SNPs are still lack of full understanding for their specific functions.

Although there have been tremendous advances in elucidating genetic risk factors underlying both familial and sporadic breast cancer, much of the genetic contribution to breast cancer etiology remains unknown. The discovery of *BRCA1* and *BRCA2* over 20 years ago remains the seminal event in the field and has paved the way for the discovery of other high-penetrance susceptibility genes by linkage analysis. The advent of genome-wide association studies made possible the next wave of discoveries, in which over 80 low-penetrance and moderate-penetrance variants were identified. Although these studies were highly successful at discovering variants associated with both familial and sporadic breast cancer, the variants identified to date explain only 50 % of the heritability of breast cancer. In this review, we look back at the investigative strategies that have led to our current understanding of breast cancer genetics, consider the challenges of performing association studies in heterogeneous complex diseases such as breast cancer, and look ahead toward the types of study designs that may lead to the identification of the genetic variation accounting for the remaining missing heritability.

## **Method**

## **Study Design and Institutional Review Board (IRB)**

According to Marshfield Clinic Research Institute, Research involving the collection or study of existing data, documents, records, pathological specimens or diagnostic specimens and these resources are publicly available or the information is recorded by the investigator in such a manner that subjected cannot be identified, directly or through identifiers linked to the subjects are qualified for IRB exemption. This study is a computational analysis to public dataset and the workflow has been shown in **Figure 1**.

**Identification of m6A-SNPs from previous published and available GWAS data**

In the past decades, more than ~70 GWAS or follow-up study have been conducted and identified more than xx breast cancer associated susceptibility variants. The most recently and largest GWAS data published by Michailidou and colleagues on Nature(3) in 2017 was collected in the study. As the largest GWAS study, 122,977 breast cancer and 105,974 controls of European Ancestry (EA) and 14,068 breast cancer and 13,104 controls of East Asian Ancestry (EAA) were enrolled and 11,792,542 reliable variants (r2>0.3 and MAF>0.5%) after imputation with 1000 Genome phase 3 sequencing reference panel. Summary statistics were downloaded (website see data availability section). m6A-SNPs were downloaded from m6Avar database(4) on November 13rd 2019. I also prepared a portable m6A-SNPs involved in the GWAS data and deposited to Github which will be easy for further analysis (<https://github.com/cnaid/gwas/tree/master/breastcancer/m6A>).

**eQTL analysis to m6A-SNPs with Genotype-Tissue Expression (GTEx) dataset**

Single-Tissue cis-QTL Data from GTEx project were downloaded from https://gtexportal.org/home/datasets on December 9th 2019 and Perl script was prepared to transfer signfi\_variant\_gene\_pairs (SVGP) id to dbSNP id. Perl script can be downloaded from above mentioned Github page.

**Differential gene expression analysis to m6A-SNPs related genes with TCGA Pan-cancer dataset**

I downloaded 11,093 gene expression quantification data derived from RNA-seq data in TCGA database (<https://portal.gdc.cancer.gov/repository>) on February 24th, 2019. The RNA-seq data covered 32 cancer types. However 9 cancer types were excluded since the low samples size for control samples (N<=1, Supplementary Table S1). Log2-transformed fragments per kilobase of transcript per million mapped reads upper quartile (FPKM-UQ) derived from HTSeq(5), is applied for differential gene expression analysis. Bayesian generalized linear model (bayesglm) from ARM package (v1.10-1) was applied for differential gene expression analysis. metafor package (v2.1-0) was applied for meta-analysis cross 23 cancer types. Cox proportional hazards regression model was applied for survival analysis to TCGA overall survival times (R survival package v0.9).

## **Result**

**Identification of m6A-associated SNPs as potential functional susceptibility variants for breast cancer**

We collected summary statistics of 11,792,565 SNPs from most recent breast cancer genome-wide association study including 137,045 case and 119,078 control of both European and East Asian ancestry(3). We then collected all 312,363 m6A-SNPs from m6A database so that we can investigate functional m6A-SNPs in breast association study. Finally, we identified 6,010 m6A-SNPs in breast cancer association study including xx located in xx, xxx located in xxx.

The first step of this study was to pick out m6A-SNPs from the RA-GWAS dataset which containing 6.6 million SNPs according to the annotation information of the 313 thousand m6A-SNPs in the m6AVar database. We found 3,883 unique m6A-SNPs which located not only in protein-coding genes, but also non-coding RNAs. Among these SNPs, 227 and 308 (476 unique) were nominally (P < 0.05) associated with RA in Asians and Europeans, respectively. Specifically, 9 and 32 (38 unique) m6A-SNPs were significant at genome-wide level (P < 5.0 × 10−8) in Asians (Figure1A) and Europeans (​Figure1B), respectively. Among these genome-wide significant m6A-SNPs, the effects of rs1046080, rs13978, rs2736158, and rs58892873 on m6A were confirmed by miCLIP/PA-m6A-Seq experiments. The effects of another 8 SNPs on m6A were confirmed by MeRIP-Seq experiments. The rest 26 fell within the low confidence categories (predicted to be associated with m6A). The effects of the predicted SNPs on m6A modification have not been validated by technical and biological experiments. Further experiments are needed to determine their effects in large scale studies (e.g., QTL mapping).

eQTL analysis to m6A-associated functional SNPs in Genotype-Tissue Expression (GTEx) dataset

Gene expression analysis to m6A-associated functional SNPs in Pan-cancer mRNA sequencing dataset.

## **Discussion**

In the past two decades, genetic etiology research to breast cancer has received inspiring achievement. However, common susceptibility variants identified by GWAS can only explain less than 20% familial relative risk indicating missing heritability problem are still require further investigation. Different research strategies have been conducted to identify novel susceptibility variants from different genetic variations including rare variants(6), copy number variation(7), even epigenetics elements(2) including miRNAs. In this study, we provided a novel angel to identify functional variants that are related to m6A methylation in UTR and intron regions.

## **Availability of data and materials**

**GWAS**: <http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/gwas-icogs-and-oncoarray-summary-results/>

**eQTL**: <https://www.gtexportal.org/home/datasets>

**TCGA**: <https://portal.gdc.cancer.gov/>

**Roadmap Project**: <http://www.roadmapepigenomics.org/data/>

**R**: <https://www.r-project.org/>

**m6A-SNPs**: <http://m6avar.renlab.org/>

**m6A-SNP-BRCA**: <https://github.com/cnaid/gwas/tree/master/breastcancer/m6A>

**SVGP2RSID**:

All other data used in the study were available upon the reader’s request.

## **Authors’ contributions**

SG designed the study and analyzed the data, prepared the draft and submission of the final manuscript.

## **Disclosure of Conflicts of Interest**

The authors declare no conflict of interest.

## **Funding**

Dr. Guo was supported by the Center for Precision Medicine Research, Marshfield Clinic Research Institute.

**Acknowledgements**

The breast cancer genome-wide association analyses were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the ‘Ministère de l’Économie, de la Science et de l’Innovation du Québec’ through Genome Québec and grant PSR-SIIRI-701, The National Institutes of Health (U19 CA148065, X01HG007492), Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710) and The European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935).

## **Reference**

1. Fan L, Chen L, Ni X, Guo S, Zhou Y, Wang C, Zheng Y, et al. Genetic variant of miR-4293 rs12220909 is associated with susceptibility to non-small cell lung cancer in a Chinese Han population. PLoS One 2017;12:e0175666.

2. Shen F, Chen J, Guo S, Zhou Y, Zheng Y, Yang Y, Zhang J, et al. Genetic variants in miR-196a2 and miR-499 are associated with susceptibility to esophageal squamous cell carcinoma in Chinese Han population. Tumour Biol 2016;37:4777-4784.

3. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, et al. Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551:92-94.

4. Zheng Y, Nie P, Peng D, He Z, Liu M, Xie Y, Miao Y, et al. m6AVar: a database of functional variants involved in m6A modification. Nucleic Acids Res 2018;46:D139-D145.

5. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166-169.

6. Zhou X, Li H, Guo S, Wang J, Shi C, Espitia M, Guo X, et al. Associations of Multiple NOTCH4 Exonic Variants with Systemic Sclerosis. J Rheumatol 2019;46:184-189.

7. Guo S, Li Y, Wang Y, Chu H, Chen Y, Liu Q, Guo G, et al. Copy Number Variation of HLA-DQA1 and APOBEC3A/3B Contribute to the Susceptibility of Systemic Sclerosis in the Chinese Han Population. J Rheumatol 2016;43:880-886.

## **Figure legends**

Figure 1.

Figure 2

Figure 3

Figure 4.