Archive of SID
Journal of Cell and Molecular Research (2020) 12 (1), 22-32 DOI: 10.22067/jcmr.v12i1.87629

Research Article

Identification of Breast Cancer Associated Putative Functional Single Nucleotide Polymorphisms in the Iranian Population through *In Silico* Analyses

Vida Nadafi Sichani[1], Seyed Alireza Emami[2], Morteza Bitaraf Sani[3]*, Mohammadreza Nassiri[4], Vinod Gopalan[5]

[1] *Department of Biology, Faculty of Science, University of Science and Art, Yazd, Iran*
[2] *Department of Biology, Faculty of Basic Sciences, Shahrekord Branch, Islamic Azad University, Shahrekord, Iran*
[3] *Animal Science Research Department, Yazd Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education & Extension Organization (AREEO), Yazd, Iran*
[4] *Recombinant Proteins Research Group, Research Institute of Biotechnology and Genetic Group, Ferdowsi University of Mashhad, Mashhad, Iran*
[5] *School of Medicine and Medical Science, Menzies Health Institute Queensland, Griffith University, Gold Coast, 4222*

Received 8 July 2020 Accepted 22 August 2020

Abstract

Previous studies have found several distinct alleles at both levels of transcriptional activity and protein-DNA binding manners in breast cancer patients vs. healthy individuals through multi-step experimental approaches. This study presents a computational-based model to investigate the regulatory potential and functional properties of disease-related non-coding single nucleotide polymorphisms (SNPs) variants through several online *in silico* tools in the Iranian population. The association between the risk of breast cancer and its putative single nucleotide polymorphisms in the Iranian population was investigated through SNPedia database and genome-wide association studies (GWAS). Furthermore, a meta-analysis was performed by Comprehensive Meta-Analysis (CMA) software. Functional analyses were carried out through LDlink, HaploReg, and RegulomeDB. The impact of each single nucleotide polymorphism on gene expression profiles and transcription factor binding sites were predicted by the RegulomeDB. "5", "6", and "1d" scores were assigned to rs3746444, rs1062577, and rs1049174 by this scoring system, respectively. RegulomeDB scores of rs3746444-*MYH7B/MIR499A* and rs1062577-*ESR1* suggested that they are not putative functional single nucleotide polymorphisms; and may not associate with significant eQTL signals. The "1d" score for rs1049174-*RP11-277P12.20* confirmed an association with the expression of the target gene. Proxy variants rs6088678 and rs2617160 have been identified using LDlink in non-coding segments. They were in strong linkage disequilibrium (LD) with single nucleotide polymorphisms rs3746444 and rs1049174, respectively. Also, non-coding variants rs6088678-*TRPC4AP* and rs2617160- *RP11-277P12.20* with high-ranked scores showed the strongest related-expression. This work provides a rapid and direct *in silico*-based approach for the identification of functional genetic variants in the breast cancer. These analyses were conducted to evaluate the association of intended SNPs with the regulatory elements of histones, DNases, motif changes, and selected eQTL signals. It can be extended to some other complex single nucleotide polymorphism-related diseases.

***Keywords*:** Epigenetics, Functional single nucleotide polymorphisms, Genome-Wide Association Studies, Linkage Disequilibrium, RegulomeDB scoring system.

Introduction

Single Nucleotide Polymorphisms (SNPs) represent the most common markers of the genome diversity among individuals (Coetzee et al., 2012). The overwhelming majority of significantly associated genetic variants identified through GWAS were drop down outside of the coding area. Hence, it is difficult to understand how specific SNP increases disease susceptibility (Meng et al., 2018). Single nucleotide polymorphisms have a crucial role in the prediction of the risk of various complicated diseases including cancer (Seyedmir et al., 2017). Cis-regulatory regions (non-coding DNA regions) comprise distal elements such as promoters, enhancers, and insulators, which regulate transcriptional activities and complex spatial and temporal gene expressions following the binding of

*Corresponding author's e-mail address: *m.bitaraf@areeo.ac.ir*

www.SID.ir
Licensed as Open Access CC-BY

Archive of SID

transcription factors (Bauer-Mehren et al., 2009).

In addition, the majority of epigenetic changes may be reversible or preventable. So, the restoration of epigenetic changes could be applied as a proper strategy for cancer treatment or prevention (Coetzee et al., 2012). There are highly advanced web-based tools with the capacity for the annotation of a specific SNP to a target gene. Also, it is possible to measure the causal risk among numerous non-coding loci before performing time-consuming validation experiments. Such experiments will enable us to accurately predict the likelihood of particular cancer risk for individuals or communities (Coetzee et al., 2012). These types of studies are based on two hypotheses: I) alterations in the regulatory areas are major determinants of gene expression modifications. II) motifs in regulatory regions exhibit a location preference (e.g. at the center of H3K27ac, H3K4me1 or DNase peak (Meng et al., 2018). It has been observed that the chromatin status of enhancers is determined by highly specific histone modification patterns which are strongly linked to cell-type-specific gene expression programs on a global scale.(Heintzman et al., 2009) Along with H3K4me1, a general signal for enhancers, H3K27ac enrichment is also dedicated to the identification of active enhancers. Sequences with high H3K4me1 enrichment, and low H3K27ac are considered as ready-to-activate enhancers and are associated with low gene expression levels (Rhie et al., 2013). Hence, in the present study, in line with these kind of experiments, a comprehensive in silico study was conducted based on the application of computational-based methods including RegulomeDB, HaploReg, and LDlink. Encyclopedia of DNA Elements (ENCODE, from ChIP-seq experiments), and Roadmap Epigenomics (from methods such as ChromHMM) were utilized as data resources (Edwards et al., 2013). We selected breast cancer as the phenotype of choice among others during genome-wide association studies (GWAS). A list of three breast cancer risk-associated SNPs was obtained from the GWAS Catalog and SNPedia. Our purpose was to determine the functional value of rs3746444, rs1062577, and rs1049174 SNPs which were obtained through wet-lab experiments in the Iranian population.

Indeed, we performed pairwise comparisons with functional proxy variants suggested by the LDlink application. LDlink (analysistools.nci.nih.gov) is a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants (Machiela et al., 2015). Due to the importance of linkage disequilibrium (LD) structures in the indigenous populations, LDlink was utilized for finding two proxy variants to be compared with query variants in the Iranian population. We found coding proxy variants with high RegulomeDB scores which do not have any functional effect on regulatory regions. On the other side, a non-coding proxy variant with low RegulomDB score was selected. RegulomeDB variant classification scheme is fully described by Boyle et al. (Boyle et al., 2012).

HaploReg v4.1 is another web-based tool for exploring the annotations and producing mechanistic hypotheses of the impact of noncoding variants on the clinical phenotypes and normal variations (Fayez et al., 2018).

Materials and Methods

Selection of SNPs

In the present study, SNPs associated with breast cancer risk in the Iranian population were identified through the SNPedia (www.snpedia.com) and GWAS Catalog (www.ebi.ac.uk›gwas). These detected SNPs include rs3746444 (Kabirizadehet al., 2015), (Jiang et al., 2015), (Zou et al., 2012), (Mu et al., 2017), (Wang et al., 2012), (Wang et al., 2012), rs1062577(Dehghan et al., 2017), (Chen et al., 2016), and rs1049174 (Ghobadzadeh et al., 2013). Different parameters including odds ratios (OD), confidence interval (CI), number of samples, author's name, and host countries for these SNPs were extracted from relevant literature to conduct a comprehensive meta-analysis. The best and most effective SNPs were selected for downstream procedures.

In-silico studies

LDlink (www.ldlink.nci.nih.gov), HaploReg (www.pubs.broadinstitute.org/mammals/haploreg/haploreg.pbp) and RegulomeDB (http://www.regulomedb.org) web tools and databases were applied to determine the functional value of desired polymorphisms. Figure 1 is outline of our processing pipeline.

www.SID.ir

Archive of SID

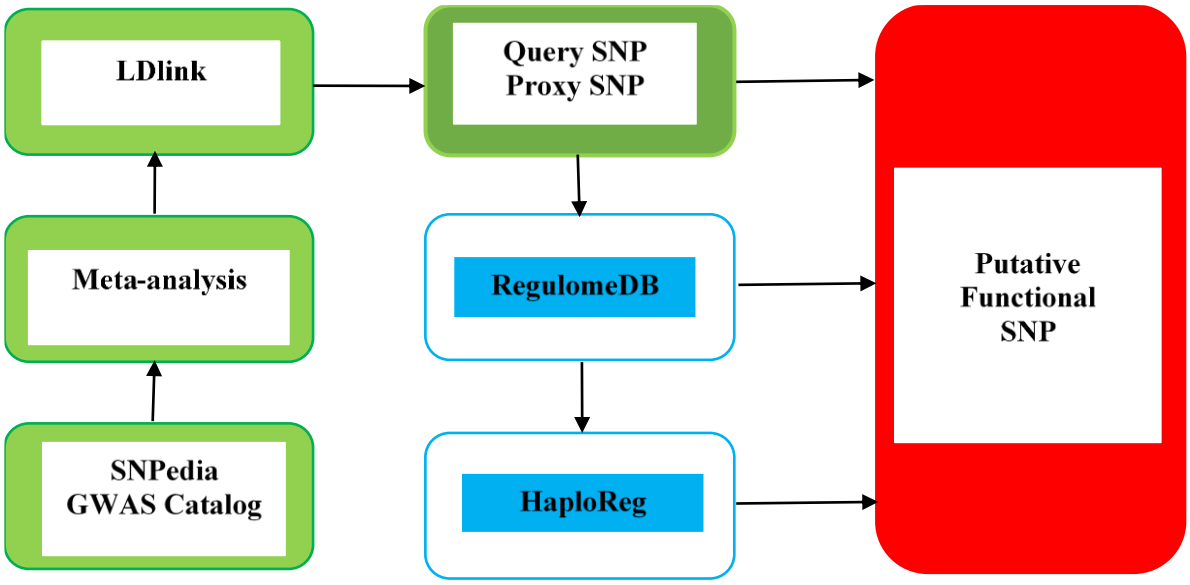

Figure 1. The pipeline consists of various key points including methods of SNP collection (SNPedia and GWAS Catalog), comprehensive meta-analysis (CMA), investigating the patterns of linkage disequilibrium across a variety of ancestral populations (LDlink), and developing the mechanistic hypothesis of the impact of non-coding variants on the clinical phenotypes (RegulomeDB and HaploReg). At the final step, we endeavored to confirm whether these SNPs are located in the regulatory segments and have functional impact on the gene expression patterns (Putative Functional SNP).
SNPedia: wiki-based bioinformatics web site that serves as a database of single nucleotide polymorphisms (SNPs). NHGRI-EBI GWAS Catalog: publicly available resource of Genome Wide Association Studies (GWAS) and their results. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. HaploReg: a tool for exploring annotations of the noncoding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease associated loci. RegulomeDB: a database that annotates SNPs with known and predicted regulatory elements in the intergenic regions of the H. sapiens genome. Known and predicted regulatory DNA elements include regions of DNase hypersensitivity, binding sites of transcription factors, and promoter regions that have been biochemically characterized to regulation transcription. Sources of these data include public datasets from GEO, the ENCODE project, and published literature. Query SNP: variant RS number - RS number for query variant. RS number must match a bi-allelic variant. Table of proxy variants: by default, the ten variants with the highest R2 values and closest distance to the query variant are displayed. External links lead to the variant RS number in dbSNP, coordinates in the UCSC Genome Browser, and regulatory information (if any) in RegulomeDB.

The LDlink web tool was used to detect proxy SNPs with strong LD (≥ 0.8) for rs3746444, rs1062577, and rs1049174. Proxy SNPs with these properties were selected for further analysis: I) coding SNPs with high RegulomeDB scores (4-6) and the least evidences for binding to regulatory proteins and participation in the gene expression regulation. II) non-coding SNPs with low RegulomeDB score (1a-1f) and the most evidences for binding to regulatory proteins and participation in the gene expression regulation. In addition, LD hap option of the LDlink was applied to evaluate haplotype frequencies between input SNPs and proxy SNPs.
Histone modifications in human tissues relevant to the breast cancer, such as breast myoepithelial primary cells (MEPs) and breast variant human mammary epithelial cells (vHMECs) (Fayez et al., 2018) were investigated by the HaploReg v4.1 tool. HaploReg was used to explore the annotations of the non-coding genome at variants on haplotype blocks, such as candidate regulatory SNPs at disease-related loci (Hamdi et al., 2018). The impact of genetic changes on different tissues and biological systems was revealed using HaploReg, RegulomeDB, and the genotype-tissue expression (GTEx) portal (Ward et al., 2011).

Results

Identification of single nucleotide polymorphisms

The meta-analysis results using CMA software are shown in the Table 1. The level of rs3746444 effect on the breast cancer predicted as "low". The interactive plot for rs3746444 represents the high density of SNPs with high LD around this variant (Figure 2).

www.SID.ir

Archive of SID

Using LDlink web tool it was confirmed that two proxy SNPs (rs3746435 and rs6088678) with strong LD 0.87 with query SNP rs3746444, are associated with the breast cancer pathogenesis (Table 2). RegulomeDB score "5" was obtained for SNP rs3746435, located in the coding region. Meanwhile, based on this scoring system, low score of "1f" was measured for rs6088678 non-coding variant, implying a higher level of functional properties (Table 2).

It was confirmed by the application of HaploReg that rs3746444 induces histone modification H3K4me1_Enh in the breast myoepithelial primary cells and is located in the DNase I hypersensitive region of the breast variant of human mammary epithelial cells. It shows that motif changes may allow the DNase I to identify the available chromatin and cuts DNA at its respective region (Table 3). Although, proxy SNP rs3746435 (missense) with score "5", did not results in any histone modifications in the examined breast cancer cell lines (Table 3). It was also shown based on the results from HaploReg tool that proxy SNP rs6088678 caused the histone modification H3K9ac_ Pro. It is located within the promoter region and transcription start sites (TSS) and was effective in the regulation of *TRPC4AP* expression (Table 3). Query SNPs or SNPs previously identified as breast cancer associated ones in the Iranian population are marked in bold.

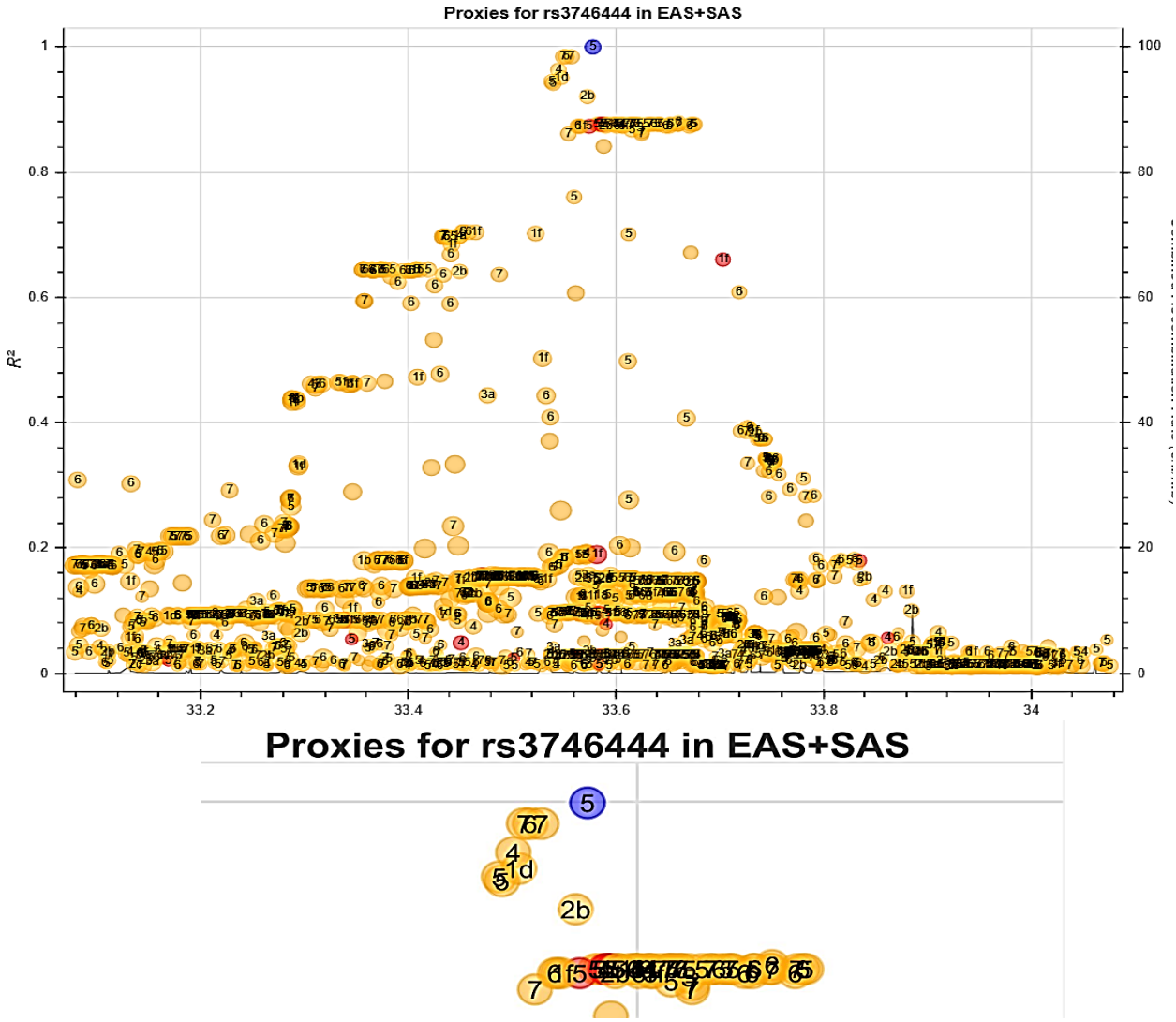

Figure 2. The interactive plot obtained for rs3746444(blue circle with RegulomeDB score=5) by the application of LDlink web tool. The complete and high-resolution chart could be viewed through the given link. (https://ldlink.nci.nih.gov/?var=rs3746444&pop=CHB%2BJPT%2BCHS%2BCDX%2BKHV%2BGIH%2BPJL%2BBEB%2BSTU%2BITU&r2_d=r2&window=500000&tab=ldproxy). Interactive plot: interactive plot of query variant(rs3746444) and all bi-allelic dbSNP variants plus or minus 500 kilobases (Kb) of the query variant(rs3746444). X axis is the chromosomal coordinates and the Y axis is the pairwise R2 value with the query variant as well as the

www.SID.ir

Archive of SID

combined recombination rate from HapMap. Each point represents a proxy variant and is colored based on function, sized based on minor allele frequency, and labeled based on regulatory potential (regulatory potential number of rs3746444=5). Hovering over the point will display detailed information on the query and proxy variants. Reference population(s)((SAS) South Asian and (EAS) East Asian): selected from the drop-down menu. At least one 1000 Genomes Project sub-population is required, but more than one may be selected. R2/D' toggle: select if desired output is based on estimated R2 or D'.

Table 1. Results from meta-analysis of association studies for rs3746444.

Reference study	Odds ratios	Lower limit	Upper limit	z value	*p* value
Wang, Y., Yang, B. and Ren, X. (2012) Hsa-miR-499 polymorphism (rs3746444) and cancer risk: a meta-analysis of 17 case–control studies. Gene 509(2): 267-272.	1.230	1.059	1.429	2.710	0.007
Mu, K., Wu, Z. Z., Yu, J. P., Guo, W., Wu, N., Wei, L. J. and Liu, J. T. (2017) Meta-analysis of the association between three microRNA polymorphisms and breast cancer susceptibility. Oncotarget 8(40): 68809.	1.170	1.025	1.336	2.319	0.020
Wang, L., Qian, S., Zhi, H., Zhang, Y., Wang, B. and Lu, Z. (2012) The association between hsa-miR-499 T> C polymorphism and cancer risk: a meta-analysis. Gene 508(1): 9-14.	1.160	0.995	1.353	1.892	0.058
Zou, P., Zhao, L., Xu, H., Chen, P., Gu, A., Liu, N. and Lu, A. (2012) Hsa-mir-499 rs3746444 polymorphism and cancer risk: a meta-analysis. Journal of biomedical research 26(4): 253-259.	1.100	1.004	1.205	2.049	0.040
Jiang, S. G., Chen, L., Tang, J. H., Zhao, J. H. and Zhong, S. L. (2015) Lack of association between Hsa-Mir-499 rs3746444 polymorphism and cancer risk: meta-analysis findings. Asian Pacific Journal of Cancer Prevention 16(1):339-344.	1.180	1.035	1.346	2.466	0.014
Kabirizadeh, S., Azadeh, M., Mirhosseini, M., Ghaedi, K. and Tanha, H. M. (2016) The SNP rs3746444 within mir-499a is associated with breast cancer risk in Iranian population. Journal of Cellular Immunotherapy 2(2): 95-97.	1.922	1.064	3.471	2.167	0.030
	1.157	1.094	1.223	5.149	0.000

Data obtained by the LDlink revealed that rs3746444, which is located in a non-coding segment, indicated a RegulomeDB score "5" (Table 2). It seems that rs3746444 does not exhibit any significant biological activity such as alterations in the transcription factors binding capacity and gene regulatory effects in the Iranian population.

Table 2. Details of putative regulatory functions of query SNPs and their associated proxy SNPs.

cVariant	LD (r^2)	LD (D')	ASN freq	Enhancer histone marks	DNase	dbSNP func annot	GEN CODE genes	RegulomeDB score	Predicted function
rs3746444	1	1	0.17	IPSC, GI, MUS	BRST, BRN, LIV	Intronic	MYH7B & MIR499A	5	TF binding or DNase peak
rs3746435	0.87	1.0	0.12	SKIN, MUS	SKIN, PLCNT	Missense	MYH7B	5	TF binding or DNase peak
rs6088678	0.87	1.0	0.12	8 tissues	-	Intronic	TRPC4AP	1f	eQTL+TF binding / DNase peak
rs1062577	1	1	0.27	-	-	3'-UTR	ESR1	6	Motif hit
rs1049174	1	1	0.60	-	8 tissues	3'-UTR	RP11-277P12.20	1d	eQTL+TF binding+any motif+ DNase peak
rs2617160	0.88	0.9	0.59	10 tissues	5 tissues	Intronic	RP11-277P12.20	1f	eQTL+TF binding / DNase peak

A comparison of several factors between query and proxy variants with high LDs has been performed in the Asian population.

ASN freq: allele abundance in Asian population. SNP functional annotation: the functional area where mentioned SNP is located. GENCODE genes: the gene region in which SNP is located. RegulomeDB score: Regulome DB is a database

www.SID.ir

Archive of SID

that scores SNPs functionality based upon experimental data. It is necessary to mention that in all tables Query SNPs are displayed in bold.

Table 3. Regulatory chromatin status from DNase and histone ChIP-Seq (Roadmap Epigenomics Consortium, 2015).

variant	Group	Description	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase
rs3746444	Epithelial	Breast Myoepithelial Primary Cells	H3K4me1_Enh	-	-	-	-
rs3746444	Epithelial	Breast variant Human Mammary Epithelial Cells (vHMEC)	-	-	-	-	DNase
rs3746435	Epithelial	Breast Myoepithelial Primary Cells	-	-	-	-	-
rs3746435	Epithelial	vHMEC	-	-	-	-	-
rs6088678	Epithelial	Breast Myoepithelial Primary Cells	-	-	-	H3K9ac_Pro	-
rs1062577	Epithelial	vHMEC	H3K4me1_Enh	-	-	-	-
rs1049174	Epithelial	Breast Myoepithelial Primary Cells	H3K4me1_Enh	-	-	-	-
rs1049174	Epithelial	vHMEC	H3K4me1_Enh	-	-	-	-
rs2617160	Epithelial	Breast Myoepithelial Primary Cells	H3K4me1_Enh	-	-	-	-
rs2617160	Epithelial	vHMEC	H3K4me1_Enh	-	-	-	DNase

Query SNPs have been compared with proxy SNPs in terms of cellular and histological position. Histone modifications that each one creates in the target cells has been investigated.
Open chromatin: DNase1 hypersensitivity. Histone modifications: H3K4me1, H3K4me3, H3K9ac, H3K27ac. It is necessary to mention that in all tables Query SNPs are displayed in bold.

Increased frequency of haplotype AGC

The LD hap analysis (http://analysistools.nci.nih.gov/LDlink/tab=ldhap) showed increased frequency (79%) of AGC haplotype among three SNPs including rs3746444, rs3746435, and rs6088678. Results indicated that when the query SNP is adenine, the proxy allele for rs3746435 and rs6088678 will be G and C, respectively. There was a very strong LD among these three SNPs (87%). Also, the abundance of the AGC haplotype was high. It was revealed that allele A in query SNP rs3746444 is more likely to be associated with allele G; while if the query allele is G, it is likely that the proxy allele would be C (Figure 3).

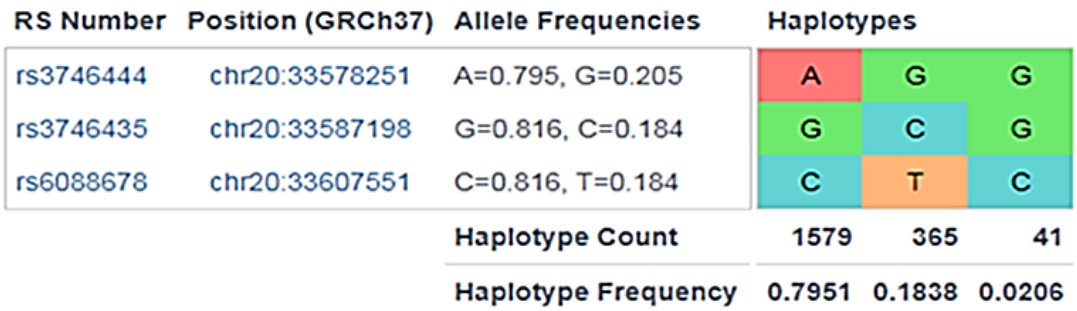

Figure 3. Haplotype Analysis of query SNP rs3746444 with two proxy SNP rs3746435 and rs6088678. Results obtained from haplotype study of SNPs using LDlink web-based tool indicate that when the query SNP is adenine, the proxy allele for rs3746435 and rs6088678 will be G and C, respectively. There is a very strong LD among these three SNPs (87%).
(http://analysistools.nci.nih.gov/LDlink/tab=ldhap)

www.SID.ir

Archive of SID

Association of SNPs with transcriptional levels of the target genes

Polymorphism rs3746444-*MYH7B/MIR499A* induces a poor transcriptional level in the breast MEPs and vHMECs, which in turn, will be resulted in the formation of a weak polycomb complex and reduced regulatory effects of the target gene. On the other side, rs3746435 induces a strong transcription in the examined cell lines. The proxy SNP rs6088678-*TRPC4AP* showed a strong transcription in addition to the score "1f" in both cell lines (Table 4). As demonstrated in Fig. 4, the non-coding proxy SNP rs6088678 with low score "1f", indicated the highest expression level in the breast tissue. Its value was equal to 40-60 Reads Per kilobase Million (RPKM) (Figure 4).

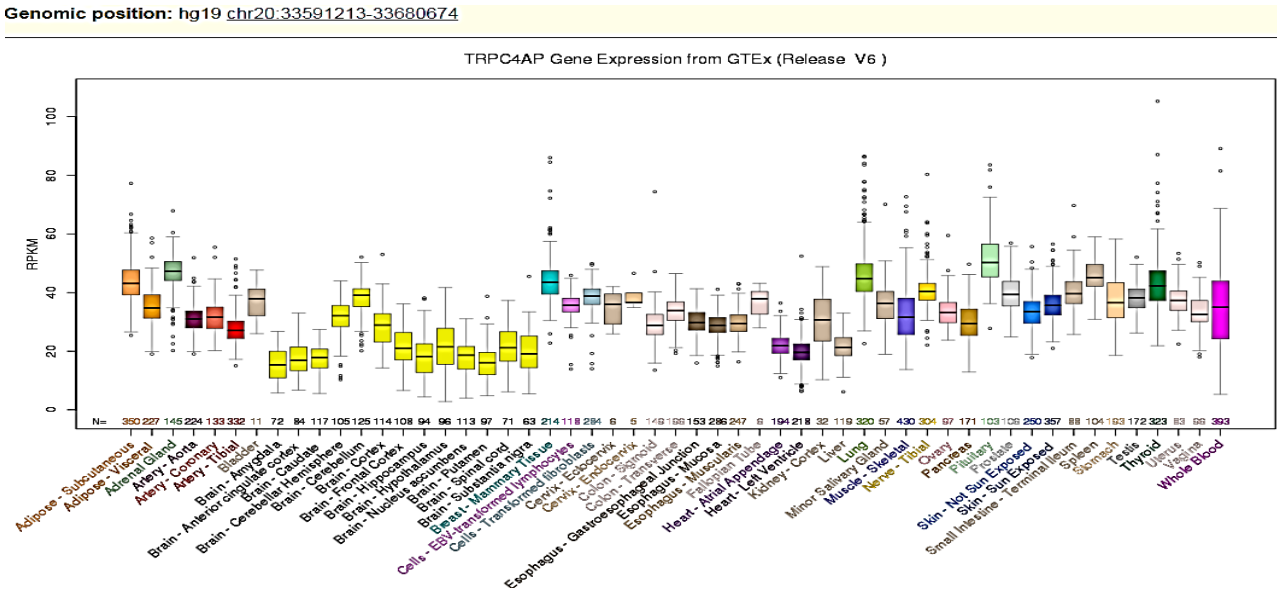

Figure 4. TRPC4AP gene expression from (GTEx) project for rs6088678.
The non-coding proxy SNP rs6088678 with low score "1f", indicates the highest expression level in the breast tissue (red arrow). Its value is equal to 40-60 Reads Per kilobase Million (RPKM).

Table 4. Genome browser, chromatin state and accessibility.

Method	SNP	Location	Chromatin State	Tissue Group	Tissue
ChromHMM	rs3746444	chr20:33575200..33578600	Weak transcription	Epithelial	Breast Myoepithelial Primary Cells
ChromHMM	rs3746444	chr20:33574000..33583000	Weak Repressed PolyComb	Epithelial	Breast variant Human Mammary Epithelial Cells (vHMEC)
ChromHMM	rs3746435	chr20:33583600..33645600	Strong transcription	Epithelial	Breast Myoepithelial Primary Cells
ChromHMM	rs3746435	chr20:33583000..33590400	Quiescent/Low	Epithelial	vHMEC
ChromHMM	rs6088678	chr20:33583600..33645600	Strong transcription	Epithelial	Breast Myoepithelial Primary Cells
ChromHMM	rs6088678	chr20:33603600..33608000	Strong transcription	Epithelial	vHMEC

www.SID.ir

Archive of SID

ChromHMM	rs1062577	chr6:152398400..152431600	Quiescent/Low	Epithelial	Breast Myoepithelial Primary Cells
ChromHMM	rs1062577	chr6:152423600..152425200	Weak transcription	Epithelial	vHMEC
ChromHMM	rs1049174	chr12:10524400..10528200	Weak transcription	Epithelial	Breast Myoepithelial Primary Cells
ChromHMM	rs1049174	chr12:10525000..10525600	Enhancers	Epithelial	vHMEC
ChromHMM	rs2617160	chr12:10544600..10549000	Enhancers	Epithelial	Breast Myoepithelial Primary Cells
ChromHMM	rs2617160	chr12:10544800..10546400	Enhancers	Epithelial	vHMEC

ChromHMM (Hidden Markov Model) is applied to annotate the non-coding genome using epigenomic information between one or multiple cell types. Using RegulomeDB web-based tool, the transcription level of Query SNPs and proxy SNPs in different tissues and cell types has been determined. It is necessary to mention that in all tables Query SNPs are displayed in bold.

The NIH genotype-tissue expression (GTEx) project was created to establish sample and data resources for studies aimed to unravel the relationships between genetic variations and gene expression levels in multiple human tissues. This track shows median gene expression levels in 51 tissues and 2 cell lines, based on the RNA-seq data from the GTEx midpoint milestone data release (V6, October 2015). This release is formed based on the data from 8555 tissue samples obtained from 570 adult post-mortem individuals.

All the regulatory features which were seen in tables were obtained from ENCODE and NIH Roadmap Epigenomics data through the UCSC Genome Browser.

SNP rs1062577

Meta-analyses were not possible for query SNP rs1062577-*ESR1* due to the limited number of studies which carried out on rs1062577 in the Asia. The interactive plot in LDlink tool revealed that there is no SNP with strong LD (≥ 0.8) around rs1062577. RegulomeDB score of "6" revealed no remarkable effect on the gene expression levels (Table 2). However, it shows the target gene (*ESR1*) expression of rs1062577 in the breast cancer tissues. There are overwhelming data on the expression of *ESR1*, up to about 50 RPKM. Indeed, this polymorphism induces H3K4me1_Enh histone modification in vHMECs (Table 3). However, there was little evidences for rs1062577 to be a functional noncoding SNP.

SNP rs1049174 versus rs2617160

The interactive plot from LDlink tool, indicated low density of SNPs with strong LD around rs1049174. The non-coding proxy SNP rs2617160, located in the intronic region with score "1d", was selected for further analysis (Table 2). There was no coding SNP with strong LD for rs1049174. Both of the query SNP rs1049174, and proxy SNP rs2617160 caused H3K4me1_Enh histone modification in the investigated cell lines and were associated with the breast cancer tissue.

In contrast, proxy SNP rs2617160, thorough the induction of motif changes, produces open chromatin regions in vHMECs. Hence, DNase I can cut DNA in its respective region (Table 3). Both rs1049174, and 2617160 which were submerged in the RegulomeDB tool in addition to the proxy SNP rs2617160 are located in *RP11-277P12.20* enhancer sites of the examined breast cancer cell lines. rs1049174 caused a poor transcriptional level in the breast MEPs and is specifically located in the enhancer of the vHMECs (Table 4).

Discussion

Previous studies demonstrated that most of the GWAS variants fall in non-coding (nc) regions. The identification of the functions of these ncSNPs remains as a major challenge. The importance of understanding the functional contributions of specific risk variants to disease pathogenesis is widely accepted (Rhie et al., 2013). The biological effects of the most already studied SNPs in the Iranian population were not strong. In the present study through the application of a set of *in silico*

www.SID.ir

Archive of SID

approaches, functional analyses were performed for previously known breast cancer risk associated SNPs in the Iranian population. The HaploReg database was established as a computer simulation tool by Ward and Kellis (Ward et al., 2011) to provide an intersects of single nucleotide variants (SNVs) with chromatin status (Ernst et al., 2010). For the first time, this work demonstrated that a comprehensive *in silico* analysis of well-known ncSNPs and regulatory regions is essential before we can attribute them to the Iranian population.

It was previously reported that rs3746444 (Kabirizadeh et al., 2016), rs1062577 (Dehghan et al., 2017), and rs1049174 (Ghobadzadeh et al., 2013) are associated with an increased risk of breast cancer in the Iranian population. We focused on non-coding proxy SNPs (LD≥0.8 with query SNPs rs3746444, rs1062577, and 1049174) which were obtained from LDlink. It was assumed that all non-coding variant SNPs which are located in the regulatory regions (promoter, enhancer, 5`UTR, 3`UTR) have a highly ranked RagulomeDB score (Table2). The meta-analysis of the rs3746444 in the Asian and Iranian population indicated a statistically significant relationship with the breast cancer by Odds Ratio(OR) = 1.15(1.09-1.22). These analyses were only possible for one SNP (Table 1).

Moreover, the regulatory effects of rs3746444-*MYH7B/MIR499A*, rs1062577-*ESR1*, and rs1049174-*RP11-277P12.20* and their related proxy SNPs were determined based on the high LD. We apply this analysis to identify the most likely functional variant among *MYH7B*, *ESR1*, and *RP11-277P12.20* genes. However, a solid framework of the functional significance of variants cannot be obtained by a single bioinformatics tool. Hence, some complementary tools were applied to perform the current study. Three computational-based tools including LDlink, HaploReg, and RegulomeDB were used for above mentioned SNPs in a combinatory mode to prioritized ncSNPs for their association with the disease status. The LD structure haplotype block for the Iranian population was not available because GWAS studies have not been performed previously in Iran. Hence, related information from the Asian population were utilized as a reference for LDlink studies.

We identified query SNP rs1049174 in 3'UTR region as the only previously wet-lab studied SNP with high ranked RegulomeDB score "1d" and validated functional effects (eQTL+TF binding+any motif+ DNase peak) (Table 2). rs1049174 caused histone modification H3K4me1 in both cancerous cell lines. It confirms that these enhancers are ready to be active.

The present study demonstrated that SNPs in the *MYH7B*, *TRPC4AP* and *RP11-277P12.20* genes (Table 2) in addition to the ncSNPs rs6088678, and rs2617160 are functionally important. Although, wet-lab experiments are essential for the validation of the results. Pairwise comparisons confirmed that intronic SNP rs6088678 (r^2 = 0.87 with rs3746444) and RegulomeDB score "1f" showed more evidences of being functional in comparison to rs3746444 (Table 2). It was shown that the rs6088678 induced histone modification H3K9ac in the breast myoepithelial primary cells (Table 3).

Due to our knowledge, this is the first association study between breast cancer susceptibility and polymorphisms of *MYH7B*, *MIR499A*, *TRPC4AP*, *ESR1* and *RP11-277P12.20* genes. These genes were selected using LDlink for the Iranian population. RegulomeDB is a powerful tool for the prediction of the regulatory potential of various variants. It is expected that the RegulomeDB web-based tool will be widely applied in the future for performing extensive association studies.

Conclusion

Considering the results of comparisons made in the present study which confirmed epigenetic properties for non-coding SNPs, the importance of these segments in the functional epigenetic studies were highlighted. Non-coding SNPs have a great impact on the binding capacity of regulatory proteins and gene expression pattern modifications as they can lead to histone modifications (Khurana et al., 2016). In order to evaluate the possible functional properties of shortlisted SNPs in the Iranian population, *in silico* analyses using LDlink, RegulomeDB and HaploReg are strongly recommended. It could be expected that our computational model could prioritize variants in the regulatory regions. Thus, it helps researchers to figure out functional variants of noncoding regions with key effects in the pathogenesis of various diseases.

Acknowledgments

The present article is part of the master's thesis of Biology and supported by the University of Science and Art. We also appreciate the efforts of Yazd Agricultural and Natural Resources Research Center members.

Conflict of interest

The authors declared no conflicts of interest.

www.SID.ir

Archive of SID

References

Bauer-Mehren, A., Furlong, L. I., Rautschka, M. and Sanz, F. (2009) From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. BMC bioinformatics 10: S6.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M. and Cherry, J. M. (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome research 22(9): 1790-1797.

Chen, L., Kang, H., Jin, G. J., Chen, X., Zhang, Q. Y., Lao, W. T. and Li, R. (2016) The association between a novel polymorphism (rs1062577) in ESR1 and breast cancer susceptibility in the Han Chinese women. Gynecological Endocrinology 32(7): 553-556.

Coetzee, S. G., Rhie, S. K., Berman, B. P., Coetzee, G. A. and Noushmehr, H. (2012) FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. Nucleic acids research 40(18): e139-e139.

Dehghan, Z., Sadeghi, S., Tabatabaeian, H., Ghaedi, K., Azadeh, M., Fazilati, M. and Bagheri, F. (2017) ESR1 single nucleotide polymorphism rs1062577 (c.* 3804T> A) alters the susceptibility of breast cancer risk in Iranian population. Gene 611: 9-14.

Edwards, S. L., Beesley, J., French, J. D. and Dunning, A. M. (2013) Beyond GWASs: illuminating the dark road from association to function. The American Journal of Human Genetics 93(5): 779-797.

Ernst, J. and Kellis, M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology 28(8): 817-825.

Fayez, A. (2018) Using postgenome-wide association study analysis; Vars2-Pic3ca-AKT is novel putative interactive pathway associated with conotruncal heart defects. Biomedical and Biotechnology Research Journal 2(4): 269.

Ghobadzadeh, S., Shams, A., Eslami, G. and Mirghanizadeh, A. (2013) Investigation of NKG2D rs1049174G> C Gene Polymorphism in Women with Breast Cancer. SSU_Journals 21(3): 291-299.

Hamdi, Y., Rekaya, M. B., Jingxuan, S., Nagara, M., Messaoud, O., Elgaaied, A. B. and Boussen, H. (2018) A genome wide SNP genotyping study in the Tunisian population: specific reporting on a subset of common breast cancer risk loci. BMC cancer 18(1): 1-14.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F. and Ching, K. A. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459(7243): 108-112.

Jiang, S. G., Chen, L., Tang, J. H., Zhao, J. H. and Zhong, S. L. (2015) Lack of association between Hsa-Mir-499 rs3746444 polymorphism and cancer risk: meta-analysis findings. Asian Pacific Journal of Cancer Prevention 16(1):339-344.

Kabirizadeh, S., Azadeh, M., Mirhosseini, M., Ghaedi, K. and Tanha, H. M. (2016) The SNP rs3746444 within mir-499a is associated with breast cancer risk in Iranian population. Journal of Cellular Immunotherapy 2(2): 95-97.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A. and Gerstein, M. (2016) Role of non-coding sequence variants in cancer. Nature Reviews Genetics 17(2): 93.

Machiela, M. J. and Chanock, S. J. (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31(21): 3555-3557.

Meng, F., Yuan, G., Zhu, X., Zhou, Y., Wang, D. and Guo, Y. (2018) Functional variants identified efficiently through an integrated transcriptome and epigenome analysis. Scientific reports 8(1): 1-13.

Mu, K., Wu, Z. Z., Yu, J. P., Guo, W., Wu, N., Wei, L. J. and Liu, J. T. (2017) Meta-analysis of the association between three microRNA polymorphisms and breast cancer susceptibility. Oncotarget 8(40): 68809.

Ong, C. T. and Corces, V. G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics 12(4): 283-293.

Rhie, S. K., Coetzee, S. G., Noushmehr, H., Yan, C., Kim, J. M., Haiman, C. A. and Coetzee, G. A. (2013) Comprehensive functional annotation of seventy-one breast cancer risk Loci. PloS one 8(5): e63925.

Seyedmir, F., Mirzaie, K. and Bitaraf Sani, M. (2017) The Studies of Decision Tree in Estimation of Breast Cancer Risk by Using Polymorphism Nucleotide. Journal of Shahid Sadoughi University of Medical Sciences 25(4): 300-310.

Wang, L., Qian, S., Zhi, H., Zhang, Y., Wang, B. and Lu, Z. (2012) The association between hsa-miR-499

www.SID.ir

Archive of SID

T> C polymorphism and cancer risk: a meta-analysis. Gene 508(1): 9-14.

Wang, Y., Yang, B. and Ren, X. (2012) Hsa-miR-499 polymorphism (rs3746444) and cancer risk: a meta-analysis of 17 case–control studies. Gene 509(2): 267-272.

Ward, L. D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research 40(D1): D930-D934.

Zou, P., Zhao, L., Xu, H., Chen, P., Gu, A., Liu, N. and Lu, A. (2012) Hsa-mir-499 rs3746444 polymorphism and cancer risk: a meta-analysis. Journal of biomedical research 26(4): 253-259.

Open Access Statement:
This is an open access article distributed under the Creative Commons Attribution License (CC-BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

www.SID.ir