Functional nsSNPs from carcinogenesis-related genes expressed in breast tissue: Potential breast cancer risk alleles and their distribution across human populations

*Sevtap Savas,[1,2,3] Steffen Schmidt,[4] Hamdi Jarjanazi[1,2,3] and Hilmi Ozcelik[1,2,3]**

[1] Fred A. Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, ON, M5G 1X5, Canada
[2] Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 600 University Avenue, Toronto, ON, M5G IX5, Canada
[3] Department of Laboratory Medicine and Pathobiology, University of Toronto, 100 College Street, Toronto, ON, M5G IL5, Canada
[4] Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
* *Correspondence to*: Tel: +1 416 586 4996; Fax: +1 416 586 8869; E-mail: ozcelik@mshri.on.ca

Date received (in revised form): 9th December 2005

Abstract

Although highly penetrant alleles of *BRCA1* and *BRCA2* have been shown to predispose to breast cancer, the majority of breast cancer cases are assumed to result from the presence of low–moderate penetrant alleles and environmental carcinogens. Non-synonymous single nucleotide polymorphisms (nsSNPs) are hypothesised to contribute to disease susceptibility and approximately 30 per cent of them are predicted to have a biological significance. In this study, we have applied a bioinformatics-based strategy to identify breast cancer-related nsSNPs from 981 carcinogenesis-related genes expressed in breast tissue. Our results revealed a total of 367 validated nsSNPs, 109 (29.7 per cent) of which are predicted to affect the protein function (functional nsSNPs), suggesting that these nsSNPs are likely to influence the development and homeostasis of breast tissue and hence contribute to breast cancer susceptibility. Sixty-seven of the functional nsSNPs presented as commonly occurring nsSNPs (minor allele frequencies $\geq$5 per cent), representing excellent candidates for breast cancer susceptibility. Additionally, a non-uniform distribution of the common functional nsSNPs among different human populations was observed: 15 nsSNPs were reported to be present in all populations analysed, whereas another set of 15 nsSNPs was specific to particular population(s). We propose that the nsSNPs analysed in this study constitute a unique resource of potential genetic factors for breast cancer susceptibility. Furthermore, the variations in functional nsSNP allele frequencies across major population backgrounds may point to the potential variability of the molecular basis of breast cancer predisposition and treatment response among different human populations.

Keywords: *breast cancer predisposition, nsSNPs, breast tissue expression, carcinogenesis-related genes, PolyPhen*

Introduction

Mutations of *BRCA1*[1] and *BRCA2*[2] confer high breast cancer risk to the carriers. Such highly penetrant mutations are only responsible for a small fraction (~5–10 per cent) of all breast cancer cases,[3,4] however, suggesting the presence of other, yet to be identified, mutations in other breast cancer predisposition genes.[5–7] Mutations in a number of genes, such as *p53*,[8] *ATM*[6] and *Chek2*,[9] have also been shown to contribute to breast cancer risk in a very small fraction of breast cancer cases. So far, no other high-penetrant breast cancer susceptibility gene has been identified; however, genetic variations including single nucleotide polymorphisms (SNPs) have been hypothesised to act as low–moderate penetrant alleles and contribute to breast cancer, as well as other complex diseases.[7,10–12]

Variations in protein sequence and function are mainly due to the non-synonymous form of SNPs (nsSNPs). The fraction of nsSNPs in the genome is relatively low (~10 per cent of all coding SNPs)[13] compared with other types, but they are

more likely to alter the structure, function and interaction of the proteins, and thus constitute a set of candidate genetic factors associated with disease predisposition.[14,15] Approximately 30 per cent of the nsSNPs are predicted to have biological consequences.[16–18] Several nsSNPs from the proteins acting in a variety of cellular pathways—such as apoptosis,[19] oxidative stress[20] and signal transduction[21]—have already been reported to be associated with an increased/decreased risk of breast cancer.

Several studies have described cancer-relevant nsSNPs;[22–25] however, to our knowledge they have not been studied in the context of expression of genes in a particular tissue. Clearly, in order for genes to be linked to a disease of a tissue, their protein products should somehow influence that particular tissue, either as exogenous proteins (such as hormones) or endogenous proteins (such as the proteins expressed in that tissue).[26,27] In this study, we have applied a bioinformatics-based strategy and identified potentially functional nsSNPs from endogenous carcinogenesis-related proteins expressed in breast tissue.

Methods

Genes

The Ensembl transcript identifiers (http://www.ensembl.org/)[28] of the genes expressed in breast tissue were retrieved from the TissueInfo database (db) (http://icb.med.cornell.edu/services/tissueinfo/query).[29] The list of carcinogenesis-related genes from 18 different categories ('DNA adduct', 'DNA damage', 'DNA replication', 'angiogenesis', 'apoptosis', 'behavior', 'cell cycle', 'cell signaling', 'development', 'gene regulation', 'transcription', 'immunology', 'metabolism', 'metastasis', 'pharmacology', 'signal transduction', 'tumor suppressors/oncogenes' and 'miscellaneous') was retrieved from the National Cancer Institute's Cancer Genome Anatomy Project Genetic Annotation Initiative ([CGAP-GAI] website [http://lpgws.nci.nih.gov/html-cgap/cgl/]).[30] The genes retrieved from the TissueInfo and the CGAP-GAI resources were then cross-referenced with each other to identify the group of carcinogenesis-related genes that are expressed in breast tissue.

nsSNPs

The nsSNPs from the group of carcinogenesis-related genes expressed in breast tissue were retrieved from dbSNP build 120 (http://www.ncbi.nlm.nih.gov/SNP/).[31] Only the nsSNPs detected in ≥ 2 chromosomes in a sample panel of ≥ 40 chromosomes were included in this study (validated nsSNPs). Seventeen nsSNPs were found in both less and more than 5 per cent of the chromosomes analysed in different sample sets; for simplicity, we have classified such nsSNPs within the nsSNP set with ≥ 5 per cent minor allele frequencies throughout this paper.

PolyPhen analysis

The PolyPhen predictions[18] were retrieved from a pre-computed dbSNP–PolyPhen resource. All PolyPhen predictions were based on either alignment of at least five similar proteins (for a more reliable prediction) or structural parameters.

Results

The results obtained in this study are summarised in Table 1 and constitute only the validated nsSNPs with a reliable prediction made by the PolyPhen prediction tool (see Methods). A total of 367 nsSNPs from 189 carcinogenesis-related genes expressed in breast tissue are presented. A total of 109 nsSNPs (28.4 per cent) from 75 genes were predicted potentially to affect the protein function (functional nsSNPs). Additionally, 61.5 per cent ($n = 67$) of the potentially functional nsSNPs represented commonly occurring nsSNPs in the population (≥ 5 per cent minor allele frequency; Table 2). In this paper, we mainly discuss the commonly occurring functional nsSNPs; however, the list of rarely occurring functional nsSNPs can also be found under the supplementary table (www.ozceliklab.com/Breast_rare_nsSNPs/).

A fraction of protein products of genes bearing commonly occurring functional nsSNPs were found to be involved in one or more carcinogenesis-related biological pathways compiled by the CGAP-GAI[30] (Table 2). Such nsSNPs were mostly found in the proteins from DNA repair (three genes, four nsSNPs); metastasis (four genes, four nsSNPs);

Table 1. Summary of the results.

	n
Genes	
Carcinogenesis-related genes	2,832
Expressed in breast tissue	981
With validated nsSNPs	189
With functional nsSNPs	75
nsSNPs	
Validated nsSNPs	367
Benign by PolyPhen	258
Functional by PolyPhen	109
With $\geq 5\%$ minor allele frequency	67
With $< 5\%$ minor allele frequency	42

Abbreviation: *n* = number; nsSNP = non-synonymous form of single nucleotide polymorphisms. Please note that only the genes and the nsSNPs for which a reliable PolyPhen prediction (based on ≥ 5 proteins in the alignment) was available are shown in this table.

Table 2. Functional and common non-synonymous form of single nucleotide polymorphisms (nsSNPs) from the breast tissue-expressed carcinogenesis-related genes.

Gene[a]	Accession number	SNP ID[b]	Amino acid change[c]	Codons[d]	Damaging allele	Damaging amino acid[e]	PolyPhen prediction	Pathway[f]
ACY1	NM_000666.1	rs2229152	R386C	cgt/tgt	t	C	Probably damaging	IM
ADD1	NM_014189.2	rs4961	G460W	ggg/tgg	t	W	Probably damaging	IM
ADD1	NM_014189.2	rs4962	N541I	aat/att	t	I	Probably damaging	IM
ADD1	NM_014189.2	rs4971	Y270N	tat/aat	a	N	Probably damaging	IM
ADM	NM_001124.1	rs5005	S50R	agc/agg	g	R	Possibly damaging	AN
ADRB2	NM_000024.3	rs1042713	G16R	gga/aga	a	R	Possibly damaging	BE, IM
ALDH2	NM_000690.2	rs671	E504K	gaa/aaa	a	K	Possibly damaging	IM, PH
APOE	NM_000041.1	rs429358	C130R	tgc/cgc	c	R	Probably damaging	IM
AXIN2	NM_004655.1	rs2240308	P50S	cct/tct	t	S	Probably damaging	DE
C2	NM_000063.3	rs4151648	R734C	cgc/tgc	t	C	Possibly damaging	IM
CD2	NM_001767.2	rs699738	H266Q	cac/caa	a	Q	Probably damaging	AN, IM, MET
CDH12	NM_004061.2	rs4371716	V68M	gtg/atg	g	V	Probably damaging	IM
CHGA	NM_001275.2	rs729940	R399W	cgg/tgg	t	W	Probably damaging	IM
CHGA	NM_001275.2	rs9658667	G382S	ggc/agc	a	S	Possibly damaging	IM
CLU	NM_001831.1	rs9331936	N317H	aac/cac	c	H	Possibly damaging	IM
CSF1	NM_000757.3	rs2229165	G438R	ggg/agg	a	R	Probably damaging	IM
CSF3R	NM_000760.2	rs3917973	M231T	atg/acg	c	T	Probably damaging	IM
CSF3R	NM_000760.2	rs3917974	Q346R	cag/cgg	g	R	Possibly damaging	IM
CSF3R	NM_000760.2	rs3917991	D510H	gac/cac	c	H	Possibly damaging	IM
CYBA	NM_000101.1	rs4673	Y72H	tac/cac	c	H	Possibly damaging	IM
CYP11B1	NM_000497.2	rs4541	A386V	gcg/gtg	c	A	Possibly damaging	PH
CYP11B1	NM_000497.2	rs5287	M160I	atg/atc	c	I	Possibly damaging	PH
CYP11B1	NM_000497.2	rs5294	Y439H	tac/cac	t	Y	Probably damaging	PH
CYP11B1	NM_000497.2	rs5312	E383V	gag/gtg	t	V	Probably damaging	PH
CYP1B1	NM_000104.2	rs1800440	N453S	aac/agc	g	S	Possibly damaging	IM, PH

(*continued*)

Table 2. *Continued.*

Gene[a]	Accession number	SNP ID[b]	Amino acid change[c]	Codons[d]	Damaging allele	Damaging amino acid[e]	PolyPhen prediction	Pathway[f]
CYP2A6	NM_000762.4	rs1801272	L160H	ctc/cac	a	H	Probably damaging	IM, PH
CYP2B6	NM_000767.3	rs2279343	K262R	aag/agg	a	K	Possibly damaging	PH
CYP2C9	NM_000771.2	rs1799853	R144C	cgt/tgt	t	C	Probably damaging	IM, PH
DAG1	NM_004393.1	rs2131107	S14W	tcg/tgg	c	S	Probably damaging	IM
ENG	NM_000118.1	rs1800956	D366H	gac/cac	c	H	Possibly damaging	AN, DE, IM, MET
EPHX1	NM_000120.2	rs1051740	Y113H	tac/cac	c	H	Possibly damaging	IM, ME, PH
ERBB2	NM_004448.1	rs1058808	P1170A	ccc/gcc	g	A	Possibly damaging	IM, ST, TS/ON
F2R	NM_001992.2	rs2230849	Y187N	tac/aac	a	N	Probably damaging	IM
FPR1	NM_002029.3	rs867228	E346A	gag/gcg	c	A	Possibly damaging	IM
FUCA2	NM_032020.3	rs3762001	H371Y	cat/tat	t	Y	Possibly damaging	IM
GAA	NM_000152.2	rs1800307	G576S	ggc/agc	a	S	Possibly damaging	IM
GBP1	NM_002053.1	rs1048425	T349S	acc/agc	g	S	Possibly damaging	CS
GYS1	NM_002103.3	rs5453	P691A	cca/gca	g	A	Probably damaging	IM
GYS1	NM_002103.3	rs5456	K130E	aag/gag	g	E	Possibly damaging	IM
GYS1	NM_002103.3	rs5461	N283S	aat/agt	g	S	Possibly damaging	IM
HK2	NM_000189.4	rs2229629	R844K	agg/aag	g	R	Possibly damaging	IM, MIS
LIG4	NM_002312.2	rs1805388	T9I	act/att	t	I	Possibly damaging	DA, DD
MC1R	NM_002386.2	rs1805005	V60L	gtg/ttg	t	L	Possibly damaging	IM
MC1R	NM_002386.2	rs1805007	R151C	cgc/tgc	t	C	Probably damaging	IM
MC1R	NM_002386.2	rs3212366	F196L	ttc/ctc	c	L	Probably damaging	IM
MMP9	NM_004994.1	rs2250889	R574P	cgg/ccg	g	R	Possibly damaging	AN, IM
MMP9	NM_004994.1	rs3918252	N127K	aac/aag	g	K	Probably damaging	AN, IM
MNDA	NM_002432.1	rs2276403	H357Y	cac/tac	t	Y	Possibly damaging	GR, TR
MUC4	NM_004532.2	rs2259292	G88D	ggc/gac	g	G	Possibly damaging	IM
NFATC1	NM_006162.3	rs754093	C751G	tgt/ggt	g	G	Probably damaging	IM

NOTCH4	NM_004557.2	rs2071282	P203L	cc**c**/ct**c**	t	L	Probably damaging	IM, TS/ON
PGM3	NM_015599.1	rs473267	D466N	g**a**t/**a**at	a	N	Possibly damaging	IM
PLAU	NM_002658.1	rs2227564	L141P	ct**g**/cc**g**	t	L	Possibly damaging	AN
PLAUR	NM_002659.1	rs4760	L317P	ct**c**/cc**c**	c	P	Possibly damaging	AN
PTGS2	NM_000963.1	rs5272	E488G	g**a**g/g**g**g	g	G	Probably damaging	IM, MIS
PTPN3	NM_002829.2	rs3793524	A90P	g**c**c/c**c**c	g	A	Probably damaging	CC, CS
SLC1A5	NM_005628.1	rs3027956	P17A	c**c**c/g**c**c	g	A	Possibly damaging	IM
STAT2	NM_005419.2	rs2066816	Q66H	ca**g**/ca**t**	t	H	Possibly damaging	IM, ST
TBXAS1	NM_001061.2	rs5760	G390V	g**g**c/g**t**c	t	V	Probably damaging	IM
TBXAS1	NM_001061.2	rs5762	R425C	c**g**c/t**g**c	t	C	Probably damaging	IM
TBXAS1	NM_001061.2	rs5770	R261G	a**g**g/g**g**g	g	G	Probably damaging	IM
TDG	NM_003211.2	rs4135113	G199S	g**g**c/a**g**c	a	S	Possibly damaging	DD
TUBA1	NM_006000.1	rs3731891	R243C	c**g**c/t**g**c	t	C	Probably damaging	CS, MET
TYR	NM_000372.2	rs1042602	S192Y	t**c**t/t**a**t	a	Y	Possibly damaging	ME
VCAM1	NM_001078.2	rs3783613	G413A	g**g**t/g**c**t	c	A	Possibly damaging	AN, CS, IM, MET
XRCC1	NM_006297.1	rs25489	R280H	c**g**t/c**a**t	a	H	Possibly damaging	DD, DR, IM
XRCC1	NM_006297.1	rs1799782	R194W	c**g**g/t**g**g	t	W	Probably damaging	DD, DR, IM

Abbreviations: AN = angiogenesis; BE = behaviour, CC = cell cycle; CS = cell signalling; DA = DNA adduct; DD = DNA damage; DE = development; GR = gene regulation; IM = immunology; ME = metabolism; MET = metastasis; MIS = miscellaneous; PH = pharmacology; ST = signal transduction; TS/ON = tumour suppressor/oncogene; TR = transcription.
All nsSNPs are with ≥5 per cent minor allele frequency.
[a] The gene symbols are as approved by the HUGO Gene Nomenclature Committee.[67]
[b] SNP identifiers (IDs) correspond to the dbSNP IDs (http://www.ncbi.nlm.nih.gov/SNP/).[31]
[c] The position of the amino acid substitution and the amino acids specified by the major and minor SNP alleles are indicated.
[d] The codons specified by the major and the minor SNP alleles are shown. The nucleotide change is underlined.
[e] One-letter codes for the amino acids that are predicted to affect the protein function by PolyPhen.
[f] The pathway(s) that the proteins are implicated in are as shown by the Cancer Genome Anatomy Project Genetic Annotation Initiative website (http://lpgws.nci.nih.gov/html-cgap/cgl/).[30]

angiogenesis (seven genes, eight nsSNPs); pharmacology (seven genes, ten nsSNPs); and immunology (38 genes, 51 nsSNPs).

We have also analysed the distribution of the commonly occurring functional nsSNPs across human populations. For simplicity, we have categorised the frequency information obtained from different dbSNP entries into three major groups: African (African and African-American), Caucasian (Caucasian and European) and Asian (Chinese and East Asian) populations. Minor allele frequencies for nsSNPs were available for at least three different human populations for 30 out of 67 commonly occurring functional nsSNPs (Table 3). Fifteen nsSNPs were found in all populations analysed ($n \geq 3$). In the case of the remaining 15 nsSNPs, five were found exclusively in one population (ADM-S50R and MMP9-N127K in African; ALDH2-E504K and MNDA-H357Y in Asian; MC1R-R151C in Caucasian). Additionally, three nsSNPs were found in Caucasian, Asian or Hispanic samples, but not in the African samples (CHGA-G382S, CYP1B1-N453S and CYP2C9-R144C). Moreover, in the case of five nsSNPs, the major and the minor alleles were different among the populations analysed (ADBR2-G16R, CDH12-V68M, ERBB2-P1170A, PGM3-D466N and SLC1A5-P17A).

Table 3. Functional and common non-synonymous form of single nucleotide polymorphisms (nsSNPs) with frequency information available from different human populations.

Gene[a]	SNP ID[b]	Amino acid change[c]	African	Asian	Caucasian	Hispanic
ADD1	rs4961	G460W	46 chr. G = 0.891 T = 0.109	48 chr. G = 0.521 T = 0.479	48 chr. G = 0.833 T = 0.167	n/a
ADM	rs5005	S50R	46 chr. C = 0.957 G = 0.043	48 chr. C = 1.000	48 chr. C = 1.000	n/a
ADRB2	rs1042713	G16R	46 chr. G = 0.609 A = 0.391	48 chr. A = 0.583 G = 0.417	46 chr. G = 0.674 A = 0.326	n/a
ALDH2	rs671	E504K	48 chr. G = 1.000	48 0 G = 0.771 A = 0.229	58 chr. G = 1.000	44 chr. G = 1.000
CDH12	rs4371716	V68M	46 chr. T = 0.674 C = 0.326	48 chr. C = 0.812 T = 0.188	48 chr. C = 0.729 T = 0.271	n/a
CHGA	rs729940	R399W	114 chr. C = 0.954 T = 0.046	88 chr. C = 0.715 T = 0.285	104 chr. C = 0.893 T = 0.107	56 chr. C = 0.769 T = 0.231
CHGA	rs9658667	G382S	114 chr. G = 1.000	88 chr. G = 0.982 A = 0.018	104 chr. G = 0.951 A = 0.049	56 chr. G = 0.941 A = 0.059
CSF3R	rs3917973	M231T	48 chr. T = 0.938 C = 0.062	48 chr. T = 1.000	58 chr. T = 0.983 C = 0.017	46 chr. T = 1.000
CSF3R	rs3917991	D510H	48 chr. G = 0.750 C = 0.250	48 chr. G = 1.000	58 chr. G = 1.000	46 chr. G = 0.935 C = 0.065
CYBA	rs4673	Y72H	48 chr. C = 0.542 T = 0.458	1480 chr. G = 0.907 A = 0.093	60 chr. C = 0.683 T = 0.317	46 chr. C = 0.783 T = 0.217
CYP1B1	rs1800440	N453S	48 chr. A = 1.000	48 chr. A = 0.958 G = 0.042	62 chr. A = 0.806 G = 0.194	46 chr. A = 0.761 G = 0.239
CYP2A6	rs1801272	L160H	46 chr. T = 1.000	46 chr. T = 1.000	60 chr. T = 0.900 A = 0.100	46 chr. T = 0.978 A = 0.022
CYP2C9	rs1799853	R144C	48 chr. C = 1.000	48 chr. C = 0.979 T = 0.021	62 chr. C = 0.871 T = 0.129	46 chr. C = 0.935 T = 0.065

(*continued*)

Table 3. *Continued.*

Gene[a]	SNP ID[b]	Amino acid change[c]	African	Asian	Caucasian	Hispanic
ENG	rs1800956	D366H	46 chr. C = 0.978 G = 0.022	1480 chr. C = 0.942 G = 0.058	46 chr. C = 1.000	n/a
EPHX1	rs1051740	Y113H	48 chr. T = 0.917 C = 0.083	84 chr. T = 0.620 C = 0.380	62 chr. T = 0.613 C = 0.387	46 chr. T = 0.587 C = 0.413
ERBB2	rs1058808	P1170A	40 chr. C = 0.775 G = 0.225	1502 chr. G = 0.514 C = 0.486	48 chr. G = 0.646 C = 0.354	n/a
FPR1	rs867228	E346A	44 chr. G = 0.818 T = 0.182	46 chr. G = 0.761 T = 0.239	48 chr. G = 0.771 T = 0.229	n/a
FUCA2	rs3762001	H371Y	44 chr. G = 0.818 A = 0.182	1282 chr. G = 0.789 A = 0.211	44 chr. G = 0.795 A = 0.205	n/a
LIG4	rs1805388	T9I	48 chr. C = 0.979 T = 0.021	48 chr. G = 0.792 A = 0.208	62 chr. C = 0.871 T = 0.129	46 chr. C = 0.848 T = 0.152
MC1R	rs1805007	R151C	42 chr. C = 1.000	40 chr. C = 1.000	46 chr. C = 0.891 T = 0.109	n/a
MMP9	rs2250889	R574P	46 chr. C = 0.870 G = 0.130	1488 chr. C = 0.688 G = 0.312	48 chr. C = 0.896 G = 0.104	n/a
MMP9	rs3918252	N127K	48 chr. C = 0.938 G = 0.062	48 chr. C = 1.000	48 chr. C = 1.000	n/a
MNDA	rs2276403	H357Y	46 chr. C = 1.000	1484 chr. C = 0.944 T = 0.056	48 chr. C = 1.000	n/a
PGM3	rs473267	D466N	46 chr. T = 0.565 C = 0.435	84 chr. C = 0.750 T = 0.250	48 chr. C = 0.688 T = 0.312	n/a
PLAU	rs2227564	L141P	48 chr. C = 0.979 T = 0.021	1492 chr. G = 0.783 A = 0.217	44 chr. C = 0.659 T = 0.341	n/a
PTPN3	rs3793524	A90P	46 chr. G = 0.522 C = 0.478	1498 chr. G = 0.628 C = 0.372	46 chr. C = 0.717 G = 0.283	n/a
SLC1A5	rs3027956	P17A	46 chr. G = 0.957 C = 0.043	42 chr. G = 0.524 C = 0.476	146 chr. C = 0.710 G = 0.290	n/a
TYR	rs1042602	S192Y	46 chr. C = 0.957 A = 0.043	48 chr. C = 1.000	48 chr. C = 0.750 A = 0.250	n/a
VCAM1	rs3783613	G413A	48 chr. G = 0.938 C = 0.062	44 chr. G = 0.977 C = 0.023	48 chr. G = 1.000	n/a
XRCC1	rs25489	R280H	48 chr. G = 0.937 A = 0.063	84 chr. C = 1.000	62 chr. G = 0.968 A = 0.032	46 chr. G = 0.957 A = 0.043

Abbreviations: chr: chromosomes; n/a: not available.

[a] The gene symbols are as approved by the HUGO Gene Nomenclature Committee.[67]

[b] SNP identifiers (IDs) correspond to the dbSNP IDs (http://www.ncbi.nlm.nih.gov/SNP/).[31]

[c] The position of the amino acid substitution and the amino acids specified by the major and minor SNP alleles are indicated. The frequency information is as in dbSNP build 123 and is based on ≥40 chromosomes. Please note that the samples annotated as African and African-American; Caucasian and European; Chinese and East Asian are combined together here and are referred to as African, Caucasian and Asian, respectively. Whenever more than one entry was available for a group, only the information from the entries with the highest number of chromosomes is included here.

Discussion

A portion of SNPs is considered to contribute to complex disease development.[7,10–12] SNPs in or around the candidate genes might be directly linked to a disease; however, not all SNPs are supposed to affect gene expression and function, so selection of those with potential effects is keenly debated.[32] Several studies have developed tools and/or systematically analysed nsSNPs to identify those that affect gene function based on evolutionary conservation or structural parameters.[16–18,33] PolyPhen[18] is one such web-based tool utilised to select the nsSNPs that are likely to affect protein function. In short, the PolyPhen predictions are based on protein alignments, structural parameters or sequence annotations. The sensitivity of PolyPhen has been reported to be approximately 82 per cent.[18]

In this study, we hypothesised that the systematic analysis of candidate genes that are expressed in the affected tissue is likely to improve and enrich the identification of disease-susceptibility alleles. Accordingly, using a bioinformatics-based strategy, we identified the functional nsSNPs from a large number of genes related to the carcinogenesis-related pathways (DNA repair, cell cycle, signal transduction, etc), which are expressed in breast tissue. We propose that these potentially functional nsSNPs can result in abnormalities at the protein level, which are likely to affect the development, metabolism and homeostasis of the breast tissue, and thus can contribute to breast cancer susceptibility.

The genes with functional nsSNPs identified in this study were from a variety of carcinogenesis-related cellular pathways. According to this information, possible biological roles for these nsSNPs may be suggested. For example, nsSNPs from angiogenesis- and metastasis-related proteins may have roles in tumour growth and the development of metastatic tumours.[34,35] Additionally, DNA repair nsSNPs may lead to the accumulation of somatic mutations and thus can participate in cancer initiation and promotion.[34–36] Furthermore, together with the DNA repair nsSNPs, the nsSNPs from the pharmacology genes may also be good candidates for the studies targeting the efficacy, differential response and adverse effect of chemo-/radiotherapy in breast cancer.[37–39] The majority of the nsSNPs were from the genes related to immunological responses (74.6 per cent), which can both suppress and promote tumorigenesis.[34] It is likely that the larger number of the functional nsSNPs in immune system-related genes is a reflection of the large number of immunology genes in the breast tissue-expressed gene set (60 per cent).

A considerable number of genes with functional nsSNPs have been previously linked to breast cancer aetiology: *ADM*,[40] *ADRB2*,[41] *APOE*,[42] *CHGA*,[43] *CSF1*,[44] *CYP1B1*,[45] *DAG1*,[46] *ENG*,[47] *EPHX1*,[48] *ERBB2*,[49] *F2R*,[50] *MMP9*,[51] *MUC4*,[52] *NFATC1*,[53] *NOTCH4*,[54] *PLAU*,[55] *PLAUR*,[55] *PTGS2*[56] and *VCAM1*.[57] Therefore, we propose that the nsSNPs in Table 2 are excellent candidates as genetic factors involved in breast cancer initiation, promotion or progression. Additionally, some of these nsSNPs may be critical for breast cancer treatment outcome.

When the distribution of the commonly occurring functional nsSNPs was analysed, differences in the major alleles and the allele frequencies across human populations were observed. For example, 15 commonly occurring nsSNPs were found in all populations, whereas another set of 15 nsSNPs was specific to particular population(s). These differences might be reflections of either the age of the allele, founder effects or the dissimilar selective pressures acting on different populations.[58,59] Most importantly, the data also indicate that a common nsSNP with a potential biological consequence in our set was equally likely to be either prevalent across different human populations or limited to some populations. Clearly, the latter prompted us to conclude that the population-specific functional nsSNPs may contribute to the genetic predisposition in individuals with a specific background. In this regard, this conclusion is consistent with previous studies in which genetic variations with significantly different allelic frequencies among populations were found to be associated with specific disease or differential drug responses.[60–65] This information may be particularly helpful to researchers in determining which nsSNPs may be relevant to utilise in specific population-based studies. In addition, although further analyses are required, it is tempting to speculate that these nsSNPs may be a part of the potential variability of the molecular basis of breast cancer predisposition and drug response among different human populations.

Data integration from several databases forms the basis of our strategy to determine functional SNPs of breast tissue-expressed genes. The quality and the quantity of the genomic data within individual databases influence the comprehensiveness of the combined data. The functional SNP list presented in this study is a result of data integration from three databases — namely, TissueInfo,[29] Ensembl,[28] and dbSNP.[31] The non-matching data fields (eg transcript identifiers) between TissueInfo, Ensembl and dbSNP have been the main source of missing data. For example, although *BRCA1* was known to have a potentially functional SNP (predicted previously), this information has not been captured because of non-matching transcript identifier information for *BRCA1* in the databases. Thus, incompatibility of data in different databases has been a rate-limiting factor for the bioinformatics-based strategies presented here. The improvement of the quality and the quantity of genomic data in the databases will prove beneficial for researching complex questions. Also, the genes presented in this paper are based on the expressed sequence tag information, which may lead to an underrepresentation of rarely expressed genes.[29,66] Data integration using other tissue expression databases is likely to enrich the quality of the data produced. Nevertheless, although it is possible that the SNPs presented here may not represent

the most comprehensive list, the SNPs identified using the proposed strategy represent a valuable resource for studying the genetic predisposition to breast cancer.

Conclusion

In conclusion, we have designed a novel strategy to identify potentially functional variants of cancer-related genes expressed in breast tissue. Our results demonstrated the presence of 109 nsSNPs with a potential biological consequence, 67 of which were frequent in human populations. We propose that, together with other genetic and environmental factors, these nsSNPs may be involved in breast cancer initiation and progression; thus, these nsSNPs represent the premium candidates as genetic variations of breast cancer predisposition. We also suggest that a considerable fraction of the nsSNPs may, in fact, be population-specific genetic variations.

Acknowledgments

The authors thank Baris Tuncertan and Mehjabeen Shariff for retrieving the data from the dbSNP and the pre-computed PolyPhen resource and Dr Michelle Cotterchio for critically reading the manuscript. This work was supported by grants (BCTR0100627) from the Susan Komen Breast Cancer Foundation, USA, and the Canadian Breast Cancer Foundation. Sevtap Savas is supported, in part, by a 'CIHR Strategic Training Program Grant — The Samuel Lunenfeld Research Institute Training Program: Applying Genomics to Human Health' fellowship.

References

1. Miki, Y., Swensen, J., Shattuck-Eidens, D. *et al.* (1994), 'A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*', *Science* Vol. 266, pp. 66–71.
2. Wooster, R., Bignell, G., Lancaster, J. *et al.* (1995), 'Identification of the breast cancer susceptibility gene *BRCA2*', *Nature* Vol. 378, pp. 789–792.
3. Hofmann, W. and Schlag, P.M. (2000), '*BRCA1* and *BRCA2* — Breast cancer susceptibility genes', *J. Cancer Res. Clin. Oncol.* Vol. 126, pp. 487–496.
4. Hodgson, S.V., Morrison, P.J. and Irving, M. (2004), 'Breast cancer genetics: Unsolved questions and open perspectives in an expanding clinical practice', *Am. J. Med. Genet. C Semin. Med. Genet.* Vol. 129, pp. 56–64.
5. Dong, C. and Hemminki, K. (2001), 'Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families', *Int. J. Cancer* Vol. 92, pp. 144–150.
6. Chenevix-Trench, G., Spurdle, A.B., Gatei, M. *et al.* (2002), 'Dominant negative ATM mutations in breast cancer families', *J. Natl. Cancer Inst.* Vol. 94, pp. 205–215.
7. Ponder, B.A. (2001), 'Cancer genetics', *Nature* Vol. 411, pp. 336–341.
8. Malkin, D., Li, F.P., Strong, L.C. *et al.* (1990), 'Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms', *Science* Vol. 250, pp. 1233–1238.
9. Meijers-Heijboer, H., van den Ouweland, A., Klijn, J. *et al.* (2002), 'Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of *BRCA1* or *BRCA2* mutations', *Nat. Genet.* Vol. 31, pp. 55–59.
10. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* Vol. 273, pp. 1516–1517.
11. Collins, A., Lonjou, C. and Morton, N.E. (1999), 'Genetic epidemiology of single-nucleotide polymorphisms', *Proc. Natl. Acad. Sci. USA* Vol. 96, pp. 15173–15177.
12. Houlston, R.S. and Peto, J. (2004), 'The search for low-penetrance cancer susceptibility alleles', *Oncogene* Vol. 23, pp. 6471–6476.
13. Reumers, J., Schymkowitz, J., Ferkinghoff-Borg, J. *et al.* (2005), 'SNPeffect: A database mapping molecular phenotypic effects of human non-synonymous coding SNPs', *Nucleic Acids Res.* Vol. 33(Database Issue), pp. D527–D532.
14. Chanock, S. (2001), 'Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease', *Dis. Markers* Vol. 17, pp. 89–98.
15. Pharoah, P.D., Dunning, A.M., Ponder, B.A. and Easton, D.F. (2004), 'Association studies for finding cancer-susceptibility genetic variants', *Nat. Rev. Cancer* Vol. 4, pp. 850–860.
16. Wang, Z. and Moult, J. (2001), 'SNPs, protein structure, and disease', *Hum. Mutat.* Vol. 17, pp. 263–270.
17. Ng, P.C. and Henikoff, S. (2002), 'Accounting for human polymorphisms predicted to affect protein function', *Genome Res.* Vol. 12, pp. 436–446.
18. Ramensky, V., Bork, P. and Sunyaev, S. (2002), 'Human non-synonymous SNPs: Server and survey', *Nucleic Acids Res.* Vol. 30, pp. 3894–3900.
19. MacPherson, G., Healey, C.S., Teare, M.D. *et al.* (2004), 'Association of a common variant of the *CASP8* gene with reduced risk of breast cancer', *J. Natl. Cancer Inst.* Vol. 96, pp. 1866–1869.
20. Menzel, H.J., Sarmanova, J., Soucek, P. *et al.* (2004), 'Association of NQO1 polymorphism with spontaneous breast cancer in two independent populations', *Br. J. Cancer* Vol. 90, pp. 1989–1994.
21. Rutter, J.L., Chatterjee, N., Wacholder, S. and Struewing, J. (2003), 'The HER2 I655V polymorphism and breast cancer risk in Ashkenazim', *Epidemiology* Vol. 14, pp. 694–700.
22. Livingston, R.J., von Niederhausern, A., Jegga, A.G. *et al.* (2004), 'Pattern of sequence variation across 213 environmental response genes', *Genome Res.* Vol. 14, pp. 1821–1831.
23. Savas, S., Kim, D.Y., Ahmad, M.F. *et al.* (2004), 'Identifying functional genetic variants in DNA repair pathway using protein conservation analysis', *Cancer Epidemiol. Biomarkers Prev.* Vol. 13, pp. 801–807.
24. Xi, T., Jones, I.M. and Mohrenweiser, H.W. (2004), 'Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function', *Genomics* Vol. 83, pp. 970–979.
25. Savas, S., Ahmad, M.F., Shariff, M. *et al.* (2005), 'Candidate nsSNPs that can affect the functions and interactions of cell cycle proteins', *Proteins* Vol. 58, pp. 697–705.
26. Ben-Shlomo, I., Vitt, U.A. and Hsueh, A.J. (2002), 'Perspective: The ovarian kaleidoscope database-II. Functional genomic analysis of an organ-specific database', *Endocrinology* Vol. 143, pp. 2041–2044.
27. Morton, C.C. (2004), 'Gene discovery in the auditory system using a tissue specific approach', *Am. J. Med. Genet. A* Vol. 130, pp. 26–28.
28. Hubbard, T., Andrews, D., Caccamo, M. *et al.* (2005), 'Ensembl 2005', *Nucleic Acids Res.* Vol. 33(Database Issue), pp. D447–D453.
29. Skrabanek, L. and Campagne, F. (2001), 'TissueInfo: High-throughput identification of tissue expression profiles and specificity', *Nucleic Acids Res.* Vol. 29, pp. E102–2.
30. Clifford, R., Edmonson, M., Hu, Y. *et al.* (2000), 'Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project', *Genome Res.* Vol. 10, pp. 1259–1265.
31. Sherry, S.T., Ward, M.H., Kholodov, M. *et al.* (2001), 'dbSNP: The NCBI database of genetic variation', *Nucleic Acids Res.* Vol. 29, pp. 308–311.
32. Daly, A.K. and Day, C.P. (2001), 'Candidate gene case-control association studies: Advantages and potential pitfalls', *Br. J. Clin. Pharmacol.* Vol. 52, pp. 489–499.
33. Sunyaev, S., Ramensky, V., Koch, I. *et al.* (2001), 'Prediction of deleterious human alleles', *Hum. Mol. Genet.* Vol. 10, pp. 591–597.
34. Jakobisiak, M., Lasek, W. and Golab, J. (2003), 'Natural mechanisms protecting against cancer', *Immunol. Lett.* Vol. 90, pp. 103–122.
35. Kirsch, M., Schackert, G. and Black, P.M. (2004), 'Metastasis and angiogenesis', *Cancer Treat. Res.* Vol. 117, pp. 285–304.

36. Mohrenweiser, H.W. (2004), 'Genetic variation and exposure related risk estimation: Will toxicology enter a new era? DNA repair and cancer as a paradigm', *Toxicol. Pathol.* Vol. 32, pp. 136–145.
37. Andreassen, C.N., Alsner, J., Overgaard, M. and Overgaard, J. (2003), 'Prediction of normal tissue radiosensitivity from polymorphisms in candidate genes', *Radiother. Oncol.* Vol. 69, pp. 127–135.
38. Watters, J.W. and McLeod, H.L. (2003), 'Cancer pharmacogenomics: Current and future applications', *Biochim. Biophys. Acta* Vol. 1603, pp. 99–111.
39. Sullivan, A., Syed, N., Gasco, M. *et al.* (2004), 'Polymorphism in wild-type p53 modulates response to chemotherapy *in vitro* and *in vivo*', *Oncogene* Vol. 23, pp. 3328–3337.
40. Oehler, M.K., Fischer, D.C., Orlowska-Volk, M. *et al.* (2003), 'Tissue and plasma expression of the angiogenic peptide adrenomedullin in breast cancer', *Br. J. Cancer* Vol. 89, pp. 1927–1933.
41. Cakir, Y., Plummer, H.K. 3rd, Tithof, P.K. and Schuller, H.M. (2002), 'Beta-adrenergic and arachidonic acid-mediated growth regulation of human breast cancer cell lines', *Int. J. Oncol.* Vol. 21, pp. 153–157.
42. Zunarelli, E., Nicoll, J.A., Migaldi, M. and Trentini, G.P. (2000), 'Apolipoprotein E polymorphism and breast carcinoma: Correlation with cell proliferation indices and clinical outcome', *Breast Cancer Res. Treat.* Vol. 63, pp. 193–198.
43. Pagani, A., Papotti, M., Hofler, H. *et al.* (1990), 'Chromogranin A and B gene expression in carcinomas of the breast. Correlation of immuno-cytochemical, immunoblot, and hybridization analyses', *Am. J. Pathol.* Vol. 136, pp. 319–327.
44. Lin, E.Y., Gouon-Evans, V., Nguyen, A.V. and Pollard, J.W. (2002), 'The macrophage growth factor CSF-1 in mammary gland development and tumor progression', *J. Mammary Gland Biol. Neoplasia* Vol. 7, pp. 147–162.
45. Spink, D.C., Spink, B.C., Cao, J.Q. *et al.* (1998), 'Differential expression of CYP1A1 and CYP1B1 in human breast epithelial cells and breast tumor cells', *Carcinogenesis* Vol. 19, pp. 291–298.
46. Sgambato, A., Migaldi, M., Montanari, M. *et al.* (2003), 'Dystroglycan expression is frequently reduced in human breast and colon cancers and is associated with tumor progression', *Am. J. Pathol.* Vol. 162, pp. 849–860.
47. Li, C., Guo, B., Bernabeu, C. and Kumar, S. (2001), 'Angiogenesis in breast cancer: The role of transforming growth factor beta and CD105', *Microsc. Res. Tech.* Vol. 52, pp. 437–449.
48. Fritz, P., Murdter, T.E., Eichelbaum, M. *et al.* (2001), 'Microsomal epoxide hydrolase expression as a predictor of tamoxifen response in primary breast cancer: A retrospective exploratory study with long-term follow-up', *J. Clin. Oncol.* Vol. 19, pp. 3–9.
49. Zhou, B.P. and Hung, M.C. (2003), 'Dysregulation of cellular signaling by HER2/neu in breast cancer', *Semin. Oncol.* Vol. 30, pp. 38–48.
50. Booden, M.A., Eckert, L.B., Der, C.J. and Trejo, J. (2004), 'Persistent signaling by dysregulated thrombin receptor trafficking promotes breast carcinoma cell invasion', *Mol. Cell. Biol.* Vol. 24, pp. 1990–1999.
51. Lee, P.P., Hwang, J.J., Murphy, G. and Ip, M.M. (2000), 'Functional significance of MMP-9 in tumor necrosis factor-induced proliferation and branching morphogenesis of mammary epithelial cells', *Endocrinology* Vol. 141, pp. 3764–3773.
52. Carraway, K.L., Price-Schiavi, S.A., Komatsu, M. *et al.* (2001), 'Muc4/sialomucin complex in the mammary gland and breast cancer', *J. Mammary Gland Biol. Neoplasia* Vol. 6, pp. 323–337.
53. Jauliac, S., Lopez-Rodriguez, C., Shaw, L.M. *et al.* (2002), 'The role of NFAT transcription factors in integrin-mediated carcinoma invasion', *Nat. Cell. Biol.* Vol. 4, pp. 540–544.
54. Politi, K., Feirt, N. and Kitajewski, J. (2004), 'Notch in mammary gland development and breast cancer', *Semin. Cancer Biol.* Vol. 14, pp. 341–347.
55. Sliva, D. (2004), 'Signaling pathways responsible for cancer cell invasion as targets for cancer therapy', *Curr. Cancer Drug Targets* Vol. 4, pp. 327–336.
56. Singh, B. and Lucci, A. (2002), 'Role of cyclooxygenase-2 in breast cancer', *J. Surg. Res.* Vol. 108, pp. 173–179.
57. O'Hanlon, D.M., Fitzsimons, H., Lynch, J. *et al.* (2002), 'Soluble adhesion molecules (E-selectin, ICAM-1 and VCAM-1) in breast carcinoma', *Eur. J. Cancer* Vol. 38, pp. 2252–2257.
58. Cavalli-Sforza, L.L. and Feldman, M.W. (2003), 'The application of molecular genetic approaches to the study of human evolution', *Nat. Genet.* Vol. 33, pp. 266–275.
59. Fay, J.C. and Wu, C.I. (2003), 'Sequence divergence, functional constraint, and selection in protein evolution', *Annu. Rev. Genomics Hum. Genet.* Vol. 4, pp. 213–235.
60. London, S.J., Lehman, T.A. and Taylor, J.A. (1997), 'Myeloperoxidase genetic polymorphism and lung cancer risk', *Cancer Res.* Vol. 57, pp. 5001–5003.
61. Evans, D.A., McLeod, H.L., Pritchard, S. *et al.* (2001), 'Interethnic variability in human drug responses', *Drug Metab. Dispos.* Vol. 29, pp. 606–610.
62. Gibson, A.W., Edberg, J.C., Wu, J. *et al.* (2001), 'Novel single nucleotide polymorphisms in the distal IL-10 promoter affect IL-10 production and enhance the risk of systemic lupus erythematosus', *J. Immunol.* Vol. 166, pp. 3915–3922.
63. Hopper, J.L. (2001), 'Genetic epidemiology of female breast cancer', *Semin. Cancer Biol.* Vol. 11, pp. 367–374.
64. Xu, C., Goodz, S., Sellers, E.M. and Tyndale, R.F. (2002), 'CYP2A6 genetic variation and potential consequences', *Adv. Drug Deliv. Rev.* Vol. 54, pp. 1245–1256.
65. Shimizu, E., Hashimoto, K. and Iyo, M. (2004), 'Ethnic difference of the BDNF 196G/A (val66met) polymorphism frequencies: The possibility to explain ethnic mental traits', *Am. J. Med. Genet. B Neuropsychiatr. Genet.* Vol. 126, pp. 122–123.
66. Wang, S.M. and Rowley, J.D. (1998), 'A strategy for genome-wide gene analysis: Integrated procedure for gene identification', *Proc. Natl. Acad. Sci. USA* Vol. 95, pp. 11909–11914.
67. Povey, S., Lovering, R., Bruford, E. *et al.* (2001), 'The HUGO Gene Nomenclature Committee (HGNC)', *Hum. Genet.* Vol. 109, pp. 678–680.