Refinement of the 22q12-q13 Breast Cancer–Associated Region: Evidence of *TMPRSS6* as a Candidate Gene in an Eastern Finnish Population

Jaana M. Hartikainen,[1,3,4] Hanna Tuhkanen,[1,3,4] Vesa Kataja,[4] Matti Eskelinen,[5] Matti Uusitupa,[2] Veli-Matti Kosma,[1,3] and Arto Mannermaa[1,6]

Abstract Although many risk factors for breast cancer are known, most of the genetic background and molecular mechanisms still remain to be elucidated. We have previously published an autosome-wide microsatellite scan for breast cancer association and here we report a follow-up study for one of the detected regions. Ten single nucleotide polymorphisms (SNP) were genotyped in an Eastern Finnish population sample of 497 breast cancer cases and 458 controls to refine the 550-kb region on 22q12-q13 and identify the breast cancer–associated gene(s) in this region. We also studied 22q12-q13 for allelic imbalance for the detection of a possible tumor suppressor gene and to see whether the breast cancer association and allelic imbalance in this region could be connected. A SNP (rs733655) in matriptase-2 gene (*TMPRSS6*) was detected to associate with breast cancer risk. The genotype frequencies of rs733655 differed significantly between cases and controls in the entire sample and in the geographically and genetically more homogeneous subsample with $P = 0.044$ and $P = 0.0003$, respectively. The heterozygous genotype TC was observed to be the risk genotype in both samples (odds ratios, 1.39; 95% confidence intervals, 1.06-1.83 and odds ratios, 2.11; 95% confidence intervals, 1.46-3.05). An associated two-marker haplotype involving SNP rs733655 (empirical $P = 0.041$) provides further evidence for breast cancer risk factor locating on 22q12-q13, possibly being *TMPRSS6*. Our results suggest that matriptase-2 gene is associated with breast cancer risk in the Eastern Finnish population.

Breast cancer is the most common of cancers among women in western countries. Susceptibility genes thus far identified (e.g., *BRCA1*, *BRCA2*, *ATM* and *CHEK2*) explain ~20% of the familial aggregation of breast cancer and the yet unidentified genes presumably are numerous and confer a moderate risk (1). Such low-penetrance susceptibility genes are likely to interact with environmental and life-style factors as well as with other genetic factors to cause disease. Linkage disequilibrium (LD) based genetic association studies are suitable tools for detecting these genes, and young, rapidly grown isolated populations may provide more help by reducing the genetic heterogeneity. For example, single nucleotide polymorphism (SNP) haplotypes in the Finnish population have been successfully used in the detection of susceptibility genes for complex disease (2). The population history makes the Eastern Finns in the late settlement region especially suitable for LD analysis and association studies (3, 4).

We have previously reported an autosome-wide microsatellite scan for LD-based association with breast cancer in this Eastern Finnish population where we found three chromosomal regions as candidate locations for genetic breast cancer risk factors (5). One of them locates on 22q12-q13, in which three microsatellites defined a 550-kb breast cancer–associated region. To refine this association in the present study, we genotyped 10 SNPs in a 516-kb region in the vicinity of the three microsatellites in a set of 497 Eastern Finnish breast cancer cases and 458 controls. On chromosome 22q12-q13, several studies have reported allelic imbalance (AI) in different tumor tissues, e.g., breast, ovary, and colon, and it has been suggested as a possible location for a tumor suppressor gene (6–10). Therefore, to find further evidence on whether there is a connection between the described AI and the association that we observed, we analyzed 45 breast cancer tumors for AI. Here, we report the results of the SNP association and AI analysis of the 22q12-q13 region as

Authors' Affiliations: Departments of [1]Pathology and Forensic Medicine, and [2]Clinical Nutrition, University of Kuopio, Departments of [3]Pathology, [4]Oncology, and [5]Surgery, Kuopio University Hospital, Kuopio, and [6]Department of Clinical Genetics, Oulu University Hospital, Oulu, Finland

Received 6/29/05; revised 11/4/05; accepted 12/16/05.

Grant support: The Finnish Cultural Foundation of Northern Savo, Special Government Funding (EVO) of Kuopio University Hospital (no. 5654113), Northern Savo Cancer Society, Emil Aaltonen Foundation, Kuopio University (Saastamoinen Foundation), Paavo Koistinen Foundation, Kuopio University Foundation, Finnish Cancer Association and European Union (Marie Curie Individual Fellowship for J.M. Hartikainen).

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.

Requests for reprints: Jaana M. Hartikainen, Department of Pathology and Forensic Medicine, University of Kuopio, P.O. Box 1627, FI-70211 Kuopio, Finland. Phone: 358-17-162754; Fax: 358-17-162753; E-mail: jaana.hartikainen@uku.fi.

doi:10.1158/1078-0432.CCR-05-1417

evidence of a breast cancer risk–associated gene located in this region.

Materials and Methods

Samples. The entire sample is a set of 497 breast cancer cases and 458 controls from the province of Northern Savo in Eastern Finland. The cases represent 96% of the total of 516 breast cancers diagnosed in Kuopio University Hospital between April 1990 and December 1995, and the 458 age and long-term area-of-residence matched controls were selected from the National Population Register during the same time period. The sample material is described in more detail in refs. (5, 11). Genomic DNA was extracted from peripheral blood lymphocytes of both cases and controls using standard procedures (12).

The entire sample set (497 cases and 458 controls) is a population sample that also includes women born outside of the Northern Savo region. In order to get as genetically homogeneous a subset of the population as possible, we selected a stratified subset of the entire sample. In the stratified sample, we included only cases born in the province of Northern Savo and their age and long-term area-of-residence matched controls. Altogether, 280 of the 497 cases and 257 of the 458 controls from the entire sample were included in the stratified subsample. The cases between the two sample sets did not differ in body mass index, tumor histology, histologic grade, stage, lymph node status, estrogen receptor status, or progesterone receptor status. Mean age at diagnosis in the entire sample was 58.9 years, and 54.3 years (median 56.3 and 53.2, respectively) in the subsample.

For the AI analysis, we used paraffin-embedded samples of breast cancer tissue from 45 of the original 49 cases included in the initial autosome scan (5). Samples were microdissected by an experienced pathologist in carefully matched tumor areas containing at least 70% tumor cells (13). For each specimen, DNA was extracted from five 10-μm-thick consecutive sections according to standard protocols (14). The Kuopio Breast Cancer Project has been approved by the joint ethics committee of Kuopio University and Kuopio University Hospital.

SNP validating. We searched SNPs in the genes on 22q12-q13 from databases (http://www.sanger.ac.uk, http://www.ncbi.nlm.nih.gov/SNP, http://snp.cshl.org/db/snp) and confirmed the frequencies of 26 suggested SNPs on denaturing high-pressure liquid chromatography (Transgenomics, Crewe, United Kingdom) according to the manufacturer's instructions. We also screened the coding, 5′-untranslated region and 3′-untranslated region sequences of one of the genes in the associated region (*RABL4*) for sequence variations using denaturing high-pressure liquid chromatography as no coding SNPs for this gene were found in the databases. For validation, we used a set of 48 British and a set of 48 Eastern Finnish samples. The variants from denaturing high-pressure liquid chromatography analyses were re-amplified and sequenced using AbiPrism 377 DNA sequencer and Sequencing Analysis 3.3 software (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions. All reaction conditions, PCR programs and primer sequences are available on request.

Ten of the validated SNPs were selected for genotyping in 497 cases and 458 controls. The criteria for SNPs to be selected were rare allele frequency of 10% or more and suitability to be detected by TaqMan assays. Coding SNPs were preferred, as well as location and distance between adjacent SNPs were considered so that the region would be covered as evenly as possible. The selected SNPs cover a 516-kb region in chromosome 22q12-q13 and adjacent SNPs are 0.9 to 148 kb apart (average distance, 57 kb; Table 1).

SNP genotyping. Genotyping was carried out using the 5′ nuclease assay (TaqMan, Applied Biosystems) in 15 μL reactions in 96-well format as previously described (15). Primer and probe sequences and annealing temperatures are shown in Table 1.

AI analysis. For the AI studies, we used 13 microsatellite markers from the initial autosome-wide scan (5), spanning ~6.8 Mb on chromosome 22q12-q13 (Fig. 1). Intermarker distances and order were obtained at http://www.sanger.ac.uk and Human Genome Browser Gateway at the University of California at Santa Cruz (http://genome.ucsc.edu). The AI analysis was conducted as previously described (13).

Statistical analyses. The significance levels for comparisons of the SNP allele and genotype frequencies between cases and controls were computed using Fisher's exact test and Monte Carlo approximation implemented in SPSS v 11.5. The consistency of the genotypes with the Hardy-Weinberg equilibrium (HWE) was calculated using the standard χ^2 test. Breast cancer–associated risks for the SNP genotypes and for rare/common allele carriers were estimated as odds ratios (OR) with 95% confidence intervals (CI) using cross-tabulation in SPSS v 11.5. Haplotype frequencies for multiple loci were estimated using the expectation-maximization algorithm (16) implemented in SNPAlyze v4.0 software. SNPAlyze was also used for testing the association of the estimated haplotypes and breast cancer. This software implements a global test for haplotype frequency differences between cases and controls. Moreover, SNPAlyze provides haplotype-specific tests which allow evaluation of all individual haplotypes. It also provides permutated (empirical) P values for the global test and individual haplotypes. All permutated P values were reached using 10,000 replicates. $P \leq 0.05$ were considered significant in all analyses.

LD between two SNPs was estimated by calculating the D' values for all 45 possible pairs of the 10 SNPs using the combined set of cases and controls (altogether, 891-931 samples for each SNP; ref. 17).

Power estimations for the SNP association studies were calculated using the Genetic Power Calculator, case-control for discrete traits at http://statgen.iop.kcl.ac.uk/gpc/cc2.html (18). In these estimations, we used $\alpha = 0.05$ and breast cancer prevalence of 0.8% (19). D' was set at 1 and the allele frequencies were assumed equal for the risk SNP and the marker SNP. Also the risks were assumed equal (1.5 or 2) for the homozygous and heterozygous high-risk allele carrying genotypes.

Results

SNP genotyping. Breast cancer association was tested by comparing the allele and genotype frequencies of the genotyped SNPs between cases and controls separately in the entire sample and in the subsample. Genotype- and allele-specific risks were also calculated for both sample sets. In the entire study sample, no significant difference in the allele frequencies between cases and controls were observed with any of the studied 10 SNPs (data not shown). With rs733655, the difference in genotype frequencies between cases and controls in the entire sample was significant with $P = 0.044$ (Table 2). With the same SNP, a near significant OR, 1.28 (95% CI, 0.99-1.50) was observed for the rare allele C carriers (TT genotype versus TC and CC combined, not shown) and the genotype-specific risk for heterozygous genotype TC was significant (OR, 1.39; 95% CI, 1.06-1.83; Table 2). Differences in genotype frequencies between cases and controls with the other SNPs were not significant in the entire sample.

In the stratified subsample (which included only cases born in the province of Northern Savo and their age and long-term area-of-residence matched controls), the most significant association was also detected at SNP rs733655 in the matriptase-2 gene (*TMPRSS6*). The allele frequencies differ between cases and controls with $P = 0.009$ (data not shown) and

Table 1. Primer and probe sequences of the genotyped SNPs

Locus RefSNP ID	Contig position nucleotide (NT.011520)	Distance between SNPs (kb)	PCR primer sequence
rs738977	16399443		F: GAGGATGCTTTTTACCAGTGGTTT
			R: AGCATTAGCGGGAAAGGATGT
rs3827351	16547755	148.3	F: TTGATAGAACTCAGAAATTAAAACCCAC
			R: CTTCAACACCTCAACAGAGTTTAACC
rs3484	16587386	39.6	F: CTCTGCCCTGAACACCCAA
			R: CGAACTGAACAGAAATGCAGGA
rs760517	16649501	62.1	F: TGCCCTCCAAGGGTTCTTC
			R: AGACACACTCGCACAGTTCACAG
rs738148	16661017	11.5	F: CCTGGAGAGGAGCTAGTAGTGACC
			R: AGAGCCCTGAGTTCCTGAGATG
rs7285064	16720514	59.5	F: GGGATTTTCTGGCAGTGAGAAA
			R: TCAGGACCTCCCGCACC
rs25095	16824543	104.0	F: CGGGTGACTTAGTGTAATGATACTCAG
			R: TGCACAGGAGGTGCGAATAC
rs733655	16885566	61.0	F: GTGTGTGCTAACCACCTACTACATGG
			R: CAGAGCCACGCCTTTCTTACC
rs228941	16914236	28.7	F: TGACCTCTTCCCTGGGTTTTCT
			R: GTGGCCATATTTGGGTTTTGG
rs228942	16915134	0.9	F: CTGTGGGTGCCCCGG
			R: CCTGCCAGGTGTACTTTACTTACGA

*Probe label: v, VIC; f, FAM; A1, allele 1 (major); A2, allele 2 (minor).
†Polymorphic bases were underlined.
‡Complementary sequence.

genotype frequencies with $P = 0.0003$ (Table 3). With rs733655, a significant OR of 1.90 (95% CI, 1.34-2.69) was observed for the rare allele C carriers (TT versus TC and CC; data not shown) and also the heterozygous genotype TC was observed to be the risk genotype with OR, 2.11 (95% CI, 1.46-3.05; Table 3). In this subsample, the allele frequencies of SNP rs7285064 in the *CSF2RB* gene also differed between cases and controls with $P = 0.034$ (data not shown) and the genotype-specific risk for the heterozygous genotype CT was significant, indicating a protective effect (OR, 0.62; 95% CI, 0.41-0.96; Table 3). Differences in genotype and allele frequencies between cases and controls with the other SNPs were not significant in the subsample.

The deviation of the genotype frequencies from the HWE was tested separately for the cases and controls in the entire sample and in the subsample. In the controls, the rs733655 genotypes deviated slightly from HWE in the entire sample and significantly in the controls of the subsample ($P = 0.041$ and 0.008, respectively), whereas the cases were in equilibrium. Other SNPs were in equilibrium in cases and controls in both sample sets (Tables 2 and 3).

LD analysis. LD between SNPs was estimated in the entire sample by calculating pairwise D' values for all 45 possible SNP pairs. In 6 of the 45 pairs, $D' \geq 0.5$ were observed (Table 4). As expected, D' was detected to decline when the distance between SNPs increased. The highest $D' = 0.99$ was obtained for a pair of SNPs that are in the same gene (rs228941 and rs228942) and are separated by 898 bp (Table 4). However, we observed moderate LD between SNPs that are separated by up to 236 kb ($D' = 0.49$-0.61).

Haplotypes. Two-marker haplotypes were estimated for adjacent SNPs that were in LD with each other ($D' \geq 0.5$). Altogether, five SNP pairs were tested (Table 5). Of these, the pair rs25095 and rs733655 (61 kb apart), showed significant difference in haplotype frequencies between cases and controls with empirical global $P = 0.041$ in the subsample (Table 5). An individual haplotype AC of this SNP combination is significantly more frequent in cases (empirical $P = 0.003$). In the entire sample, this haplotype AC is also more frequent in cases with borderline significance, $P = 0.055$ (empirical $P = 0.073$), but the global P for the rs25095 and rs733655 haplotypes is not significant (Table 5). With other tested two-marker haplotypes, the global empirical P for difference between cases and controls was not significant.

Allelic imbalance. AI was detected in 82% (37 of 45) of the tumors at 1 or more of the 13 microsatellite marker loci studied on chromosome 22q12-q13. Altogether, five microsatellite markers, *D22S1142*, *D22S924*, *D22S1177*, *D22S445*, and *D22S279*, showed significant AI (Fig. 1). Two of these markers, *D22S1177* and *D22S445*, locate in the breast cancer–associated 550-kb region (5). However, marker *IL2RB* being the closest marker to *TMPRSS6* gene, where the significant SNP association was detected, did not exhibit significant AI.

Discussion

Here, we report the results of a further association study using 10 SNPs to refine a 516-kb region in a previously detected breast cancer–associated 550-kb region on chromosome

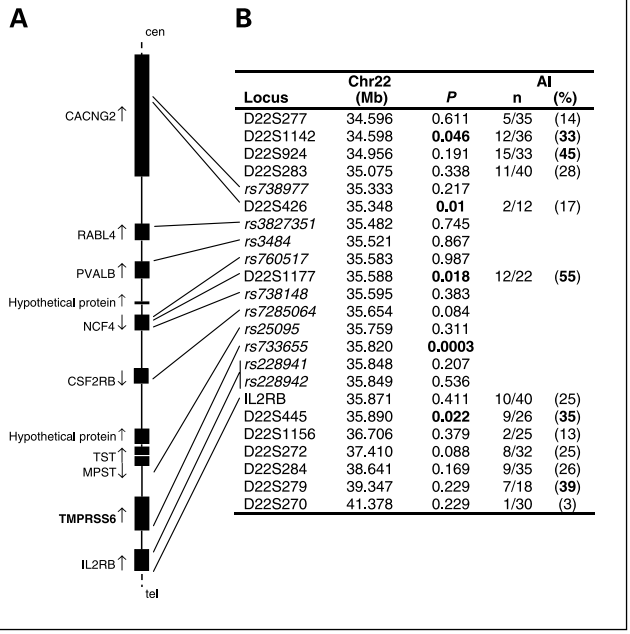
Table 1. Primer and probe sequences of the genotyped SNPs (Cont'd)

Locus RefSNP ID	Probe sequence*†	Annealing temperature (°C)
rs738977	A1: CGGCTGCGGGTATTTCATCCAT (v) ‡	62
	A2: CCGGCTGCAGGTATTTCATCCATT (f) ‡	62
rs3827351	A1: CCAAAACATGAGTCACCAAAGGCCA (v) ‡	62
	A2: CAAAACAGGAGTCACCAAAGGCCAGT (f) ‡	62
rs3484	A1: CTCGGCCCCTCTTGCCACC (v) ‡	62
	A2: CTCGGCCCCTCTCGCCAC (f) ‡	62
rs760517	A1: TCCAAAACCTCCCCCAGGCC (f)	60
	A2: TCCAAAACTTCCCCCAGGCCA (v)	60
rs738148	A1: TTAGCCAGTTTGGCAGTTTCCATTTTACC (v)	64
	A2: TTTAGCCAGTTTAGCAGTTTCCATTTTACCCT (f)	64
rs7285064	A1: CTCACTCCACTCGCTCCAGATCCC (f) ‡	62
	A2: CTCACTCCACTCACTCCAGATCCCG (v) ‡	62
rs25095	A1: ACTTTCTATTGCAATTGTTTCCCCAACG (v) ‡	62
	A2: TTCTATTGCAATCGTTTCCCCAACG (f) ‡	62
rs733655	A1: TGCCTCCCCTTGTGAAGCTGACA (v)	62
	A2: CCTCCCCTCGTGAAGCTGACAGTG (f)	62
rs228941	A1: CTCCCCTCCCCGTCCACAGG (f)	62
	A2: CCTCCCCCTCCACAGGGCA (v)	62
rs228942	A1: CACCCTCATCAGGGTCTTCCTCTGAG (f) ‡	62
	A2: ACACCCTCATCAGGTTCTTCCTCTGAGTAG (v) ‡	62

22q12-q13 (5) and AI analysis for a larger region on 22q12-q13 for the detection of a possible tumor suppressor gene(s). A SNP in matriptase-2 gene (*TMPRSS6*) was detected to associate with breast cancer risk. Our results suggest that the matriptase-2 gene is a candidate for breast cancer risk factor in this Eastern Finnish population.

Significant breast cancer association was detected with SNP rs733655. This SNP locates in the intronic sequence of *TMPRSS6* gene which encodes a membrane-bound serine proteinase 6 called matriptase-2. Matriptase-2 is a member of a family of type II transmembrane serine proteinases which have a possible role in cancer development. Matriptase-2 has the ability to degrade extracellular matrix components, suggesting that it may participate in some of the matrix-degrading processes occurring in both normal and pathologic conditions, including cancer progression (20). Because matriptase-2 has only recently been discovered, little is known about its physiologic function(s). Matriptase-2 is expressed in normal breast tissue and the expression is elevated in breast cancer (invasive ductal carcinoma; ref. 21), which points out that matriptase-2 is not expected to be a tumor suppressor. Our AI analysis provided further support for this because AI was not detected in the vicinity of the *TMPRSS6* gene. This also indicates that the detected AI seems to be a separate event from the original breast cancer association (Fig. 1), i.e., hereditary risk factor, whereas AI may indicate the involvement of a nonhereditary factor. However, in line with other studies (6–10), the observed AI on the 22q12-q13 region does not exclude the existence of a tumor suppressor, other than *TMPRSS6*.

The rs733655 allele frequencies that we observed among controls are similar to those expected in the National Center



Locus	Chr22 (Mb)	*P*	AI n	AI (%)
D22S277	34.596	0.611	5/35	(14)
D22S1142	34.598	**0.046**	12/36	(**33**)
D22S924	34.956	0.191	15/33	(**45**)
D22S283	35.075	0.338	11/40	(28)
rs738977	35.333	0.217		
D22S426	35.348	**0.01**	2/12	(17)
rs3827351	35.482	0.745		
rs3484	35.521	0.867		
rs760517	35.583	0.987		
D22S1177	35.588	**0.018**	12/22	(**55**)
rs738148	35.595	0.383		
rs7285064	35.654	0.084		
rs25095	35.759	0.311		
rs733655	35.820	**0.0003**		
rs228941	35.848	0.207		
rs228942	35.849	0.536		
IL2RB	35.871	0.411	10/40	(25)
D22S445	35.890	**0.022**	9/26	(**35**)
D22S1156	36.706	0.379	2/25	(13)
D22S272	37.410	0.088	8/32	(25)
D22S284	38.641	0.169	9/35	(26)
D22S279	39.347	0.229	7/18	(**39**)
D22S270	41.378	0.229	1/30	(3)

Fig. 1. The breast cancer – associated region on 22q12-q13 combining the results of SNP association and AI analyses, as well as the original microsatellite association. *A,* gene map of the SNP-covered 516-kb region. Black boxes, genes; arrows, direction of transcription of each gene; diagonal lines, approximate location of the studied SNP (in italics) and microsatellite markers in this region. *B,* locations of the studied markers on chromosome 22 (http://genome.ucsc.edu), *P* values for breast cancer association and observed AI. *P* values for difference in allele frequencies between cases and controls for microsatellites (from ref. 5) and *P* values for the difference in genotype frequencies between cases and controls for the SNPs (in the subsample). For AI, the number of tumors with AI/total number of informative tumors (and AI percentage) at each studied locus. Significant *P* values and AI% in boldface.

for Biotechnology Information RefSNP database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Display&DB=snp). In the RefSNP European test population sample, the HWE probability is 0.05, indicating that in normal situations, the genotype frequencies deviate from those expected under HWE. This deviation was also detected in our controls group. The allele frequencies among our cases differ from those expected (RefSNP) as genotypes that we observed do not deviate from HWE. If the number of controls by genotype is replaced by the numbers expected under HWE, the difference in genotype frequencies between cases and controls becomes nonsignificant in the entire sample (P = 0.338), but still remains significant in the subsample (which includes only cases born in the province of Northern Savo and their age and long-term area-of-residence matched controls; P = 0.013). Also, the risk for the heterozygous genotype TC becomes nonsignificant in the entire sample but remains significant in the subsample, although reducing from 2.11 to 1.67 (CI, 1.17-2.39). Therefore, we consider that the detected difference between the case and control populations is real. The rare allele C carriers had increased risk of breast cancer in the subsample (OR, 1.90; 95% CI, 1.34-2.69) and nearly significant risk in the entire sample (OR, 1.28; 95% CI, 0.99-1.50) but the rare homozygous genotype CC did not associate significantly with breast cancer. This may be partly explained by too small number of observed rare homozygotes alone to show significant increase in risk (Tables 2 and 3). When validating the SNPs for this study, we sequenced 20 samples for rs733655 and no sequence variation other than rs733655 was observed in the region of the designed TaqMan primer and probe sequences. Genotyping assays were optimized and allele calling was unambiguous as allele controls were used. Also, 141 samples from altogether six plates were re-genotyped (original genotype was confirmed) and thus, a genotyping error is also excluded. Although it is common to exclude markers which deviate from HWE from association studies, it is possible to miss true risk factors by doing so. A recent report described such a polymorphism in the urokinase-type plasminogen activator gene associating with Alzheimer's disease (22).

A common way to enhance the power of genetic case-control association studies is to combine the statistical power

Table 2. Genotype frequencies and genotype-specific risks in the entire sample

SNP	*n*		Genotype	Genotype counts		Genotype frequencies		*P* for HWE test		χ^2 *P* for difference in genotype frequencies	Genotype-specific risk	
	Cases	Controls		Cases	Controls	Cases	Controls	Cases	Controls		OR	CI
rs738977	486	430	CC	340	301	0.700	0.700	0.414	0.655	0.982		
			CT	130	116	0.267	0.270				0.99	0.74-1.33
			TT	16	13	0.033	0.030				1.09	0.52-2.30
rs3827351	486	425	AA	253	228	0.521	0.537	0.594	0.776	0.839		
			AC	192	165	0.395	0.388				1.05	0.80-1.38
			CC	41	32	0.084	0.075				1.15	0.70-1.90
rs3484	484	435	AA	155	116	0.320	0.266	0.540	0.567	0.198		
			AG	233	223	0.480	0.513				0.78	0.58-1.06
			GG	98	96	0.200	0.221				0.76	0.52-1.1
rs760517	480	433	CC	228	193	0.475	0.446	0.701	0.576	0.641		
			CT	203	196	0.423	0.453				0.88	0.67-1.15
			TT	49	44	0.102	0.101				0.94	0.6-1.48
rs738148	480	440	GG	254	229	0.529	0.520	0.716	0.442	0.527		
			GA	193	172	0.402	0.391				1.01	0.77-1.33
			AA	33	39	0.069	0.089				0.76	0.46-1.25
rs7285064	483	430	CC	375	327	0.776	0.761	0.751	0.988	0.784		
			CT	102	96	0.211	0.223				0.93	0.68-1.27
			TT	6	7	0.013	0.016				0.75	0.25-2.25
rs25095	484	416	AA	135	106	0.279	0.255	0.341	0.689	0.438		
			AG	251	212	0.519	0.510				0.93	0.68-1.27
			GG	98	98	0.202	0.235				0.79	0.54-1.15
rs733655	485	443	TT	246	254	0.507	0.573	0.321	0.041	0.044		
			TC	205	152	0.423	0.343				1.39	1.06-1.83
			CC	34	37	0.070	0.084				0.95	0.58-1.56
rs228941	487	440	GG	240	234	0.493	0.532	0.566	0.934	0.502		
			GC	208	173	0.427	0.393				1.17	0.89-1.54
			CC	39	33	0.080	0.075				1.15	0.70-1.89
rs228942	484	444	CC	307	302	0.634	0.680	0.812	0.528	0.329		
			CA	158	126	0.326	0.284				1.23	0.93-1.63
			AA	19	16	0.040	0.036				1.17	0.59-2.31

Table 3. Genotype frequencies and genotype-specific risks in the stratified subsample

SNP	*n*		Genotype	Genotype counts		Genotype frequencies		*P* for HWE test		χ^2 *P* for difference in genotype frequencies	Genotype-specific risk	
	Cases	Controls		Cases	Controls	Cases	Controls	Cases	Controls		OR	CI
rs738977	279	247	CC	212	174	0.760	0.705	0.052	0.881	0.217		
			CT	58	67	0.208	0.271				0.71	0.47-1.06
			TT	9	6	0.032	0.024				1.23	0.43-3.53
rs3827351	279	242	AA	145	125	0.520	0.516	0.193	0.847	0.745		
			AC	106	97	0.380	0.401				0.94	0.65-1.36
			CC	28	20	0.100	0.083				1.21	0.65-2.25
rs3484	278	249	AA	86	72	0.309	0.289	0.763	0.914	0.867		
			AG	135	123	0.486	0.494				0.92	0.62-1.37
			GG	57	54	0.205	0.217				0.88	0.54-1.44
rs760517	274	246	CC	130	115	0.474	0.467	0.396	0.501	0.987		
			CT	113	103	0.413	0.419				0.97	0.67-1.40
			TT	31	28	0.113	0.114				0.98	0.55-1.73
rs738148	273	248	GG	148	122	0.542	0.492	0.648	0.418	0.383		
			GA	104	100	0.381	0.403				0.86	0.60-1.23
			AA	21	26	0.077	0.105				0.67	0.36-1.24
rs7285064	277	245	CC	227	181	0.819	0.739	0.747	0.701	0.084		
			CT	47	60	0.170	0.245				0.62	0.41-0.96
			TT	3	4	0.011	0.016				0.6	0.13-2.71
rs25095	278	239	AA	82	66	0.295	0.276	0.283	0.767	0.311		
			AG	146	117	0.525	0.490				1	0.67-1.51
			GG	50	56	0.180	0.234				0.72	0.44-1.19
rs733655	278	250	TT	125	152	0.449	0.608	0.082	0.008	0.0003		
			TC	132	76	0.475	0.304				2.11	1.46-3.05
			CC	21	22	0.076	0.088				1.16	0.61-2.21
rs228941	280	249	GG	135	130	0.482	0.522	0.670	0.087	0.207		
			GC	121	107	0.432	0.430				1.09	0.76-1.55
			CC	24	12	0.086	0.048				1.93	0.92-4.01
rs228942	278	251	CC	172	167	0.619	0.665	0.445	0.856	0.536		
			CA	96	76	0.345	0.303				1.23	0.85-1.77
			AA	10	8	0.036	0.032				1.21	0.47-3.15

Table 4. *D′* values for pairwise LD (cases and controls combined, entire sample)

	rs738977	rs3827351	rs3484	rs760517	rs738148	rs7285064	rs25095	rs733655	rs228941	rs228942
rs738977										
rs3827351	0.56									
rs3484	0.05	0.59								
rs760517	0.08	0.26	0.17							
rs738148	0.12	0.04	0.10	0.57						
rs7285064	0.14	0.06	0.08	0.16	0.18					
rs25095	0.04	0.08	0.12	0.15	0.06	0.25				
rs733655	0.13	0.28	0.26	0.04	0.28	0.09	0.49			
rs228941	0.02	0.03	0.13	0.10	0.24	0.19	0.24	0.05		
rs228942	0.23	0.08	0.02	0.07	0.24	0.61	0.01	0.02	0.99	

Table 5. Haplotype frequencies in the entire sample and in the subsample

SNP pair	D'	Entire sample							
		Global P*		Haplotype	Statistics for individual haplotypes				
					Frequency			P †	
		χ^2	Empirical		Overall ‡	Cases§	Controls‖	χ^2	Empirical
rs738977-rs3827351	0.56	0.207	0.410						
				C A	0.578	0.575	0.581	0.824	0.831
				C C	0.259	0.261	0.256	0.849	0.856
				T A	0.143	0.138	0.150	0.472	0.502
				T C	0.020	0.026	0.013	0.042	0.209
rs3827351-rs3484	0.59	0	0.323						
				A G	0.408	0.408	0.429	0.393	0.600
				A A	0.316	0.337	0.301	0.113	0.132
				C A	0.225	0.255	0.217	0.056	0.212
				C G	0.051	0	0.053	0	0.393
rs760517-rs738148	0.57	0.055	0.133						
				C G	0.442	0.443	0.442	0.910	0.926
				T G	0.282	0.290	0.272	0.423	0.449
				C A	0.236	0.239	0.232	0.783	0.798
				T A	0.040	0.028	0.054	0.007	0.039
rs25095-rs733655	0.49	0.243	0.323						
				G T	0.411	0.400	0.424	0.310	0.326
				A T	0.319	0.316	0.321	0.847	0.840
				A C	0.212	0.230	0.191	0.055	0.073
				G C	0.058	0.054	0.064	0.411	0.491
rs228941-rs228942	0.99	0.258	0.253						
				G C	0.711	0.696	0.729	0.135	0.126
				C A	0.197	0.211	0.180	0.110	0.110
				C C	0.092	0.093	0.091	0.884	0.863
				G A	0	0	0		

*Global P value (χ^2 and empirical/permutated) for haplotype frequency difference cases versus controls for given SNP pair.
† P value (χ^2 and empirical/permutated) for haplotype frequency difference of individual haplotype cases versus controls.
‡ Frequency for given haplotype in cases and controls combined.
§Frequency for given haplotype in cases.
‖Frequency for given haplotype in controls.

of associated markers by estimating haplotypes (23). Haplotypes are dependent on the distance and number of meiotic recombinations between studied markers. In our study, we constructed haplotypes for the five marker pairs that were in LD with $D' \geq 0.5$. One pair involving the most associated SNP, rs733655 in the matriptase-2 gene, showed significant difference in haplotype frequencies between cases and controls, enhancing the association with breast cancer in this chromosomal region. Haplotype analysis was not feasible across the whole 550 kb region as the SNPs were not linked tightly enough to allow identification of common multimarker haplotype. However, the finding of an associated haplotype increases the interest on matriptase-2 gene and the next step is to type additional SNPs covering the *TMPRSS6* gene aiming to identify a haplotype and further locate the possible functional change in the gene. To resolve the global importance of *TMPRSS6* in genetic risk of breast cancer, the association also has to be tested in other populations. Another SNP, rs7285064 in the *CSF2RB* gene, showed a moderate association with breast cancer and further evidence concerning the importance of this gene in breast cancer is necessary.

The power to detect risk affecting alterations in LD association studies depends on the allele frequencies of the markers and the strength and length of the LD between the alteration and studied marker. According to the power calculations, our sample set of 497 cases and 458 controls has >95% power to detect a risk allele that is in perfect LD with the marker allele ($D' = 1$) and has a relative risk of 2. The power to detect a risk allele that is not in perfect LD with the marker allele ($D' = 0.5$) and has a relative risk of 2 varies between 44% and 74% with the 10 SNPs. In the stratified sample, the power for the detected risk (2.11 for the TC and 1.16 for the TT genotype) with SNP rs733655 is 83%. Within a short distance (1 kb), we detected highly significant LD ($D' = 1$) and even at distances >200 kb, D' is ≥ 0.5 at some instances. It is presumable that LD does not remain at the constant level across the whole 516 kb region and some associations may therefore have been missed. Furthermore, our negative results do not rule out association

Table 5. Haplotype frequencies in the entire sample and in the subsample (Cont'd)

Global *P**		Haplotype	Subsample				
			Statistics for individual haplotypes				
			Frequency			*P*†	
χ^2	Empirical		Overall‡	Cases§	Controls‖	χ^2	Empirical
0.455	0.636						
		C A	0.588	0.587	0.591	0.907	0.905
		C C	0.267	0.278	0.252	0.357	0.391
		T A	0.126	0.120	0.130	0.626	0.663
		T C	0.019	0.015	0.027	0.177	0.381
0.659	0.719						
		A G	0.407	0.392	0.426	0.296	0.300
		A A	0.306	0.315	0.295	0.522	0.515
		C A	0.235	0.236	0.235	0.919	0.916
		C G	0.051	0.057	0.044	0.404	0.450
0.041	0.104						
		C G	0.436	0.442	0.429	0.681	0.702
		T G	0.281	0.296	0.265	0.270	0.292
		C A	0.241	0.237	0.246	0.713	0.715
		T A	0.042	0.025	0.060	0.006	0.030
0.016	0.041						
		G T	0.395	0.385	0.410	0.433	0.452
		A T	0.321	0.299	0.343	0.143	0.155
		A C	0.224	0.264	0.180	0.002	0.003
		G C	0.060	0.052	0.067	0.325	0.438
0.307	0.287						
		G C	0.711	0.690	0.735	0.126	0.109
		C A	0.199	0.213	0.183	0.241	0.215
		C C	0.090	0.097	0.082	0.426	0.347
		G A	0	0	0		

involving other nearby SNPs and positive results do not necessarily indicate the discovery of the causal SNP but a marker in LD with a true causal SNP located some distance away. Of the genes located in this 516 kb region, matriptase-2 is the most promising candidate for a breast cancer risk gene. Other genes in the studied region have not been reported to associate with (breast) cancer.

In conclusion, the initial autosome-wide scan (5), as well as the present SNP association analysis, has shown that a breast cancer–associated risk factor locates on 22q12-q13. The results here are one step forward in the process of identifying the causative gene and a functional variant within and imply that *TMPRSS6* is the gene of interest in the studied Eastern Finnish population. Thus, the possibility of matriptase-2 involvement in cancer progression is further supported and a more specific investigation for the association between matriptase-2 and breast cancer, as well as the identification of the causal genetic variant is needed.

References

1. Thompson D, Easton D. The genetic epidemiology of breast cancer genes. J Mammary Gland Biol Neoplasia 2004;9:221–36.
2. Laitinen T, Polvi A, Rydman P, et al. Characterization of a common susceptibility locus for asthma-related traits. Science 2004;304:300–4.
3. Norio R. Finnish disease heritage I: characteristics, causes, background. Hum Genet 2003a;112:441–56.
4. Norio R. Finnish disease heritage II: population prehistory and genetic roots of Finns. Hum Genet 2003b;112:457–69.
5. Hartikainen JM, Tuhkanen H, Kataja V, et al. An autosome wide scan for linkage disequilibrium based association in sporadic breast cancer cases in Eastern Finland: three candidate regions found. Cancer Epidemiol Biomarkers Prev 2005;14:75–80.
6. Allione F, Eisinger F, Parc P, Noguchi T, Sobol H, Birnbaum D. Loss of heterozygosity at loci from chromosome arm 22q in human sporadic breast carcinomas. Int J Cancer 1998;75:181–6.
7. Iida A, Kurose K, Isobe R, et al. Mapping of a new target region of allelic loss to a 2-cm interval at 22q13.1 in primary breast cancer. Genes Chromosomes Cancer 1998;21:108–12.
8. Bryan EJ, Thomas NA, Palmer K, Dawson E, Englefield P, Campbell IG. Refinement of an ovarian cancer tumour suppressor gene locus on chromosome arm 22q and mutation analysis of CYP2D6, SREBP2 and NAGA. Int J Cancer 2000;87:798–802.
9. Castells A, Gusella JF, Ramesh V, Rustgi AK. A region of deletion on chromosome 22q13 common to human breast and colorectal cancers. Cancer Res 2000;60:2836–9.
10. Hirano A, Emi M, Tsuneizumi M, et al. Allelic losses of loci at 3p25.1, 8p22, 13q12, 17p13.3, and 22q13 correlate with postoperative recurrence in breast cancer. Clin Cancer Res 2001;7:876–82.
11. Männistö S, Pietinen P, Pyy M, Palmgren J, Eskelinen M, Uusitupa M. Body size indicators and risk of breast cancer according to menopause and estrogen receptor status. Int J Cancer 1996;68:8–13.
12. Vandenplas S, Wiid J, Grobler Rabie A, et al. Blot hybridization of genomic DNA. J Med Genet 1984;21:164–72.
13. Tuhkanen H, Anttila M, Kosma VM, et al. Genetic alterations in the peritumoral stromal cells of malignant and borderline epithelial ovarian tumors as indicated by allelic imbalance on chromosome 3p. Int J Cancer 2004;109:247–52.
14. Karjalainen JM, Kellokoski JK, Mannermaa AJ, et al. Failure in post transcriptional processing is a possible inactivation mechanism of AP-2α in cutaneous melanoma. Br J Cancer 2000;82:2015–21.
15. Kuschel B, Auranen A, McBride S, et al. Variants in DNA double-strand break repair genes and breast cancer susceptibility. Hum Mol Genet 2002;11:1399–407.

16. Slatkin M, Excoffier L. Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. Heredity 1996;76:377–83.
17. Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 1995;29:311–22.
18. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 2003;19:149–50.
19. Finnish Cancer Registry. Cancer Statistics for Finland, March 2, 2004. http://www.cancerregistry.fi/.
20. Velasco G, Cal S, Quesada V, Sanchez LM, Lopes-Otin C. Matriptase-2, a membrane-bound mosaic serine proteinase predominantly expressed in human liver and showing degrading activity against extracellular matrix proteins. J Biol Chem 2002;277:37637–46.
21. Overall CM, Tam EM, Kappelhoff R, et al. Protease degradomics: mass spectrometry discovery of protease substrates and the CLIP-CHIP, a dedicated DNA microarray of all human proteases and inhibitors. Biol Chem 2004;385:493–504.
22. Ertekin-Taner N, Ronald J, Feuk L, et al. Elevated amyloid β protein (Aβ42) and late onset Alzheimer's disease are associated with single nucleotide polymorphisms in the urokinase-type plasminogen activator gene. Hum Mol Genet 2005;14:447–60.
23. Rioux JD, Daly MJ, Silverberg MS, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 2001;29:223–8.