An Autosome-Wide Scan for Linkage Disequilibrium–Based Association in Sporadic Breast Cancer Cases in Eastern Finland: Three Candidate Regions Found

Jaana M. Hartikainen,[1,4] Hanna Tuhkanen,[1,4] Vesa Kataja,[4] Alison M. Dunning,[6] Antonis Antoniou,[7] Paula Smith,[7] Airi Arffman,[8] Mia Pirskanen,[2] Douglas F. Easton,[7] Matti Eskelinen,[5] Matti Uusitupa,[3] Veli-Matti Kosma,[1,9] and Arto Mannermaa[1,9,10]

Departments of [1]Pathology and Forensic Medicine, [2]Neuroscience and Neurology, and [3]Clinical Nutrition, University of Kuopio; Departments of [4]Oncology and [5]Surgery, Kuopio University Hospital, Kuopio, Finland; [6]Cancer Research UK Human Cancer Genetics Research Group, Department of Oncology, and [7]Cancer Research UK Genetic Epidemiology Unit, Strangeways Research Laboratory, University of Cambridge, Cambridge, United Kingdom; [8]Fujitsu, Fujitsu, Finland; [9]Department of Pathology, Centre for Laboratory Medicine, Tampere University Hospital, Tampere, Finland; and [10]Department of Clinical Genetics, Oulu University Hospital, Oulu, Finland

Abstract

Breast cancer is the most common of cancers among women in industrialized countries. Many of breast cancer risk factors are known, but the majority of the genetic background is still unknown. Linkage disequilibrium–based association is a powerful tool for mapping disease genes and is suitable for mapping complex traits in founder populations. We report the results of a two-stage, autosome-wide scan for LD with breast cancer. Our aim was to identify genetic risk factors for sporadic breast cancer in an eastern Finnish population. Our case-control set is from the province of northern Savo in the late-settlement area of eastern Finland. This population is relatively young and genetically homogeneous. We used 435 autosomal microsatellite markers spaced by an average of 10 cM in a set of 49 breast cancer cases and 50 controls. In the first-stage scan, we found 21 markers in LD with breast cancer (*P*s = 0.003-0.046, Fisher's exact test). In the second-stage scan with markers flanking 21 positive loci, four significant markers were found (*P*s = 0.013-0.046, Fisher's exact test). Haplotype analysis using global score method with two, three, or four markers also revealed four positive marker combinations (simulated *P* for global score = 0.003-0.021). Our results suggest breast cancer–associated regions on 3p26, 11q23, and 22q13.1 in an eastern Finnish population. (Cancer Epidemiol Biomarkers Prev 2005;14(1):75–80)

Introduction

Breast cancer is the most common cancer among women in industrialized countries. It is estimated that 1 in 10 women in Finland will have breast cancer at some point in her life. Every year, >3,500 new breast cancer cases are diagnosed in Finland (population 5.2 million), and the number is increasing yearly (Finnish Cancer Registry, http://www.cancerregistry.fi). Mutations in *BRCA1* (1, 2) and *BRCA2* (3, 4) genes are known to confer a high lifetime risk of breast cancer. However, it is estimated that they explain only ~15% of the observed familial relative risk of breast cancer (5). In addition, in a study investigating the genetic models of the non-*BRCA1*/*BRCA2* familial clustering of breast cancer, the findings suggest that several common, low-penetrance genes may account for the residual familial aggregation of breast cancer (6). Therefore, it is likely that other breast cancer susceptibility genes remain to be identified. It is recognized that a woman with a first-degree relative with breast cancer has twice the risk of developing the disease herself, indicating the presence of low-penetrance genes for breast cancer susceptibility in the general population (7). Such low-penetrance susceptibility genes are likely to interact with environmental and lifestyle factors as well as with other genetic factors to cause disease.

Case-control genetic association studies are a more efficient way of identifying low- and moderate-penetrance risk-altering genes than linkage studies of families. Linkage disequilibrium (LD) or allelic association is based on the assumption that the affected share a genetic variant/mutation, which is so close to the marker that the probability of a recombination event to occur between them is minimal. In young (15-25 generations) isolated populations, the number of meioses is relatively low and LD is thought to extend further than in older populations (8). In addition, the rate of recombination varies between chromosomes and within chromosomes (9). LD analysis has been used successfully to discover genes for several diseases of the Finnish disease heritage (10-15) as well as in other isolated populations and complex diseases (8, 16-18).

Finland has been inhabited for ~10,000 years. The size of the Finnish population has been estimated to have been ~2,500 to 10,000 until 3,300 years ago, and only during the last 2,000 to 2,500 years, especially in the most recent 10 to 12 generations, the number of people has grown rapidly to the present-day 5.2 millions. There has been very little immigration in Finland in the last 80 to 100 generations of expansion, and repeated population bottlenecks have occurred, the last one as late as the early 18th century. A relatively small group of founders from the early-settlement area populated the late-settlement area (the eastern and northern parts of Finland) only after 1500. The population of the late-settlement area has been expanding for 20 to 25 generations in isolation (mainly because of distance) and is characterized by subisolates with ≤50,000 inhabitants (19-22). This population history makes the eastern Finns an especially suitable material for LD analysis (23, 24). Due to the founder effect, many diseases, mainly autosomal recessive, are highly prevalent in Finland and rare elsewhere (20). In cancer predisposition syndromes (e.g., hereditary nonpolyposis colorectal cancer), dominant founder

Received 3/8/04; revised 6/7/04; accepted 7/19/04.

Grant support: Finnish Cultural Foundation of Northern Savo, Special Government Funding (EVO) of Kuopio University Hospital (5654113), Northern Savo Cancer Society, Emil Aaltonen Foundation, Kuopio University (Saastamoinen Foundation), Paavo Koistinen Foundation, Kuopio University Foundation, and European Union Marie Curie Individual Fellowship (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

mutations have spread and become highly enriched and specific for certain areas (25, 26). The same effect can be seen with *BRCA1* and *BRCA2* mutations that predispose to breast and ovarian cancer (27), although in the Finnish breast and ovarian cancer population the proportion of *BRCA1*/*BRCA2* mutations is lower than initially expected and some other genes also are likely to be involved in breast cancer risk and development (28). Thus, in a homogeneous founder population, low-penetrance susceptibility genes or alleles are also enriched.

We report the first published autosome-wide microsatellite screen for LD with breast cancer predisposing alleles. Our aim was to find new genetic moderate- or low-penetrance risk factors for sporadic breast cancer by identifying chromosomal regions that are associated with breast cancer by screening them for LD in a breast cancer case-control set from eastern Finland. We screened the autosomes with 366 polymorphic microsatellite markers and further analyzed 69 markers flanking significant loci and calculated differences between frequencies of the estimated two-, three-, and four-marker haplotypes in case and control groups.

Materials and Methods

Cases and Controls. Our sample material is a carefully selected case-control set from the province of northern Savo in the late-settlement area of eastern Finland. Cases for this study were ascertained through the Kuopio Breast Cancer Project (29, 30). The material for Kuopio Breast Cancer Project was collected from women (with any suspected breast disease) who came to Kuopio University Hospital (hospital district of 1 million people) for breast examination between April 1990 and December 1995. Of these women, 516 were diagnosed and histologically confirmed to have breast cancer. The Kuopio Breast Cancer Project material also includes randomly selected and individually matched age (within ±5 years) and area-of-residence controls from the National Population Register for each case of breast cancer. For the LD study, we have selected from this case-control material 49 cases who did not have a strong family history of cancer according to the initial interview and who were born around Kuopio in four rural communities whereto moving has been minimal. We also included 50 controls matched for age and long-term area of residence. Age of onset of breast cancer ranged between 30 and 82 years, median age of onset being 54.0 years. Seventy-three percent of the cases had breast cancer at or after age 50. The median age of the controls was 54.5 years (range, 30-76 years). Each patient gave informed written consent for participation in the study. The joint ethics committee of Kuopio University and Kuopio University Hospital has approved the Kuopio Breast Cancer Project.

Genotyping. Genomic DNA was extracted from peripheral blood lymphocytes using standard procedure (31). For the first-stage screening of the autosomes, we used 366 markers from the sixth version of the Weber lab microsatellite marker set (http://research.marshfieldclinic.org/genetics/sets/Set6ScreenFrames.htm), markers spanning ~10 cM (range, 1-20 cM) apart in the genome. The average heterozygosity of the markers was 0.76. The microsatellite repeats were PCR amplified (single or multiplex) with fluorescent-labeled primers using standard procedures (32) and pooled for size determination with an automated ABIPrism 310 Genetic Analyzer (Applied Biosystems, Foster City, CA). In the second-stage scan, an ABIPrism 377 DNA Sequencer (Applied Biosystems) was used. The number of samples in this study was designed so that it was possible to run all controls, 45 of the cases and a blank control in one run that takes 96 samples on ABIPrism 310. This was done to minimize possible variations in the sample running conditions (e.g., by using the same capillary for all samples). The remaining four of the cases were run first on another run to recheck that marker is fine. The runs were analyzed using GeneScan 2.1 software (Applied Biosystems). All samples included an internal GeneScan-500 [TAMRA] size standard (Applied Biosystems) according to which the size of the fragment was determined. When LD with breast cancer was observed, the second-stage scan with additional markers from the Genome Database (http://www.gdb.org/) located close to those with significant (or near significant) *P*s was done. Altogether, we used 69 additional markers. Intermarker distances and order were obtained from the databases at the Sanger Institute (http://www.sanger.ac.uk), Human Genome Browser Gateway at the University of California at Santa Cruz (http://genome.ucsc.edu), Cedar Genetics at the University of Southampton (http://cedar.genetics.soton.ac.uk), and Marshfield Center for Medical Genetics (http://research.marshfieldclinic.org/genetics).

Statistical Methods

Allele Frequencies. Significance levels for comparisons of the allele frequencies between cases and controls were computed using Fisher's exact test and Monte Carlo approximation implemented in SPSS version 11.5. χ^2 and Fisher's exact tests were also done to compare the frequency of the possible associated allele versus the frequency of all other alleles pooled together between cases and controls. Breast cancer–associated risks for the associated alleles of the significant markers were estimated as odds ratios with 95% confidence intervals using 2 × 2 cross-tabulation in SPSS version 11.5.

Genotype Frequencies and Hardy-Weinberg Equilibrium. Significance levels for comparisons of the genotype frequencies between cases and controls were computed using Fisher's exact test and Monte Carlo approximation implemented in SPSS version 11.5. Deviation of the genotype distributions from Hardy-Weinberg equilibrium (HWE) was checked for the combined set (cases and controls) in the markers that were in LD with breast cancer using HWE and Associate software (http://linkage.rockefeller.edu/ott/linkutil.htm). $P < 0.05$ was considered as significant deviation from the HWE When $P < 0.05$ was detected, the HWE was calculated separately for cases and controls.

Haplotype Analysis. Haplotype analyses testing two-, three-, or four-marker haplotypes for significant difference in haplotype frequencies between cases and controls were done on 27 markers that were in LD with breast cancer (21 significant and 2 borderline markers from the first-stage scan and 4 significant from the second-stage scan) and 44 additional (second-stage) markers next to them in altogether 25 chromosomal regions. All possible combinations, including the associated marker with one, two, or three of its adjacent markers (altogether 76 marker combinations), were tested. Haplotype frequencies were computed using the "haplo.score" software (33). Haplo.score uses unphased genotype data, and via the Expectation Maximization algorithm (34), it estimates the haplotype frequencies. This software implements a global score test for haplotype frequency differences between cases and controls. Moreover, the program provides haplotype-specific tests, which allow evaluation of all individual haplotypes when the global score test is significant. To improve the χ^2 approximation of the score test, we used a threshold level of 0.005 and omitted haplotypes with frequency of <0.5%. Marker combinations where the global score test for differences in haplotype frequencies between cases and controls reached significance (at $P = 0.05$ level) were investigated further by simulations (10,000). This involved simulating the case-control status of the individuals in the study and performing the test each time. The simulations were carried out using haplo.score. In all statistical tests, *P*s <0.05 were considered significant, and all *P*s are unadjusted for multiple comparisons.

Results

Allele Frequencies. In the first-stage scan, we detected 21 markers on 16 chromosomes to have difference in allele frequencies between cases and controls (*P*s = 0.003-0.049), indicating that these markers are in LD with breast cancer. We also detected two markers associated with breast cancer with borderline significance (*P*s = 0.057 and 0.054; Table 1). On 5 of the 16 chromosomes (1-3, 17, and 19), more than one marker was in LD with breast cancer in the first-stage scan, but none of these markers were adjacent. We did not detect any markers associated with breast cancer on 6 of the 22 chromosomes. The presented *P*s are unadjusted for multiple comparisons, although it is unclear how the adjusting for multiple comparisons should be best done in genome-wide screenings. None of the *P*s reaches 0.05/366, which would be the statistical significance level according to conservative Bonferroni correction. We selected 23 loci (21 with the most significant *P*s and 2 with borderline significance) on 16 chromosomes for the second-stage scan where altogether 69 additional flanking markers were analyzed. At 4 of these 69 markers, allele frequencies differed between cases and controls (*P*s = 0.013-0.046, Fisher's exact test; Table 1). We further carried out single-allele tests using 25 of the 27 markers (25 significant and 2 near-significant) and testing the alleles from markers where the difference between cases and controls in the count of the tested allele was ≥10. We tested 39 alleles of 25 markers and found a significant association with 28 alleles, a borderline significance with 2 tested alleles and 8 were not significant (Table 1). The odds ratios of the associated alleles indicated 9 protective (odds ratio, 0.13-0.47) and 21 risk (odds ratio, 1.84-12.86) alleles (Table 1).

Genotype Frequencies and HWE. The genotype frequencies between cases and controls differed significantly with six markers: *D3S3053*, *D7S1818*, *D19S178*, *D19S254*, *D22S1177*, and *D22S445* (data not shown). A significant deviation ($P < 0.05$) from HWE was detected with three markers in the combined set. This effect is most pronounced for marker *D1S179* where cases, controls, and the combined set deviate from HWE (Table 1). The deviation at marker *D1S179* is probably due to the large number of alleles (15). Of the other

Table 1. Associated markers and alleles in the first-stage and second-stage scans

Marker	Location (cM)*	No. different alleles and genotypes observed		*P*		Associated allele size (bp)	Statistics for the associated allele	
		Alleles	Genotypes	HWE[†]	Allele frequency difference[‡]		*P*[§]	Odds ratio (95% confidence interval)
ATA29C07	247.23	8	20	0.982	0.014	257	0.054	1.836 (1.019-3.308)
D1S179	252.12	15	27	0.000	0.038	152	0.002	12.855 (1.625-101.672)
D2S410	125.18	8	21	0.419	0.013	160	0.002	2.978 (1.471-6.029)
						165	0.018	3.079 (1.221-7.762)
D2S1384	200.43	6	14	0.847	0.039	140	0.131	0.579 (0.295-1.133)
						144	0.061	1.860 (0.998-3.466)
D2S434	215.78	6	17	0.795	0.016	268	0.003	0.376 (0.198-0.715)
D2S338	250.54	10	26	0.602	0.046	NA	NA	NA
D3S3050	14.16	5	11	0.455	0.034	234	0.096	1.712 (0.940-3.120)
D3S3053	181.87	6	12	0.080	0.004	223	0.010	2.420 (1.253-4.674)
						227	0.002	4.712 (1.681-13.212)
D4S2394	128.08	9	24	0.851	0.029	237	0.135	1.619 (0.897-2.923)
						252	0.015	2.313 (1.195-4.475)
D7S1818	69.56	6	13	0.813	0.033	185	0.004	0.424 (0.237-0.758)
D8S592	125.27	6	12	0.600	0.045	150	0.007	0.450 (0.254-0.799)
D9S925	32.24	10	30	0.854	0.039	182	0.021	0.446 (0.228-0.872)
D10S2327	100.92	7	13	0.645	0.041	200	0.182	1.535 (0.855-2.755)
						220	0.036	2.342 (1.067-5.139)
D11S1986	105.74	14	41	0.167	0.023	192	0.048	2.308 (1.019-5.226)
						223	0.019	2.989 (1.188-7.522)
						235	0.043	0.404 (0.174-0.939)
D13S796	93.52	7	19	0.299	0.057	149	0.047	2.010 (1.032-3.916)
						162	0.025	0.400 (0.184-0.871)
D14S749	108.22	7	17	0.763	0.003	167	0.120	1.667 (0.897-3.098)
						171	0.052	1.850 (1.022-3.347)
						175	0.007	3.421 (1.373-8.525)
D16S422	111.12	13	40	1	0.012	206	0.019	3.434 (1.197-9.853)
D17S969	27.75	5	9	0.457	0.049	125	0.007	2.221 (1.251-3.944)
						129	0.109	1.725 (0.910-3.272)
D17S809	74.45	9	26	0.013	0.010	250	0.001	0.468 (0.402-0.545)
D19S178	68.08	12	31	0.947	0.025	173	0.004	3.279 (1.437-7.485)
D19S254	100.61	9	19	0.889	0.034	132	0.030	1.941 (1.095-3.442)
						136	0.017	3.532 (1.229-10.149)
D20S109	74.47	15	46	0.071	0.054	NA	NA	NA
D20S887[¶]	72.27	9	29	0.198	0.041	249	0.075	2.154 (0.949-4.887)
D22S426[¶]	36.07	6	13	0.006	0.013	217	0.0002	6.545 (2.156-19.871)
D22S1177[¶]	37.47	8	25	0.379	0.018	180	0.034	2.218 (1.078-4.563)
						182	0.003	0.128 (0.028-0.579)
D22S445	39.87	6	15	0.135	0.022	103	0.006	0.298 (0.126-0.703)
						115	0.120	1.571 (0.895-2.756)
D22S1142[¶]	30.88	9	27	0.778	0.046	183	0.030	2.788 (1.101-7.062)

*Genetic distance (sex average) in cM from p telomere (http://research.marshfieldclinic.org/genetics/Physical_Maps/).

[†] *P* from χ^2 test (Associate program) for cases and controls combined.

[‡] *P* from Fisher's exact test for the difference in allele frequencies cases versus controls.

[§] *P* from Fisher's exact test for associated allele versus all other alleles pooled together.

[¶] Markers in the second-stage scan.

two markers, *D17S809* had significant deviation in cases, and at *D22S426*, the controls are slightly off from equilibrium (P = 0.03). If the HWE Ps were adjusted for multiple comparisons, then *D17S809* and *D22S426* combined sets and *D22S426* controls would not deviate from HWE.

Haplotype Analysis. Haplotype analysis testing two-, three-, or four-marker haplotypes for significant difference between cases and controls was done on 76 marker combinations. We found four marker combinations with a global test score of ≤0.05 and simulated Ps of 0.003 to 0.021 (*D3S3630-D3S1050*, *D11S1986-D11S1347*, *D22S426-D22S1177*, and *D22S1177-D22S445*) and a further three with near significant global score (Table 2). The number of observed haplotypes with frequency of >0.5% with these 4 of 76 marker combinations with global score $P \leq 0.05$ varied from 22 to 45 (Table 2). Among these altogether 119 individual haplotypes, we found four haplotypes significantly more frequent in cases and seven haplotypes significantly more frequent in controls (Ps = 0.007-0.038 and simulated Ps = 0.004-0.030; Table 2). Table 2 shows the frequencies of significant individual haplotypes (with >0.5% frequency), and it should be noted that in some instances the frequency is quite small. However, in all but one of the shown haplotypes, the frequency is >5% in either cases or controls.

Discussion

We have conducted a microsatellite scan of the autosomes for LD with breast cancer in an eastern Finnish population. In the first-stage scan, we found 21 markers on 21 different chromosomal regions in LD with breast cancer. In the follow-up scan, four more markers were in LD with breast cancer, resulting in two regions showing enhanced evidence of existing genetic breast cancer risk factors by association with adjacent markers. We conducted haplotype analysis with markers from 25 chromosomal regions, and it suggests that we have three positive candidate regions in this two-stage autosome-wide scan. We propose that three chromosomal regions, on 3p26, 11q23 and 22q13.1, are associated with breast cancer in our study population. On chromosomes 3 and 11, two markers restrict these regions to 600 and 900 kb, respectively. On chromosome 22, three markers span an area of 550 kb and divide it to regions of 250 and 300 kb. In this region on 22q13, we observed two separate haplotypes, but it could also be one associated 550-kb region.

Testing 366 markers at the 0.05 significance level is predicted to give 18 positive associations by chance alone (366 × 0.05). Because we found 21 positives in the first stage, it is reasonable to propose that some of these represent true associations. Haplotype analysis using global score method and markers in and around associated chromosomal sites revealed four positive regions, which fits well with this assumption. Estimating the number of possible false-positive results in the second-stage scan is more complicated because the markers were selected next to the markers that are in LD with breast cancer and therefore are not independent. The Bonferroni correction was not used here, as it is very conservative and we view these results as a guide for further investigation and not statistically definite significant associations. Thus, we wanted to follow-up all possible associations in the second-stage scan. In addition, it is unclear how the adjusting for multiple comparisons should be best done in genome-wide screenings (8), and a method that identifies wrong positives has not been introduced. When studying complex diseases, it is expected that there will be multiple loci affecting the trait, and a very strict control of the number of false positives (such as the Bonferroni method) will result in a considerable loss of power to identify secondary signals (8). We are aware that all possible breast cancer–associated loci may not have been recognized in our screen due to the longer distance between markers in some chromosomal regions and a limited sample set. In addition, at the time when this scan was initiated, the exact physical

Table 2. Significant and near significant haplotypes

Marker combination	Distance (cM)[*]	Distance (kb)[†]	Global score P[‡]	Simulated P[§]	No. observed haplotypes[‖]	Statistics for significant haplotypes Frequency					
						Haplotype[¶]	Overall[**]	Cases[††]	Controls[‡‡]	P[§§]	Simulated P[‖‖]
D3S3630-D3S1050	3.76	560	0.012	0.016	22	182-227	0.037	0	0.085	0.017	0.014
						182-234	0.046	0.085	0	0.019	0.018
						186-223	0.018	0	0.058	0.035	0.018
D11S1986-D11S1347	0	908	0.046	0.003	45	235-177	0.095	0.049	0.128	0.033	0.030
						227-177	0.071	0.113	0.022	0.034	0.030
D22S426-D22S1177	1.4	240	0.055	0.021	24	215-182	0.050	0.011	0.127	0.005	0.004
						213-186	0.046	0	0.081	0.027	0.030
						211-180	0.033	0.069	0	0.005	0.005
D22S1177-D22S445	2.4	303	0.058	0.018	28	182-115	0.057	0.020	0.115	0.009	0.006
						188-103	0.013	0	0.016	0.034	0.014
						180-115	0.084	0.160	0.014	0.007	0.004
D2S436-D2S410	7.02	9,471	0.077	NA	32	184-160	0.153	NA	NA	0.005	NA
D2S410-D2S363	0	1,154	0.095	NA	29	169-252	0.051	NA	NA	0.050	NA
						171-254	0.036	NA	NA	0.047	NA
						160-252	0.083	NA	NA	0.046	NA
D3S3725-D3S3053	0	60	0.068	NA	32	92-227	0.028	NA	NA	0.027	NA
						88-223	0.024	NA	NA	0.040	NA

[*] Genetic distance (sex average) between markers in cM (Marshfield).

[†] Distance between markers in kb (Marshfield).

[‡] P for global test score for the difference in haplotype frequencies cases versus controls from haplo.score.

[§] Simulated P for the global score from haplo.score.

[‖] Number of haplotypes observed with frequency >0.5%.

[¶] Individual haplotypes (allele size in bp) with significant difference in allele frequencies between cases and controls.

[**] Frequency for given haplotype in cases and controls combined.

[††] Frequency for given haplotype in cases.

[‡‡] Frequency for given haplotype in controls.

[§§] P for frequency difference of individual haplotype cases versus controls from haplo.score.

[‖‖] Simulated P for frequency difference of individual haplotype cases versus controls from haplo.score.

locations of some markers were not available. As a consequence, some markers in the second-stage scan are not as close to markers that were in LD in the first-stage scan as initially expected. This, in turn, could lead to false-negative results. Therefore, a denser marker map and a bigger sample set could probably have led to an improved result of the scan. Of the 366 markers in the first-stage scan, 59% were spaced by ≤10 cM, 30% by 11 to 15 cM, and 11% were spaced by 16 to 20 cM. Less than half of the screened chromosomes harbored one or two gaps of 18 to 20 cM. However, in this ~20-generation-old founder population, relatively few meioses have occurred and LD should extend beyond the distances between adjacent markers in the first-stage and second-stage scans in general (8). In a subisolate of Kuusamo in the late-settlement region, up to 50% of microsatellites showed LD at distances 3.5 to 7.5 cM and 30% at distances >7.5 cM (35). Our study population was selected from a small rural area in the late-settlement region and could possess the genetic characteristics more similar to the Kuusamo population than the late-settlement population in general. It is also worth noting that with a 10 cM map LD between marker and risk allele needs to extend 5 cM, as our scan is not for LD between markers but between marker and risk allele that lies somewhere between two markers spaced by 10 cM in average.

It has been estimated that, in single nucleotide polymorphism (SNP) association studies of complex disease, increasing the sample size is more prudent than increasing the number of SNPs (36). Although a lot of discussion and estimations of the sample size and number of SNPs needed have been published (e.g., refs. 37, 38), to our knowledge thus far, only one genome-wide SNP scan has been conducted using 28,000 SNPs in a set of 272 breast cancer cases and 276 controls (39). In a study comparing the extent of LD using SNP and microsatellite markers in Finnish populations, the results showed that a single microsatellite was more informative than the combined information from 3 to 5 SNPs (35). Microsatellites also detect LD over longer distances than SNPs (8). However, it is difficult to estimate accurately how sample size or marker density improves a genome-wide microsatellite scan, as studies on comparisons of changing marker density or sample size in microsatellite scans have not been published. We used Fisher's exact test with Monte Carlo approximation to handle the limited capacity of the used sample size. In addition, simulated *P*s were calculated for haplotype estimations.

Our scan was conducted to gain knowledge of possible genetic breast cancer risk factors on population basis. We selected the cases who did not have a strong family history of breast cancer from a small geographic region for optimal homogeneous gene pool and as small effective number of unrelated founders as possible (37). We searched for low-penetrance risk factors for breast cancer that do not show up clearly in families. In addition, when the cases are a homogeneous group from small geographic region, the subset of genetic risk factors they carry could be also homogeneous and the number of them is reduced. The genetic diversity of this population is believed to be lower than that of the Finnish population in general, and this feature has been the basis for mapping several genes and even genes for complex diseases (40). The case would be different if we had a more heterogeneous population. When this work was initiated in 1996, it was not recognized as it is according to the current knowledge that it could be useful to have family material even in association studies (41). Due to the nature of our population, this study falls between traditional LD study and family linkage study. Being a young isolated population, this group of eastern Finns could be seen as a sample set between a family set and a group of nonrelatives. However, the rate of genetic relationship is low, as the sample population is consistent with HWE.

BRCA1 and *BRCA2* genes are known high-risk genes for familial breast cancer, and recently, it has been proposed that *BRCA1* might have a tumor suppressor function also in sporadic breast cancer (42, 43). Polymorphic variants in high-risk breast cancer susceptibility genes (e.g., *BRCA* genes) are also candidates for low-penetrance alleles that alter breast cancer risk in population. Indeed, *BRCA2* gene N372H polymorphism has been reported to confer 1.31-fold risk of breast cancer (44). This study included a sample population also from eastern Finland. Interestingly, the N372H-associated breast cancer risk among the studied populations was lowest in the Finns. In our scan, we did not detect LD with breast cancer at a *BRCA1* intragenic (intron 20) marker, *D17S855* (*P* = 0.480), or at the markers flanking *BRCA1* (*D17S1299* and *D17S579*) that locate 2.2 Mb (~1 cM) and 1.6 Mb (~0.6 cM) from *BRCA1*, respectively (*P*s = 0.225 and 0.330). On chromosome 13, we analyzed markers flanking *BRCA2* gene (*D13S260* and *D13S1493*) and neither of them showed LD with breast cancer (*P*s = 0.290 and 0.955, respectively). These markers are 1.6 Mb (2.15 cM) apart and locate 450 kb and 1 Mb from *BRCA2*. These results could arise due to small sample size, but they may also indicate the fact that low-penetrance *BRCA* variants are not involved in our set of breast cancer cases. The core haplotypes of different *BRCA1* and *BRCA2* mutations detected in Finland extend 1.6 to 15.5 and 2.7 to 3.2 cM, respectively (27). This suggests that LD around these genes extends far enough for us to detect LD with breast cancer by using microsatellites within those distances from the genes if common low-penetrance *BRCA* variants were strongly involved.

We have identified three potential candidate regions for genetic breast cancer risk factors. These regions include both known and unknown or predicted genes. Moreover, some of the genes are proposed to be cancer related and some of them could be candidates for breast cancer–associated genes based on the proposed function of the gene product. On 3p26, there are three genes, two of which are not very well characterized or properly named. Evidence from loss of heterozygosity studies suggests that a few candidate tumor suppressor genes lie on chromosome 3p (45) but none of them locates in the 600-kb region of the significant haplotype found in our scan. On 11q23, the 900-kb haplotype region includes 13 named genes and 4 hypothetical genes. Chromosome 11q21-24 region is also known for loss of heterozygosity in several different cancers, including breast cancer. In the 550-kb region of the two significant haplotypes, 22q13.1 locates seven named genes, two noncharacterized and two hypothetical genes. The results for the three-marker haplotype *D22S426-D22S1177-D22S445* was not significant (global score, *P* = 0.19). This could be due to dividing of the risk to several haplotypes of which none is significantly associated alone, as with this marker combination we observed 51 haplotypes with frequency of >0.5%. Therefore, we cannot exclude the possibility of two separate regions and risk factors on 22q13.

We conclude that in the two-stage autosome-wide scan of 435 autosomal microsatellite markers we detected 25 markers in LD with breast cancer (with unadjusted *P*s < 0.05) and breast cancer association with two or more adjacent markers in two chromosomal regions. Haplotype analysis identified three regions, on 3p26, 11q23, and 22q13.1, with difference in haplotype frequencies between cases and controls with four marker combinations. Based on previous publications of genes in these regions, a candidate for breast cancer risk-altering gene may lie in one of these regions. The third stage in this scan will be the follow-up investigation and thorough analysis of the four regions picked up in the haplotype analysis, confirming the associations and identifying the risk-altering genetic change.

References

1. Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. Science 1990;250:1684–9.
2. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266: 66–71.

3. Wooster R, Neuhausen SL, Mangion J, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science 1994;265:2088–90.
4. Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 1995;378:789–92.
5. Peto J, Collins N, Barfoot R, et al. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. J Natl Cancer Inst 1999;91:943–9.
6. Antoniou AC, Pharoah PDP, McMullan G, et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br J Cancer 2002;86:76–83.
7. Pharoah PDP, Day NE, Duffy S, Easton DF, Ponder BAJ. Family history and the risk of breast cancer: a systematic review and meta-analysis. Int J Cancer 1997;71:800–9.
8. Ophoff RA, Escamilla MA, Service SK, et al. Genomewide linkage disequilibrium mapping of severe bipolar disorder in a population isolate. Am J Hum Genet 2002;71:565–74.
9. Lynn A, Kashuk C, Petersen MB, et al. Patterns of meiotic recombination on the long arm of human chromosome 21. Genome Res 2000;10:1319–32.
10. Järvelä I. Infantile neuronal ceroid lipofuscinosis (CLN1): linkage disequilibrium in the Finnish population and evidence that variant late infantile form (variant CLN2) represents a nonallelic locus. Genomics 1991;10:333–7.
11. Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 1992;2:204–11.
12. Kere J, Sistonen P, Holmberg C, de la Chapelle A. The gene for congenital chloride diarrhea maps close to but is distinct from the gene for cystic fibrosis transmembrane conductance regulator. Proc Natl Acad Sci 1993;90:10686–9.
13. Höglund P, Sistonen P, Norio R, et al. Fine mapping of congenital chloride diarrhea gene by linkage disequilibrium. Am J Hum Genet 1995;57:95–102.
14. Varilo T, Savukoski M, Norio R, Santavuori P, Peltonen L, Järvelä I. The age of human mutation: genealogical and linkage disequilibrium analysis of the CLN5 mutation in the Finnish population. Am J Hum Genet 1996;58:506–12.
15. Visapää I, Salonen R, Varilo T, Paavola P, Peltonen L. Assignment of the locus for Hydrolethalus syndrome to a highly restricted region on 11q23-25. Am J Hum Genet 1999;65:1086–95.
16. Toda T, Miyake M, Kobayashi K, et al. Linkage-disequilibrium mapping narrows the Fukuyama-type congenital muscular dystrophy (FCMD) candidate region to <100 kb. Am J Hum Genet 1996; 59:1313–20.
17. Snarey A, Thomas S, Schneider MC, et al. Linkage disequilibrium in the region of the autosomal dominant polycystic kidney disease gene (PKD1). Am J Hum Genet 1994;55:365–71.
18. Lee N, Daly MJ, Delmonte T, et al. A genomewide linkage-disequilibrium scan localizes the Saguenay-Lac-Saint-Jean cytochrome oxidase deficiency to 2p16. Am J Hum Genet 2001; 68:397–409.
19. Nevanlinna HR. The Finnish population structure. A genetic and genealogical study. Hereditas 1972;71:195–236.
20. Norio R, Nevanlinna HR, Perheentupa J. Hereditary diseases in Finland; rare flora in rare soul. Ann Clin Res 1973;5:109–41.
21. Workman PL, Mielke JH, Nevanlinna HR. The genetic structure of Finland. Am J Phys Anthropol 1976;44:341–67.
22. Norio R. Finnish Disease Heritage II: population prehistory and genetic roots of Finns. Hum Genet 2003;112:457–69.
23. Soininen AM. Pohjois-Savon asuttaminen keski-ja uuden ajan vaihteessa. In: Historiallisia tutkimuksia LVIII. Helsinki: Suomen Historiallinen Seura; 1981.
24. Pirinen K. Rajamaakunta asutusliikkeen aikakautena. In: Savon historia II. Pieksämäki: Kustannuskiila OY; 1982. p. 1534–617.
25. Nyström-Lahti M, Sistonen P, Mecklin JP, et al. Close linkage to chromosome 3p and conservation of ancestral founding haplotype in hereditary nonpolyposis colorectal cancer families. Proc Natl Acad Sci 1994;91:6054–8.
26. Moisio AL, Sistonen P, Weissenbach J, de la Chapelle A, Peltomäki P. Age and origin of two common MLH1 mutations predisposing to hereditary colon cancer. Am J Hum Genet 1996;59:1243–51.
27. Sarantaus L, Huusko P, Eerola H, et al. Multiple founder effects and geographical clustering of BRCA1 and BRCA2 families in Finland. Eur J Hum Genet 2000;8:757–63.
28. Vehmanen P, Friedman LS, Eerola H, et al. Low proportion of BRCA1 and BRCA2 mutations in Finnish breast cancer families: evidence for additional susceptibility genes. Hum Mol Genet 1997;6:2309–15.
29. 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.
30. Mitrunen K, Jourenkova N, Kataja V, et al. Steroid metabolism gene CYP17 polymorphism and the development of breast cancer. Cancer Epidemiol Biomarkers Prev 2000;9:1343–8.
31. Vandenplas S, Grobler-Rabie A, Bredner K, et al. Blot hybridization of genomic DNA. J Med Genet 1984;21:164–72.
32. Sheffield VC, Weber JL, Buetow KH, et al. A collection of tri- and tetranucleotide repeat markers used to generate high quality, high resolution human genome-wide linkage maps. Hum Mol Genet 1995;4:413–9.
33. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association of traits with haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425–34.
34. Slatkin M, Excoffier L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity 1996;76:377–83.
35. Varilo T, Paunio T, Parker A, Perola M, Meyer J, Terwilliger JD, Peltonen L. The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. Hum Mol Genet 2003;12:51–9.
36. Huang Q, Fu Y, Boerwinkle E. Comparison of strategies for selecting single nucleotide polymorphisms for case/control association studies. Hum Genet 2003;113:253–7.
37. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 1999;22:139–44.
38. Abecasis GR, Noguchi E, Heinzmann A, et al. Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 2001;68:191–7.
39. Kammerer S, Roth R, Nelson MR, et al. Identification of susceptibility genes for sporadic breast cancer in a genome-wide association study. ASHG Annual Meeting, poster 381, 2003.
40. Ekelund J, Lichtermann D, Hovatta I, et al. Genome-wide scan for schizophrenia in the Finnish population: evidence for a locus on chromosome 7q22. Hum Mol Genet 2000;9:1049–57.
41. Antoniou AC, Easton DF. Polygenic inheritance of breast cancer: implications for design of association studies. Genet Epidemiol 2003;25:190–202.
42. Bianco T, Chenevix-Trench G, Walsh DC, Cooper JE, Dobrovic A. Tumour-specific distribution of BRCA1 promoter region methylation supports a pathogenetic role in breast and ovarian cancer. Carcinogenesis 2000;21:147–51.
43. Staff S, Isola J, Tanner M. Haplo-insufficiency of BRCA1 in sporadic breast cancer. Cancer Res 2003;63:4978–83.
44. Healey CS, Dunning AM, Teare DM, et al. A common variant in BRCA2 is associated with both breast cancer risk and prenatal viability. Nat Genet 2000;26:362–4.
45. Yang Q, Yoshimura G, Mori I, Sakurai T, Kakudo K. Chromosome 3p and breast cancer. J Hum Genet 2002;47:453–9.

An Autosome-Wide Scan for Linkage Disequilibrium–Based Association in Sporadic Breast Cancer Cases in Eastern Finland: Three Candidate Regions Found

Jaana M. Hartikainen, Hanna Tuhkanen, Vesa Kataja, et al.

Cancer Epidemiol Biomarkers Prev 2005;14:75-80.