Published in final edited form as:
Nat Genet. 2009 May ; 41(5): 579–584. doi:10.1038/ng.353.

A multi-stage genome-wide association in breast cancer identifies two novel risk alleles at 1p11.2 and 14q24.1 (*RAD51L1*)

Gilles Thomas[1], **Kevin B. Jacobs**[1,2,3], **Peter Kraft**[4], **Meredith Yeager**[1,3], **Sholom Wacholder**[1], **David G. Cox**[4,5], **Susan E. Hankinson**[5], **Amy Hutchinson**[1,3], **Zhaoming Wang**[1,3], **Kai Yu**[1], **Nilanjan Chatterjee**[1], **Montserrat Garcia-Closas**[1], **Jesus Gonzalez-Bosquet**[1], **Ludmila Prokunina-Olsson**[1], **Nick Orr**[1], **Walter C. Willett**[5,6], **Graham A. Colditz**[7], **Regina G. Ziegler**[1], **Christine D. Berg**[8], **Saundra S. Buys**[9], **Catherine A. McCarty**[10], **Heather Spencer Feigelson**[11], **Eugenia E. Calle**[11], **Michael J. Thun**[11], **Ryan Diver**[11], **Ross Prentice**[12], **Rebecca Jackson**[13], **Charles Kooperberg**[12], **Rowan Chlebowski**[14], **Jolanta Lissowska**[15], **Beata Peplonska**[16], **Louise A. Brinton**[1], **Alice Sigurdson**[1], **Michele Doody**[1], **Parveen Bhatti**[1], **Bruce H. Alexander**[17], **Julie Buring**[18], **I-Min Lee**[18], **Lars J Vatten**[19], **Kristian Hveem**[19], **Merethe Kumle**[20], **Richard B. Hayes**[1], **Margaret Tucker**[1], **Daniela S. Gerhard**[21], **Joseph F. Fraumeni Jr.**[1], **Robert N. Hoover**[1], **Stephen J Chanock**[1], and **David J. Hunter**[1,4,5,6,22]

[1] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20892 [2] Bioinformed Consulting Services, Gaithersburg, MD 20877 [3] Core Genotyping Facility, Advanced Technology Program, SAIC-Frederick Inc., NCI-Frederick, Frederick, MD 21701 [4] Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115 [5] Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115 [6] Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts [7] Washington University School of Medicine, St. Louis, MO 63110 [8] Division of Cancer Prevention, NCI, NIH, DHHS, Bethesda, MD 20892 [9] Department of Internal Medicine, University of Utah, Salt Lake City, UT 84132 [10] The Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI 54449 [11] Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA 30329 [12] Fred Hutchinson Cancer Research Center, Seattle, WA 98195 [13] Division of Diabetes, Endocrinology and Metabolism, The Ohio State University Medical Center, Columbus, OH 43210 [14] Harbor-University of California at Los Angeles Medical Center, Torrance, CA 90509 [15] Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland [16] Nofer Institute of Occupational Medicine, Łódź, Poland [17] Division of Environmental Health Science, School of Public Health, University of Minnesota, Minneapolis, MN 55455 [18] Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115 [19] Department of Public Health, Norwegian University of Science and Technology, Trondheim, Norway [20] Institute of Community Medicine, University of Tromso, Tromso, Norway [21] Office of Cancer Genomics, NCI, NIH, DHHS Bethesda, MD 20892 [22] Broad Institute of Harvard and MIT, Cambridge, MA 02142

Abstract

The Cancer Genetic Markers of Susceptibility (CGEMS) initiative has conducted a three-stage genome-wide association study (GWAS) of breast cancer in 9,770 cases and 10,799 controls. In

[*] Correspondence to: David J. Hunter, Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, Telephone: 617-432-2252, Fax: 617-432-1722, dhunter@hsph.harvard.edu.

Stage 1, we genotyped 528,173 single nucleotide polymorphisms (SNPs) in 1,145 cases of invasive breast cancer among postmenopausal white women, and 1,142 controls; in Stage 2, 24,909 SNPs with low p values observed in Stage 1 were analyzed in 4,547 cases and 4,434 controls. In Stage 3 we investigated 21 loci in 4,078 cases and 5,223 controls with low p values from Stage 1 and 2 combined. Two novel loci achieved genome-wide significance. A pericentromeric SNP on chromosome 1p11.2, rs11249433, ($p=6.74 \times 10^{-10}$ adjusted genotype test with 2 degrees of freedom) resides in a large block of linkage disequilibrium neighboring *NOTCH2* and *FCGR1B* and is predominantly associated with estrogen receptor-positive breast cancer. A second SNP, rs999737 on chromosome 14q24.1 ($p=1.74 \times 10^{-7}$), localizes to *RAD51L1*, a gene in the homologous recombination DNA repair pathway, a prior candidate pathway for breast cancer susceptibility. We confirmed previously reported markers on chromosome 2q35, 5q11.2, 5p12, 8q24, 10q26, and 16q12.1. Our results underscore the importance of large-scale replication in the identification of low penetrance breast cancer alleles.

Epidemiologic investigation of breast cancer has identified a number of environmental and lifestyle risk factors (e.g., age at menarche and menopause, parity, age at first birth, body mass index and exogenous hormone use)[1]. Breast cancer is nearly twice as frequent in first degree-relatives of women with the disease than in relatives of women without such a history, suggesting an important contribution of inherited susceptibility. Established causal variants from before the GWAS era account for only a small fraction of sporadic breast cancers. Established associations include high penetrance germline mutations segregating in high-risk pedigrees, most notably in *BRCA1* and *BRCA2*[2,3]; a handful of rare susceptibility variants with lower penetrance identified in DNA repair and apoptosis genes[4–8]; only one locus with a minor allele frequency larger than 5% (*CASP8*) was found using the candidate gene approach in association studies[9].

Genome-wide association studies have identified multiple new common genetic variants influencing breast cancer risk. Easton et al. analyzed genotypes from 390 cases enriched for a strong family history of breast cancer and 364 controls with 227,876 SNPs and followed the top 10,405 SNPs in a two-stage replication study (primarily conducted in population-based studies of unrelated subjects), resulting in the identification of 5 loci (10q26 *(FGFR2)*, 16q12.1 *(TNRC9)*, 5q11.2 *(MAP3K1)*, 8q24 and 11p15.5 (*LSP1*)) based on large-scale follow-up studies[10]. In the initial report from the NCI Cancer Genetic Markers of Susceptibility (CGEMS) initiative, based on a follow-up of the top ten SNPs from the Stage 1 GWAS, we independently identified SNPs in intron 2 of *FGFR2* as associated with breast cancer at genome-wide significant levels[11]. Subsequently, the *FGFR2* locus was also identified in an Icelandic population[12] and a locus at 2q35 was also reported to confer susceptibility to estrogen receptor [ER] positive breast cancer[12]. Finally, combined analysis of a promising signal using the three published GWAS led to the identification of an additional locus on 5p12[13]. Power calculations based on the available sample sizes (390–1,791 cases) in the three GWAS efforts, suggest each has limited power to detect the low observed relative risks (RRs of 1.1–1.3 per allele) at conventional levels of genome-wide significance ($p < 5 \times 10^{-7}$)[14]. Thus, it is likely that a high proportion of the susceptibility loci have not yet been detected.

In Stage 1 of CGEMS, we genotyped 1,145 cases post-menopausal women of European ancestry with invasive breast cancer and 1,142 matched controls nested within the prospective Nurses' Health Study cohort[11]. This stage used 528,173 SNPs that were estimated to be correlated with an r2>0.8 to approximately 90% of the common HapMap Phase II SNPs. We report here a follow-up of this first stage. In Stage 2, we attempted to genotype 30,448 SNPs in 4,547 cases and 4,434 controls from four different studies (Table 1). These SNPs were selected using a stepwise procedure (Supplementary Methods); the majority were chosen by an hypothesis-free (agnostic) strategy while approximately one fifth of the SNPs were selected by alternative approaches fully reported in the supplementary methods and described below.

Briefly, for Stage 2, 22,136 SNPs were first selected based on a p-value less than 0.05 in a logistic regression model using a two-degree of freedom (df) score test with indicator variables for heterozygous and homozygous carriers and four continuous variables representing principal components of population stratification. The 2-df score test was chosen because it makes minimal assumptions for the underlying genetic model. This set of SNPs was complemented with 2,773 SNPs with a p-value less than 0.06 in tests of dominant, recessive or multiplicative models that were not already included by virtue of their p-value in the score test (each test has 1 df - see Supplementary Methods). In the 'agnostic' category, SNPs with low p-values in strong linkage disequilibrium ($r^2 \geq 0.8$) were removed. We selected an additional 1,436 'agnostic' SNPs not included in the two previous criteria based on a 2-SNP test that conditioned each SNP on a neighboring SNP, if this improved the p-value relative to the single SNP-statistics by an order of magnitude. Loci marked by SNPs previously established by GWAS were further explored with a dense set of 1,711 SNPs. Also included were 3,788 SNPs drawn from candidate genes in previously proposed pathways or identified in an analysis of suggested interaction with variants in intron 2 of the *FGFR2* gene. Finally, to monitor population stratification, 1,508 SNPs with low pair-wise linkage disequilibrium were included[15].

A total of 30,278 SNPs (92.1%) provided reliable genotypes according to our quality control metrics (see Supplemental Methods). We removed subjects with greater than 20% admixture of non-European origin based on analysis using the STRUCTURE program[16]. We conducted a principal component analysis (PCA) using the SNPs chosen to monitor population stratification and there was minimal evidence of population stratification observed between cases and controls; the distribution of the p-values for the association statistics with a 2 degree-of-freedom test unadjusted for population heterogeneity was close to the expected distribution under the null hypothesis[17]. The inflation factor, λ, 1.010 was reduced to 1.009 when the first four principal components were included as covariates in the association test. A joint analysis of the genotypes[18] in the first and second stages was performed using an age, study design and population stratification-adjusted multinomial regression analysis (2 df test).

In the combined analysis of the initial scan with the second stage, we note that markers in 6 of the reported 7 loci identified in prior GWAS studies were strongly associated with breast cancer risk in post-menopausal women (Table 2). SNPs in 2q35, 5q11.2 (*MAP3K1*), 5p12, 8q24, 10q26 (*FGFR2*) and 16q12.1 (*TOX3/TNRC9*) provided strong signals (Table 2 and Supplemental Table 1); in some cases, an alternative SNP to the originally reported SNP provided a smaller p value (see below). The lowest p value for a marker at 11p15.5 (*LSP1*, rs3817198) was minimally significant ($p= 3.87 \times 10^{-2}$, trend test with 1 df- see Supplemental Table 1) but its allele-specific odd ratio was similar to that reported previously (heterozygote odds ratio [OR] 1.04; 95% CI 1.00 to 1.09; homozygote OR 1.09; 95% CI 1.00–1.19 in our combined three-stage analysis. For the single candidate gene variant that had previously been reported as genome-wide significant, the results for rs1045485 in *CASP8* ($p=5.47 \times 10^{-2}$, trend test with 1 df) were also consistent with previous findings (heterozygote OR 0.96; CI 95% 0.91–1.00; homozygote OR 0.92; CI 95% 0.84–1.00). After Stage 2, no indication of association ($p_{2df}=0.50$) was observed for rs2107425 in the *H19* region, previously associated at lower level of significance by Easton et al. (reported $p_{trend}=2 \times 10^{-5}$)[10]. A GWAS in American Jewish women of Ashkenazi background had identified a locus on chromosome 6 (rs2180341) with a MAF of 0.21 and a per allele OR of 1.41 ($p= 3.0 \text{ x} 10^{-8}$)[19]. In CGEMS, SNP rs9398840, which was strongly correlated with rs2180341 ($r^2=1.0$) in the CEU HapMap population was not significantly associated ($p_{2df}=0.58$) and not taken into Stage 2.

Stage 3 included a set of 24 SNPs, 21 of which were based on a preliminary combined analysis of the first two stages, in 4,078 cases and 5,223 controls drawn from five studies (Tables 1 and 2). Specifically, we examined 16 promising novel regions based on the lowest p values of the preliminary data build with one SNP. Two novel regions were examined with two SNPs apiece. In a region of 3p24.1, two SNPs, rs724244 and 4973768, separated by 170 kb (r^2 =0.35) each had low p values. In region 1p34.2 because of difficulty in the assay design, two SNPs, separated by 40 kb and in strong LD were selected (r^2= 0.88). In the region of the two SNPs in 5p12, in which rs4415084 and rs10941679 were recently reported by Stacey et. al., we advanced two more SNPs, rs7716600 and rs2067980, separated by 100 kb (r^2= 0.50) (Figure 1)[13] Thus, the 5p12 region was explored with four SNPs. For Stage 3, rs3817198 in *LSP1* was also added to the set because of a prior publication[10].

The results of Stage 3 are remarkable for only four SNPs. Two novel SNPs, rs11249433 in the pericentromeric region of chromosome 1, and rs999737 in the candidate gene, RAD51-like 1 gene *(RAD51L1)* on chromosome 14q24.1, reached genome-wide significance in the combined analysis of all three stages (Table 3). Two of the SNPs in 5p12, rs7716600 and rs4415084, confirmed the previously reported signals.

The results of a combined joint adjusted analysis of the initial genome-wide scan plus two stages of follow-up provide conclusive statistical significance for an association with a novel marker, rs11249433 located in the pericentromeric region of the short arm of chromosome 1 ($p = 6.74 \times 10^{-10}$) (Figure 1 and Table 3). Pericentromeric regions are known to be recombination-poor regions and thus it is not surprising to observe that rs11249433 maps to large block of linkage disequilibrium. The definition of the block is difficult to determine for two reasons: (1) its close proximity to the centromere and (2) presence of a SNP desert of approximately 220kb which is immediately distal to the block (Figure 2A). The block contains several pseudogenes, and a member of the highly paralogous low affinity Fc gamma receptor family, *FCGR1B*. Distal to the SNP desert is the promoter of *NOTCH2*, a gene recently shown to be associated with type 2 diabetes[20]. Some epidemiological studies have suggested an association between type 2 diabetes and post-menopausal breast cancer[21]. Further mapping and subsequent functional work is required to provide plausibility for the association signal observed with rs11249433.

The second novel marker, rs999737 is in a gene in prior candidate pathway for breast cancer susceptibility, the double-strand break repair/homologous recombination pathway, *RAD51L1* (also known as *RAD51B*) on chromosome 14q24.1 ($p = 1.74 \times 10^{-7}$) (Table 3). The SNP maps to a 70Kb LD block defined by two recombination hotspots and is entirely contained within intron 12 of the gene (Figure 2B and Supplemental Figure 1). Its gene product is one of five paralogs that interact directly with that of the *RAD51* gene, that catalyzes key reactions in homologous recombination[22]. A polymorphism in the 5'UTR of *RAD51* has recently been identified as a genetic modifier of outcome in women with deleterious *BRCA2* mutations[23]. A copy number variation on chromosome 14q24.1 that includes the *RAD51L1* has been observed repeatedly in pedigrees with Li-Fraumeni syndrome, suggesting a possible contribution of this locus to the spectrum of cancers (that includes breast cancer) observed in this hereditary syndrome[24]. Further work is warranted to dissect the genetic signal and investigate potential functional variants.

Tumor estrogen receptor (ER) status was available for 6,386 cases[25]. Figure 3 shows the results of the analysis for the two novel SNPs, rs11249433 (chromosome 1) and rs999737 (chromosome 14) by estrogen receptor status. The association with rs11249433 is more apparent for ER+ compared to ER− breast cancer (Supplementary Tables 2, 3 and 4). The observed difference was significant in a case/case comparison (trend p value = 0.001), suggesting that the chromosome 1 locus could be more important in ER+ breast cancer susceptibility. Although there was also some evidence for a stronger association with ER+ disease for the chromosome 14 SNP, rs999737, it was not significant (trend p value = 0.20). An analysis stratified by age did not demonstrate any significant differences for the two SNPs, though it should be emphasized that the majority of cases are post-menopausal women.

Given the initial genome coverage of the CGEMS study using the Illumina HumanHap500 platform and the number of cases and controls investigated, it is unlikely that many more common loci with relative risks comparable to *FGFR2* will be discovered for the European population. The present study has confirmed strong association signals for 6 genomic regions previously reported and identified novel associations at genome-wide significance for markers on chromosome 1p11.2 and 14q24.1. In addition, we provide supportive evidence for two loci, previously associated with genome-wide significance, namely, 2p24.1 (*CASP8)* and 11p15.5 *(LSP1)*. Though the direction and magnitude of the association signal is consistent with prior reports, our study indicates that larger data sets are required to identify at genome-wide significance levels loci with smaller estimated per allele effect sizes, especially SNPs with low MAF or for which the per allele OR is estimated to be 1.1 or less. Moreover, our study suggests the value of combining scans for discovery with subsequent follow-up in large data sets, such as CGEMS and Breast Cancer Association Consortium (BCAC)[9–11]. The individual genotype data for the Stage 1 CGEMS GWAS in 1,145 cases and 1,142 controls, and the aggregate data for Stages 1, 2 and 3 are available to researchers registered after approval by the NCI Data Access Committee (DAC) through the CGEMS portal (http://cgems.cancer.gov).

To date, GWAS for breast cancer have been conducted largely among women of European ancestry, mainly with ER+ tumors. Well-designed scans in other populations should yield additional loci, some of which could be population-specific. Additional scans of ER−ve tumors will be needed to find loci specific to this subtype. Together these findings should accelerate the effort to dissect the genetic signals observed in multi-stage GWAS in an effort to nominate variants for further investigation of their biological basis. The evidence for two new associations presented in this study pinpoints genomic regions that could elucidate novel etiologic pathways contributing to the development of breast cancer. Carriage of the multiple loci reported so far, together with additional loci to be identified in follow-up of this and other studies, should refine estimates of the increased risk of sporadic breast cancer associated with inherited genetic loci, although the clinical utility of these estimates has yet to be determined[26,27]

Methods (678)

Initial Genome-wide Scan Genotyping

Briefly, this study reports the follow-up genotyping of studies based on the previously reported genome-wide scan conducted in the prospective Nurses' Health Study using the Human Hap500 Infinium Assay (Illumina) in 1,145 cases of women with post-menopausal breast cancer and 1,142 controls [11]. The details are reported elsewhere[11]. Quality control metrics included removal of samples with call rates under 90% and SNP assays with call rates under 95%. Subjects with more than 15% admixture of non-European background were removed from the analysis.

Replication Samples

In Stage 2, we genotyped 30,278 SNPs in four follow-up studies of women of European background with breast cancer totaling 4,547 cases and 4,434 controls drawn from the American Cancer Society Cancer Prevention Study II, the Prostate, Lung, Colon and Ovarian Screening Trial, part of the available Polish Breast Cancer Study and the observational arm of the Women's Health Initiative. In Stage 3, we genotyped 24 SNPs in 4,078 cases of breast cancer in women of European background and 5,223 controls drawn from the CONOR Norwegian cohort, the remaining cases and controls of the Polish Breast Cancer Study, the U.S. Radiologic Technologists Study, the Nurses' Health Study II, and the Women's Health Study. These studies were approved by the appropriate institutional review boards.

Replication Genotyping

In Stages 2 and 3, we genotyped 18,282 unique subjects (excluding validation samples and study duplicates) passing sample handling quality control metrics in the Core Genotyping Facility of the National Cancer Institute. For NHS II and WHS, the 24 SNPs of Stage 3 were genotyped at the DF/HCC Genotyping Core at the Harvard School of Public Health, Boston, MA. Stage 2 was genotyped using a custom-designed iSelect assay from Illumina with content described above; 9,804 samples were attempted (including known duplicates). Using quality control measures, samples were removed with call rates under 90% and SNPs with call rates under 95%. Fitness for Hardy-Weinberg proportion was assessed for each SNP in unique controls subjects only but was not used to exclude SNP assays (see Supplemental Methods). In Stage 3, we genotyped 9,301 unique subjects for 24 TaqMan assays (ABI) selected on the criteria described above using custom designed assays that were subsequently optimized in the SNP500Cancer initiative.

A small fraction (less than 2%) of subjects who were successfully genotyped in Stage 2 were excluded from analysis due to one of the following reasons: 1. Unanticipated interstudy or intrastudy duplicates; 2. Unanticipated non-European admixture of greater than 20% (e.g., African or East Asian; notably, in Stage 1, the threshold for non-European admixture was 15%); and/or 3. Incomplete covariate data.

In Stage 2, a total of 16,715 discordant genotypes were detected out of a possible 7,255,923 genotype comparisons (237 duplicate pairs and one triplicate) yielding a discordance rate of 0.23%. Infinium cluster plots for notable SNPs are included in Supplemental Methods.

For the 24 SNPs analyzed in Stage 3, we validated genotype calls determined by Infinium HumanHap500 and custom iSelect assay by comparing TaqMan results in the entire Polish Breast Cancer Study. 1,110 samples were genotyped with both platforms and the overall concordance rate was 99.52% (see Supplemental Materials for results).

Analysis

For the follow-up replication studies, all one-SNP analyses were conducted using unconditional logistic regression, adjusted for age in ten year intervals and study. For Stages 1 and 2, four continuous covariates were included to account for population heterogeneity based on principal component analysis of genotype correlations. Separate analyses were conducted according to the individual studies, the pooled replication studies in Stage 2 and Stage 3 and for all studies combined. Genotype effects were modeled individually, and a single-SNP score test with two degrees of freedom was computed. To enable comparison with other published GWAS, a Cochran-Armitage trend test was also performed. To explore a possible difference in effect between estrogen-positive and estrogen-negative breast cancer, separate analyses were conducted for ER+ and ER− cases, using a trend test with 1 degree of freedom..

Informatics

We used GLU (Genotyping Library and Utilities version 1.0), a suite of tools available as an open-source application for management, storage and analysis of GWAS data. STRUCTURE and EIGENSTRAT programs were used to assess population heterogeneity (see URLs below)

URLs:

CGEMS portal: http://cgems.cancer.gov/

CGF: http://cgf.nci.nih.gov/

EIGENSTRAT: http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm

GLU: http://code.google.com/p/glu-genetics/

SNP500Cancer: http://snp500cancer.nci.nih.gov/

STRUCTURE: http://pritch.bsd.uchicago.edu/structure.html

Tagzilla: http://tagzilla.nci.nih.gov/

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Nurses' Health Studies are supported by NIH grants CA 65725, CA87969, CA49449, CA67262, CA50385 and 5UO1CA098233. The authors thank Barbara Egan, Lori Egan, Helena Judge Ellis, Hardeep Ranu, and Pati Soule for assistance, and the participants in the Nurses' Health Studies.

The WHI program is supported by contracts from the National Heart, Lung and Blood Institute, NIH. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at http://www.whi.org

The ACS study is supported by UO1 CA098710. We thank Cari Lichtman for data management and the participants on the CPS-II. The U.S. Radiologic Technologists Study (USRT) is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS.

The PLCO study is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. The authors thank Dr Philip Prorok, Division of Cancer Prevention, National Cancer Institute; the Screening Center investigators and staff of the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) and Mr. Tim Sheehy, and staff at SAIC-Frederick. Most importantly, we acknowledge the study participants for their contributions to making this study possible. The authors thank the radiologic technologists who participated in the study; Jerry Reid of the American Registry of Radiologic Technologists for continued support of the study; Diane Kampa and Allison Iwan of the University of Minnesota for study coordination and data collection; Dr. Bill Kopp and staff at SAIC-Frederick for biospecimen processing; and Laura Bowen of Information Management Systems for data management.

References

1. Colditz, GA.; Baer, HJ.; Tamimi, RM. Breast Cancer. In: David, S.; Fraumeni, JF., editors. Cancer Epidemiology and Prevention. Oxford University Press; New York, USA: 2006. p. 995-1012.
2. Miki Y. A strong candidate for the breast and ovarian-cancer susceptibility gene BRCA1. Science 1994;266:66–71. [PubMed: 7545954]
3. Wooster R. Identification of the breast cancer susceptibility gene BRCA2. Nature 1995;378:789–792. [PubMed: 8524414]
4. Rahman N. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nature Genet 2007;39:165–167. [PubMed: 17200668]
5. Meijers-Heijboer H. Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nature Genet 2002;31:55–59. [PubMed: 11967536]
6. Erkko H. A recurrent mutation in PALB2 in Finnish cancer families. Nature 2007;446:316–319. [PubMed: 17287723]
7. Renwick A. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature Genet 2006;38:873–875. [PubMed: 16832357]

8. Seal S. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nature Genet 2006;38:1239–1241. [PubMed: 17033622]

9. Cox A. A common coding variant in CASP8 is associated with breast cancer risk. Nature Genetics 2007;39:688.

10. Easton DF, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447:1087–93. [PubMed: 17529967]

11. Hunter DJ, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39:870–874. [PubMed: 17529973]

12. Stacey SN, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007;39:865–869. [PubMed: 17529974]

13. Stacey SN, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2008;40:703–706. [PubMed: 18438407]

14. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78. [PubMed: 17554300]

15. Yu K, et al. Population Substructure and Control Selection in Genome-Wide Association Studies. PLoS ONE 2008;3:e2551. [PubMed: 18596976]

16. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003;164:1567–87. [PubMed: 12930761]

17. Price AL. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909. [PubMed: 16862161]

18. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 2006;38:209–213. [PubMed: 16415888]

19. Gold B, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci U S A 2008;105:4340–5. [PubMed: 18326623]

20. Staiger H, et al. Novel meta-analysis-derived type 2 diabetes risk loci do not determine prediabetic phenotypes. PLoS ONE 2008;3:e3019. [PubMed: 18714373]

21. Xue F, Michels KB. Diabetes, metabolic syndrome, and breast cancer: a review of the current evidence. Am J Clin Nutr 2007;86:s823–35. [PubMed: 18265476]

22. Li X, Heyer WD. Homologous recombination in DNA repair and DNA damage tolerance. Cell Res 2008;18:99–113. [PubMed: 18166982]

23. Antoniou AC, et al. RAD51 135G-->C modifies breast cancer risk among BRCA2 mutation carriers: results from a combined analysis of 19 studies. Am J Hum Genet 2007;81:1186–200. [PubMed: 17999359]

24. Shlien A, et al. Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. Proc Natl Acad Sci U S A 2008;105:11264–9. [PubMed: 18685109]

25. Garcia-Closas M, et al. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genet 2008;4:e1000054. [PubMed: 18437204]

26. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 2008;358:2796–803. [PubMed: 18579814]

27. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. J Natl Cancer Inst 2008;100:978–9. [PubMed: 18612128]

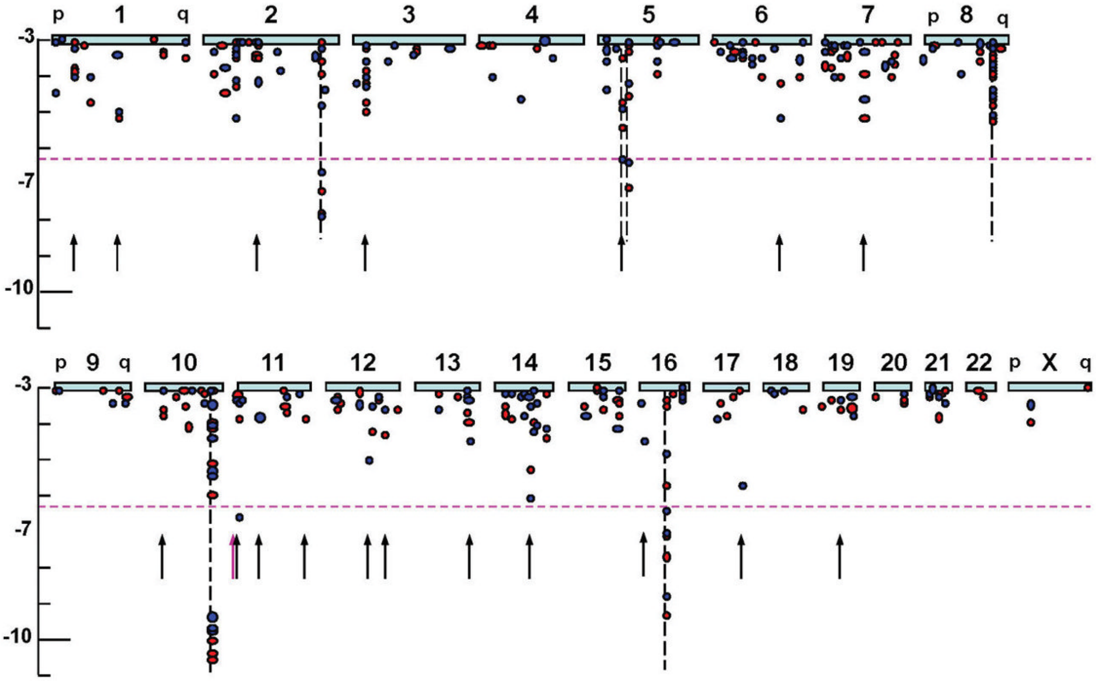
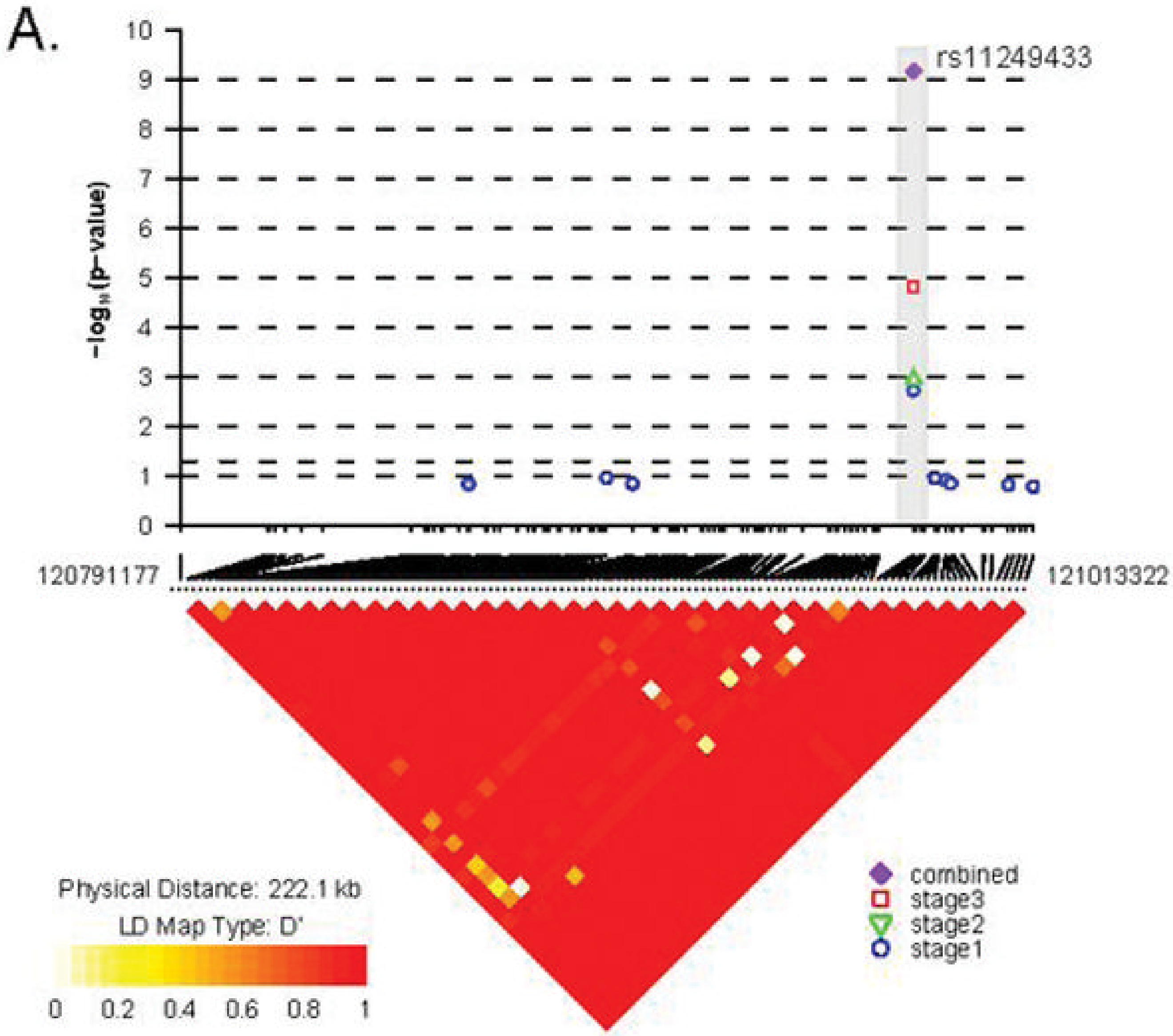


Figure 1. Results of Combined Stage 1 and 2

This figure includes the most promising SNP associations based on a combined analysis of Stage 1 and Stage 2. A joint analysis of the genotypes was performed using an age, study design and population stratification-adjusted logistic regression analysis (2 df test). Dashed vertical lines indicate loci previously reported in GWAS [10–13,19]. The horizontal magenta line denotes the range of genome-wide significance ($p < 5 \times 10^{-7}$). Black vertical arrows indicate loci explored in Stage 3 chosen on the basis of the p value. The magenta vertical arrow point to rs3817198 in the *LSP1* gene. Blue dots denote the results of the genotype test and red dots denote the trend test.


A.
10
9
8
7
6
5
4
3
2
1
0
-log10(p-value)
rs11249433
120791177
121013322
Physical Distance: 222.1 kb
LD Map Type: D'
0
0.2
0.4
0.6
0.8
1
combined
stage3
stage2
stage1

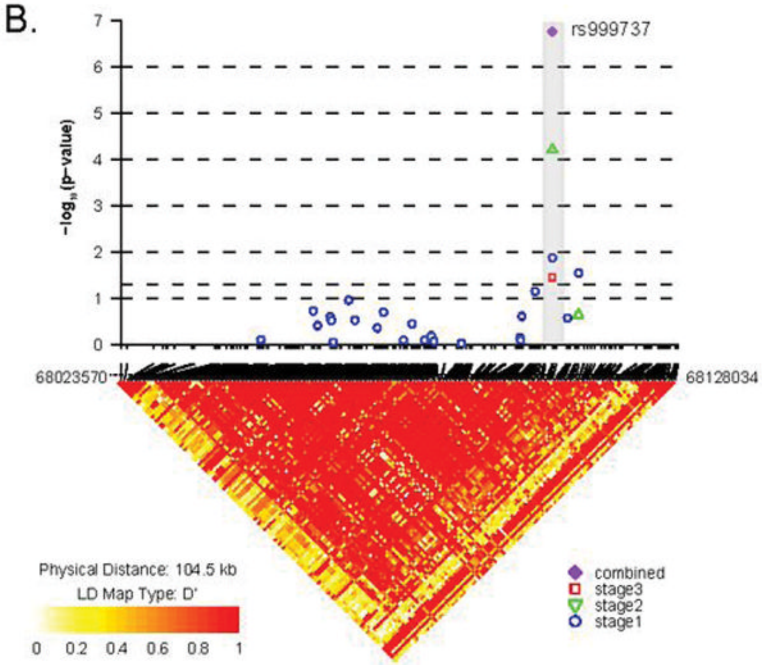


Figure 2. Linkage Disequilibrium plot of Two Novel Loci

Both panels present the LD plots (using D') for novel loci based on SNPs with MAF > 5% using HapMap Stage II individuals of European background (n=60 unrelated individuals). Above the plots are the results of the three individual Stages and the combined analysis for the SNPs reaching genome-wide significance. Panel A. Chromosome 1 region marked by rs11249433 and bounded by SNPs between chr1:120,400,700 −121,060,765. Note that one side is closely anchored to the centromere while the region distal to the centromere is bounded by a "SNP desert" of approximately 220 kb. Panel B. Chromosome 14q24.1 region marked by rs999737 and the block resides in the intron between two exons, of which the last has been observed in one of the three splice variants observed. Note that the SNP is located in an intron exclusive to the longest predicted transcript of *RAD51L1*.

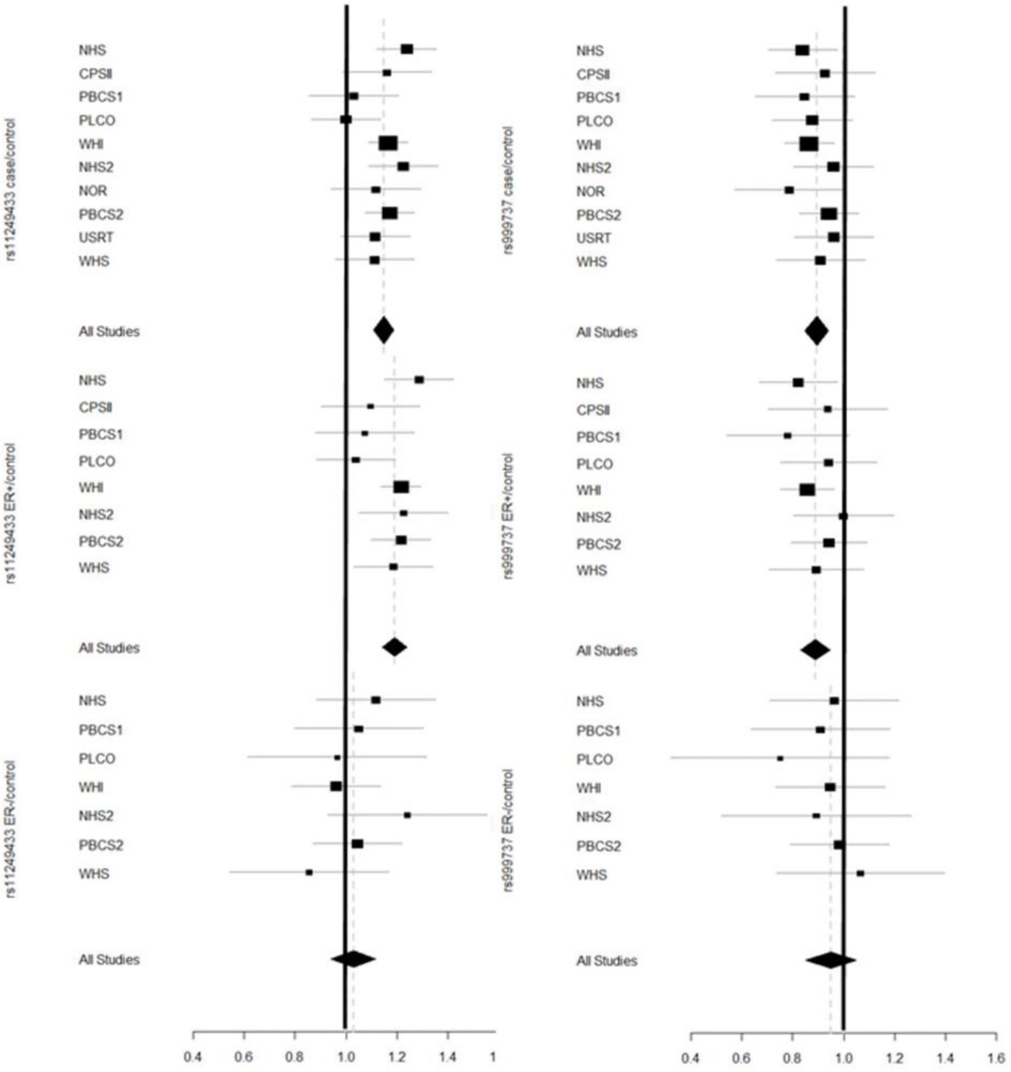


Figure 3. Forest plot for Overall, and ER+ and ER− Analysis, for rs 1124933 and rs999737

The results of the Overall Pooled analysis, and case-control analyses for ER+ cases, and ER −ve cases, were generated using a trend test with one degree of freedom. The figure includes per allele odd ratio (log additive/multiplicative model) for each study. For the overall analysis, the P-heterogeneity values are for rs1124933 P=0.44, and for rs999737 P=0.79. Data were available for estrogen-receptor status in 6,586 cases.

Table 1

Three-stage study design

	Controls	**Cases**
Stage 1 (528,173 SNPs)		
NHS1	1,142	1,145
Stage 2 (30,278 SNPs)		
CPSII	543	535
PBCS1	506	669
PLCO	975	948
WHI	2,410	2,395
Total Stage 2	4,434	4,547
Stage 3a (24 SNPs)		
CONOR	498	516
WHS	701	696
NHS2	1,243	619
USRT	998	780
PBCS2	1,783	1,467
Total Stage 3	5,223	4,078
Total Stages 1 – 3 Combined	**10,799**	**9,770**

Nine studies have participated in this multi-stage GWAS. Cases are represented with solid bars and the controls are represented by stippled bars. Note that part (26.6%, corresponding to 669 cases and 506 controls, designated as PBSC1) of the Polish Breast Cancer Study (PBCS) was genotyped using the custom iSelect Infinium (Illumina) and the remaining samples (73.4%. corresponding to 1,467 cases and 1,783 controls, designated as PBSC2) were genotyped in Stage 3.

Table 2

Results of Previously Reported Loci

				Genotype p-value*			Combined			
Chromosome band	**Proposed Candidate**	**SNPID+**	**Risk allele (freq)^**	**Stage 1**	**Stage 2**	**Stage 3**	**Controls/ cases**	**Genotype p-value**	**OR het (95% CI)**	**OR hom (95% CI)**
10q26.13	*FGFR2*	rs2981579	T (41%)	4.36×10^{-5}	1.22×10^{-6}	-	5283, 5439	1.79×10^{-10}	1.17 (1.07–1.27)	1.46 (1.30–1.62)
16q12.1	*TOX3*	rs3803662	T (27%)	5.3×10^{-2}	6.82×10^{-9}	-	5281, 5434	1.11×10^{-9}	1.16 (1.07–1.27)	1.55 (1.34–1.78)
5q11.2	*MAP3K1*	rs16886165	G (15%)	3.1×10^{-2}	1.17×10^{-5}	-	5283, 5440	5.00×10^{-7}	1.23 (1.12–1.35)	1.65 (1.30–2.10)
8q24.21		rs1562430	A (57%)	1.44×10^{-2}	4.74×10^{-4}		5285, 5440	1.28×10^{-5}	0.84 (0.77–0.92)	0.79 (0.71–0.89)
2q35		rs13387042	A (51%)	1.10×10^{-2}	1.48×10^{-6}		5285, 5433	2.10×10^{-8}	0.80 (0.73–0.87)	0.74 (0.67–0.83)
11p15.5	*LSP1*	rs3817198	C (32%)	5.36×10^{-1}	1.16×10^{-1}	4.34×10^{-1}	10316, 9408	6.51×10^{-2}	1.02 (0.96–1.08)	1.12 (1.02–1.23)
5p12		rs4415084	T (41%)	1.5×10^{-3}	1.6×10^{-2}	1.6×10^{-2}	10293, 9367	4.53×10^{-5}	1.09 (1.03–1.17)	1.20 (1.11–1.31)
5p12		rs10941679	G (26%)	-	-	5.5×10^{-3}	5490, 4575		1.12 (1.03–1.22)	1.20 (1.03–1.41)

[*] Adjusted genotype test with 2 df

[+] SNPID corresponds to dbSNP ID (http://www.ncbi.nlm.nih.gov/projects/SNP/)

[^] Estimated from controls in the combined (Stages 1–3) analysis

The results of the genotype and trend tests, both adjusted and unadjusted are presented in Supplemental Table 1.

Table 3

Novel SNPs in CGEMS

				Genotype p-value			Combined (stages 1–3)			
Chromosome band	Proposed Candidate	SNPID[*]	Risk allele (freq)[+]	Stage 1	Stage 2	Stage 3	Controls/ cases	Genotype p-value	OR het (95% CI)	OR hom (95% CI)
1p11.2		rs11249433	C (39%)	1.86×10^{-3}	1.11×10^{-3}	1.49×10^{-5}	10263, 9335	6.74×10^{-10}	1.16 (1.09–1.24)	1.30 (1.19–1.41)
14q24.1	*RAD51L1*	rs999737	C (76%)	1.31×10^{-2}	6.18×10^{-5}	3.49×10^{-2}	10298, 9395	1.74×10^{-7}	0.94 (0.88–0.99)	0.70 (0.62–0.80)
5p12	*MRPS30*	rs7716600	A (22%)	5.01×10^{-3}	7.66×10^{-5}	2.18×10^{-2}	10321, 9400	2.2×10^{-5}	1.10 (1.04–1.17)	1.28 (1.13–1.45)
5p12	*MRPS30*	rs2067980	G (16%)	1.63×10^{-2}	5.75×10^{-4}	6.14×10^{-1}	10309, 9391	1.24×10^{-3}	1.08 (1.02–1.15)	1.29 (1.09–1.52)

[*] SNPID corresponds to dbSNP ID (http://www.ncbi.nlm.nih.gov/projects/SNP/)

[+] Estimated from controls in the combined (Stages 1–3) analysis

The two additional 5p12 markers were chosen to explore the region reported[13]. One SNP assay for rs930395 did not design adequately, so a surrogate with LD=1.0 was substituted, rs7716600.