## SI Text

Microarray output was analyzed by using the Bayesian robust linear modeling using the Mahalanobis distance (BRLMM) algorithm: www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf. This is a modification of a published algorithm that normalizes fluorescent signals across multiple chips and makes inference across multiple SNPs to render more accurate calls (1). Only those 435,632 SNPs present both on the EAv3 and the Affymetrix 500K commercial array that had dbSNP rs numbers were included in further analyses. Of these, 221,233 SNPs were present on the NspI array, and 214,399 SNPs were present on the StyI array. Mapping of the SNPs to the May 2004 (NCBI build 35, hg 17) coordinates of the human genome sequence was possible through the annotation files supplied by Affymetrix: www.affymetrix.com/support/technical/byproduct.affx?product = 500k.

Detailed examination of SNP distribution and performance characteristics in an AJ population is described in ref. 2. Most of the statistical data discussed in that technical publication is summarized on a browser available at: http://theta2.ncifcrf.gov/cgi-bin/gbrowse/dbo/.

**Analytic Pipeline Workflow.** ***Affymetrix genotyping and data assembly.*** The workflow was as follows: cell intensity files or CEL files, from 500K-Affymetrix-chip mapped DNA were obtained. CEL files are binary files containing the fluorescence intensities for each probe on the microarray. Those files together with SNP-specific annotation files obtained from the Affymetrix web site were processed in two steps into a six-column prettybase file by using PERL scripts.

Affymetrix Power Tools (APT) includes an open-source linux command line program named apt-probeset-genotype software that implements the BRLMM genotyping algorithm (Bayesian Robust Linear Model with Mahalanobis distance classifier) that was run on all CEL files. A recently developed software tool (Chiamo) generated by the Wellcome Trust Consortium partially mitigates the discordance between BRLMM genotype calls and fluorescence intensity values (see J. Marchini, C. Spencer, Y. Teo, and P. Donnelly, *A Bayesian Hierarchical Mixture Model for Genotype Calling in a Multi-Cohort Study*, available at www.stats.ox.ac.uk/~marchini/software/gwas/chiamo.html), but this tool was not available at the time of our initial analysis. A related development is the Birdseed and BRLMM-P algorithm, developed jointly by Affymetrix and collaborators at the Broad Institute; this tool is not yet available for application to our data. The output files were assembled as two text files: CALL and CONF for each chip – for SNP calls and confidences. All SNP values are labeled with SNP Affymetrix ID. Our subsequent PERL script converts retrieved data into prettybase (PB) format. PB file format is based on the prettybase format, developed by the Seattle SNPs project, http://pga.gs.washington.edu. The prettybase file can be text manipulated using unix commands, PERL scripts, and regular expressions.

***Assembly of Illumina GoldenGate Data into Prettybase and Scorpio format.*** Five-column prettybase files were formed after export of a full data table from the BeadStudio Software. Although Full Data Table Exports included call confidence data (GenTrain Score), these were not used during this study. Instead, SNPs with GenTrain Scores <0.5 were excluded from further statistical workup. Manual reclustering was performed as needed in BeadStudio 3.1. Scorpio files for entry into Sapio Exemplar were created using a second PERL script from the five-column prettybase files.

Prettybase files were analyzed statistically by using SAS/Genetics, SPSS, PHASE, and R as described (3). StatTransfer, UltraEdit, and CRiSP were used to expedite file movement between analytic programs. Scorpio files were analyzed by using Sapio Sciences Exemplar. Briefly, SAS software converted the PB files into SAS datasets for further analysis. The genotypes were recoded so that the most common allele in a reference population (reference allele) was assigned a value of 0, and the other allele (variant allele) was assigned a value of 1. For each marker in each population summary, statistics were computed that included the number of observed alleles (1 or 2), the allele and genotype counts and frequencies, Hardy-Weinberg disequilibrium coefficient ($D$), inbreeding coefficient ($F_{IS}$), and the $\chi^2$ and $P$ value for the $\chi^2$ test for Hardy-Weinberg equilibrium (HWE). The genotype counts were used to compute exact HW $P$ values by the method of Wigginton *et al.* (4).

Several genetic distance statistics were computed for the populations including the Fixation Index of the Subpopulation within the Total ($F_{ST}$) (5) and Nei's standard genetic distance measure ($D_s$) (6). In addition, several information theory-based statistics were computed including entropy for admixed populations (7), Kullback-Leibler divergence (8), and informativeness for assignment statistic (In) of Rosenberg (2003) (9).

Three $\chi^2$ tests comparing marker allele and/or genotype frequencies were implemented by using SAS/Genetics software. They consist of the allele test (Pearson $\chi^2$ test on table of allele counts by disease status), the genotype test (Pearson $\chi^2$ test on table of genotype counts by disease status), and the trend test (Armitage test on table of genotype counts by disease status). Equivalent exact tests were implemented through SAS PROC FREQ. In addition, Fisher's exact test was used to test recessive and dominant models. Odds ratios were computed for the allele test by PROC ALLELE and for the trend test by PROC LOGISTIC.

**Determination of SNP Frequencies and Hardy-Weinberg Departures.** In a recent manuscript, several SNP characteristics were compared between the AJ and CEU subjects (2). One major conclusion from that work is that the AJ and CEU populations were entirely differentiable from each other on the basis of principal components analysis (PCA) (10). Population differentiation statistics from comparison of the CEU and AJ, with triangle plots of each are available on our public browser described in a subsequent section. For the X-chromosome analyses of HWE and computation of genetic measures, only the 30 females in the CEU data set were included.

**PCA.** Data from 60 CEPH-derived CEU subjects that were also part of the HapMap Project (www.hapmap.org) were included in the analysis for reference in the PCA. These 60 individuals were the parents of 30 trios, with the children excluded so that the genotype data could be considered independently. These experiments were performed by Affymetrix (www.affymetrix.com/support/technical/sample_data/500k_hapmap_ genotype_data.affx). As with the AJ population, CEU genotypes were determined by using the BRLMM algorithm. For each of the 90 CEU individuals, there were 489,913 genotype calls available from the HapMap web site (HapMap version 20, July, 2006). When compared with the commercial Affymetrix chip, there was 99.4% concordance (43,638,269 genotype calls agreed). In making this assessment, there were 67,394 absolutely discordant genotype calls (called heterozygous in HapMap but homozygous by Affymetrix array or vice versa), leading to mean call confidence scores for AJ and CEU that were within 3% of each other but significantly different. These differences and slight discordance probably reflect the large number of SNPs analyzed and use of the Affymetrix commercial arrays for the CEU analysis and early release arrays for the AJ analysis.

To assess possible population stratification confounding comparison of cases and controls, we used the numerical methods developed by Price (10) and implemented in the Eigenstrat package. Rather than use the second half of the package to adjust association *P* values for population stratification, we used only the PCA portion of the package and graphed the result of principal component 1 versus principal component 2. Research subject participation in the study was adjusted to those forming discrete clusters away from the CEU in the PCA. In addition to these graphs, we applied this PCA to a subset of the Ashkenazi controls in a separate analysis (2), and to the data in the SNP set of the current study reduced to minimize no calls (the 167,676 SNPs described above).

Although we also corrected for population stratification using the Eigenstrat PCA exclusion method, we made efforts to formulate an additional correction using the genomic control method of Roeder and Devlin (11). To do so, we used the less significant 90% of SNPs to define a median $\chi 2$ for trend and then adjusted to find $\lambda$. In both the initial survey of 435,632 SNPs and the “filtered” survey of 167,676 SNPs, a fractional $\lambda$ ($0 < \lambda < 1$) was found, suggesting that there was no stratification in our PCA-filtered sample, and further correction via the genomic control method would be inappropriate. Thus the statistics in Table 1 are adequately corrected for population stratification, and the excess of observed versus expected $\chi^2$ *P* values provides a good estimate of the number of SNPs that are significantly associated with breast cancer in our familial survey. Only SNPs in HWE were used to compile the table. SNPs within HWE were determined to be those above $P > 0.02$, based on a quantile-quantile graph evaluation (SI Fig. 3).

**Browser and Software.** The generic genome browser, gbrowse, (www.gmod.org/?q = node/71), was used to create a visual display of project data. This visual display is publicly available (http://theta2.ncifcrf.gov/cgi-bin/gbrowse/gold1/) and can be accessed from the main page of that URL under the rubric "Projects." The data on that browser includes summary association statistics, summary population divergence statistics, Hardy-Weinberg statistics, and an ability to download genotypes from the studies discussed here.

The R statistical package was used extensively for summary graphics. Data analyses were performed with PHASE, Sapio Science's Exemplar, SAS, SAS/Genetics, SPSS, PLINK, or R.

ACKNOWLEDGMENTS: We thank Drs. Scott Cain and Lincoln Stein for help and support in implementing the gbrowse data viewer and Dr. Jack Collins and Toni Harbaugh of the Advanced Biomedical Computing Center in Frederick, MD, for computational resources. Control samples were provided by the New York Cancer Project supported by the Academic Medicine Development Company of New York. This work was supported in part by federal funds from the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research, and from SAIC-Frederick under contract N01-CO-12400. The content of this publication does not necessarily reflect the views of the Department of Health and Human Services nor does its mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This research was also supported by Breast Cancer Research Foundation Grant 13740 (to K.O.), Susan Komen Foundation Grant R01-BCTR0601361 (to K.O.), the Lymphoma Foundation, and the Niehaus, Southworth, Weissenbach Cancer Research Fund. T.K. was supported by the Koodish and Frankel Fellowships.

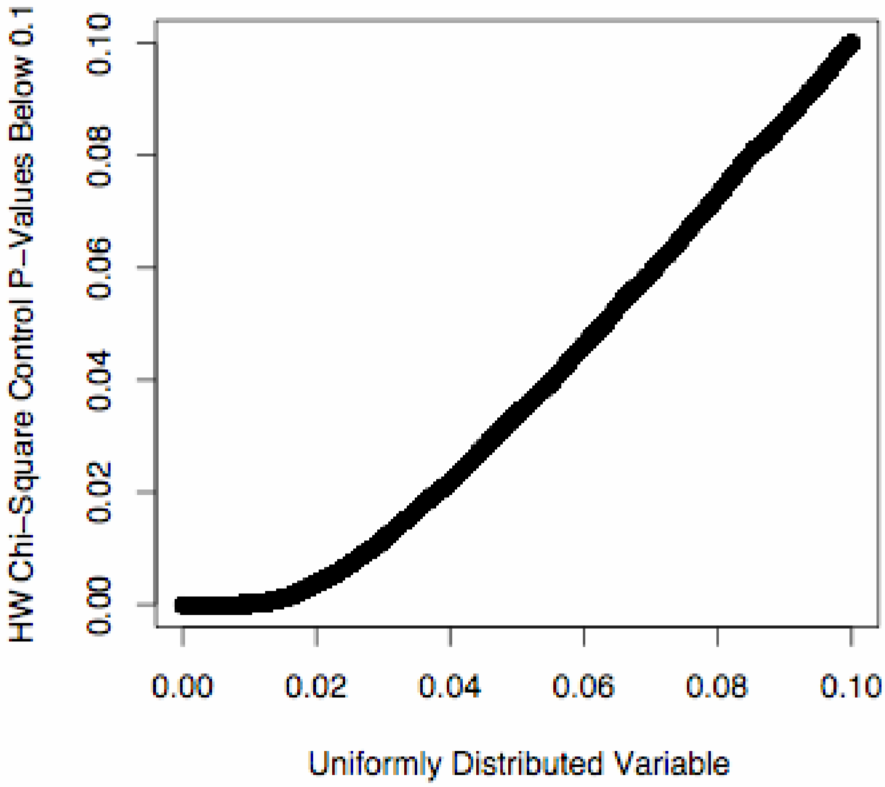

**Fig. 3.** Quantile-quantile plot of Hardy=Weinberg $\chi^2$ key values <0.1 against a uniformly distributed variable. Note linear portion of the graph >0.02.

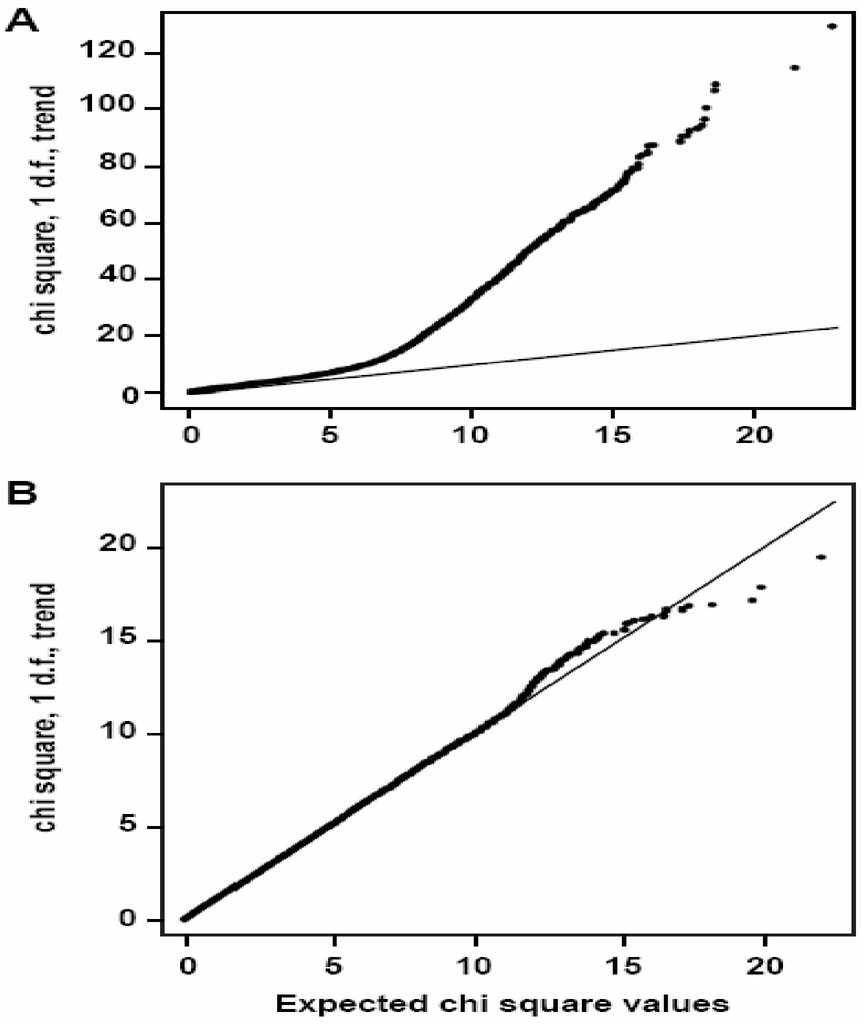

**Fig. 4.** Quantile-quantile plots for the Cochran-Armitage 1 d.f. $\chi^2$ trend test. (*A*) In the initial survey of SNPs in 249 familial cases and 299 controls at the 391,467 SNPs that did not deviate from Hardy-Weinberg equilibrium and with a nonzero trend test score. (*B*) For the same research subjects, at 150,071 SNPs where call rates were 99.6% or greater. Under the null hypothesis of no association at any locus, the points would be expected to follow the gray line.

1. Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays.*Bioinformatics* 22:7-12.

2. Olshen A, *et al.* (2008) Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping. *BMC Genet* 9:14.

3. Gold B, *et al.* (2006) Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nat Genet* 38:458-462.

4. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 76:887-893.

5. Weir B (1996) *Genetic Data Analysis 2: Methods for Discrete Population Genetic Data* (Sinauer, Sunderland, MA).

6. Nei M (1972) Genetic distance between populations. *Am Nat* 106:283-292.

7. Smith MW, *et al.* (2004) A high-density admixture map for disease gene discovery in African Americans. *Am J Hum Genet* 74:1001-1013.

8. Cover T, Thomas J (1991) *Elements of Information Theory* (Wiley, New York).

9. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) *Am J Hum Genet* 73:1402-1422.

10. Price AL, *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904-909.

11. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55:997-1004.