Supplementary Materials and Methods

Gene resequencing

All candidate genes were resequenced in order to establish a comprehensive catalogue of their common genetic variants. This work was performed at three BPC3 collaborating genome centers: USC/Broad Institute, CEPH, and NCI. Exons, intron/exon junctions and evolutionarily conserved (at least 80% homology with mouse sequence over at least 200bp) sequences in introns and sequences up to 30kb 5' of transcription start and 10kb 3' of translation end of each gene were sequenced in a panel of 95 advanced breast cancer cases from the MEC and EPIC (19 of each ethnic group represented in the study: African American, Latino, Japanese, Native Hawaiian, and Caucasian). SNPs with minor allele frequency greater than 5% in any of the five ethnic groups or greater than 1% overall were selected for further work.

SNP Selection and Genotyping

SNP selection and genotyping was done at the same genome centers where resequencing took place. The SNP selection and genotyping at each center is described below.

USC/Broad Institute: High density genotyping of SNPs identified by resequencing complemented by HapMap data was conducted in 14 candidate genes at USC/Broad in three sample panels: 1) 70 European Americans from the MEC; 2) 30 CEU trios from the HapMap project, and 3) 14 additional CEPH trios. Genotyping was performed at the Broad Institute using Sequenom, Taqman and Illumina platforms.

Tag SNPs were selected from a combined data set of these three panels and HapMap data in two phases using a pairwise r^2 procedure implemented in TagSNPs (http://www-rcf.usc.edu/~stram/tagsnps.html). In phase 1, we selected tag SNPs to capture alleles with minor allele frequencies down to 2% (at USC) or 3% (at CEPH) with an $r^2 \geq 0.8$. To increase genotyping success rates in the case-control samples, we also limited the selection of tags to those with Illumina design scores >0.4. In phase 2, we selected additional tags from a combined data set comprised of phase II HapMap data for the CEU trios plus genotype data that we generated in the same panel (#2 above). Tags from phase 1 were forced in and additional tag SNPs were chosen until each SNP (with a frequency of $\geq$5%) in the phase 2 data set had at least one pairwise surrogate with an $r^2 \geq 0.8$. We also selected >1 tag to represent bins that contain >8 proxies. Details regarding the genotype data and tag SNP selection procedure can be found at: http://www.uscnorris.com/Core/DocManager/DocumentList.aspx?CID=13 and http://cgf1.nci.nih.gov/cohort.cfm.

CEPH: A total of 23 genes were studied at CEPH. High density genotyping was done on all SNPs identified with more than one occurrence in the sequencing panel and on all SNPs (MAF ≥5%) described in public databases (30 kb up stream from the start codon and 10 kb down stream of the stop codon). Samples tested were:

- 190 DNAs from MEC for SNPs identified by resequencing (quality controls),
- 89 CEPH trios including 28 CEU trios from the HapMap project
- 1064 DNAs from the Human Genome Diversity Project-CEPH panel (51 populations from all over the world, http://www.cephb.fr/en/hgdp/)
- 58 DNAs from Hawaiian origin.

Genotyping was performed at the CEPH, Paris, France, the Core Genotyping Facility, Intramural Research Program of the NCI and the Centre National de Génotypage, Evry, France using TaqMan and Illumina platforms.

TagSNP selection was performed in the same way as described for USC/Broad.

NCI: High density genotyping was conducted across the remaining 6 genes in three panels based on sequence analysis and available HapMap Stage II data. Genotyping of SNPs identified by resequencing was performed based on the panel of the SNP500Cancer and HapMap CEU (60 unrelated subjects). Within regions of strong LD, haplotype frequencies estimates were constructed from genotype data using the expectation-maximization (E-M) algorithm of Excoffier and Slatkin (Slatkin and Excoffier, 1996). The squared correlation (R_h^2) between the true haplotypes (h) and their estimates were calculated as described by Stram et al (2003). Tagging SNPs for the BPC3 case-control samples were then chosen by finding the minimum set of SNPs that would have $R_h^2 \geq 0.7$ for all common haplotypes with an estimated frequency of $\geq$ 5%.

Slatkin M, Excoffier L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity. 1996 Apr;76 (Pt 4):377-83.

Stram, D. O., Haiman, C. A., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E., and Pike, M. C. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered, 55: 27-36, 2003.

Supplementary material

Supplementary Methods. Detailed description of SNP selection methodology.

Supplementary Table 1. SNPs identified by resequencing of candidate genes in a multiethnic panel.

Supplementary Table 2. SNP and sample success rates and quality control.

Supplementary Table 3. Associations between SNPs genotyped in steroid hormone-related (3a) and IGF-I-related (3b) genes and breast cancer risk. The columns in this table describe the following: gene name, NCBI dbSNP rs number, chromosome, position on the chromosome (NCBI build 36), genotypes, numbers of cases and controls for each of the three genotypes, odds ratios for heterozygotes and for homozygotes for the rare allele (referred to the homozygotes for the common allele), with 95% confidence interval, p-value of the trend test, p-value of the test with two indicator variables (2 d.f. test), stratum, hormone whose levels were found associated with the specific SNP in previous work, reference for the SNP-hormone association. Definitions of strata are as follows: All_Women=all cases and their controls; No_Carci=all cases and their controls except *in situ* carcinomas and their controls; Caucasian=all cases of Caucasian origin and their controls; Advanced=all invasive cases having regional or distal metastases and their controls; Non-advanced=all invasive cases without metastasis; gt55=all cases who were 55 or older at recruitment and their controls; le55=all cases who were younger than 55 at recruitment.

Supplementary Table 4. Associations between SNPs genotyped in candidate genes and breast cancer risk in all study subjects, stratified by BMI. The columns in this table describe the following: gene name, NCBI dbSNP rs number, possible genotypes, numbers of cases and controls for each of the three genotypes, odds ratios for heterozygotes and for homozygotes for the rare allele (referred to the homozygotes for the common allele), with 99% confidence interval, p-value of the test for heterogeneity, BMI stratum (<25, between 25 and 30, ≥30), p-value of the trend test for each stratum of BMI.

Supplementary Table 5. Associations between SNPs genotyped in candidate genes and breast cancer risk in all study subjects, stratified by ER status. Column names are as in supplementary table 3.

Supplementary Table 6. Associations between SNPs imputed in candidate genes and breast cancer risk. Column names are as in supplementary table 3.