Cancer Genetic Markers of Susceptibility (CGEMS) Breast Cancer Genome-Wide Association Scan

The CGEMS data portal provides public access to summary results for approximately 528,000 SNPs genotyped in the CGEMS breast cancer scan (using the Illumina HumanHap550 chip, Illumina, San Diego, CA) in 1,145 breast cancer patients and an equivalent number of controls (n=1,142). Analysis of nearly 550,000 SNP genotypes per subject provides approximately 90% coverage of common SNPs based on HapMap Phase 2 with minor allele frequency (MAF) greater than 5% in the European population and a linkage disequilibrium coefficient threshold of $r^2>0.8$ using the TagZilla program (http://tagzilla.nci.nih.gov/)[1-3].

Summary CGEMS data can be viewed via the CGEMS data portal and downloaded in bulk via https://caintegrator.nci.nih.gov/cgems. Access to a subset of the individual raw phenotype and genotype data analyzed in the Nurses' Health Study will be possible for scientific research purposes only. Registration by the individual investigator and the supporting institution will be required because of privacy concerns. The accessible data will include genotypes from the Genome-Wide Association Scan (GWAS) and a set of covariates, namely, age (in 5 intervals: <55, 55-59, 60-64, 65-69, 70-74, and >74), family history of cancer (yes/no), and disease phenotype (control, case diagnosed with invasive breast cancer). Access to additional covariate data will be possible through established data sharing policies of NHS (http://www.channing.harvard.edu/nhs).

Study Population:

The Nurses' Health Study[4] (NHS) is a longitudinal study of 121,700 women enrolled in 1976. The CGEMS case-control study is derived from 32,826 participants who provided a blood sample between 1989 and 1990 and were free of diagnosed breast cancer at blood collection and followed for incident disease until May 2004. Cancer follow-up in the NHS was conducted by personal mailings and searches of the National Death Index. It is estimated that the percentage of true cancers captured by this system is greater than 90%. Permission was requested from all participants diagnosed with cancer to review medical records to confirm the diagnoses and obtain additional information on tumor histology, staging, and other characteristics. All study participants who were menopausal at blood draw with a confirmed diagnosis of invasive breast cancer and had sufficient stored blood available for DNA extraction at the time of case and control selection were included as cases in the CGEMS project. Controls were matched to cases based on age, blood collection variables (time, date, and year of blood collection, as well as recent (<3 months) use of postmenopausal hormones), ethnicity (all cases and controls are self-reported Caucasians), and menopausal status (all cases and controls were menopausal at blood draw).

Informed consent was obtained from all participants. The study was approved by the Institutional Review Board of the Brigham and Women's Hospital, Boston, MA, USA.

Sample handling:

DNA samples were received from the NHS bio-repository and visually inspected for adequate fluid in individual tubes. Three measurements of quantification were performed according to the standard procedures at the Core Genotyping Facility of the National Cancer Institute[5]. These include pico-green analysis, optical density spectrophotometry and real time PCR (http://cgf.nci.nih.gov/dnaquant.cfm). Samples were also analyzed with 15 short tandem repeats and the Amelogenin marker in the Identifiler™ Assay (ABI, Foster City, CA). All samples advanced to genotype analysis completed no less than 13 of the 15 micro-satellite markers.

After final review and sample handling, a total of 1,183 cases DNAs, and 1,185 controls DNAs were selected for genotyping in CGEMS. 93 DNAs were aliquoted twice and five DNAs were aliquoted three times, resulting in the addition of 103 redundant DNAs from the NHS useful for quality control. Finally, 23 external non-NHS quality control (QC) DNAs were added. Thus a total of 2,494 DNA samples were attempted for genotyping.

Selection of SNPs: Genotyping of the CGEMS Breast Cancer Study was performed at the Core Genotyping Facility using the Sentrix® HumanHap550 genotyping assay according to a protocol designated by the manufacturer[6-8].

Quality control

Initial Assessment of sample completion rates

A total of 555,352 SNP genotype assays were attempted on the 2,494 DNA samples using the Illumina HumanHap550 assay. If the completion rate for a sample was below 94%, then the sample was assayed a second time. Samples that did not meet the 94% completion threshold after a second attempt were excluded from further analysis. 59 samples from NHS (30 cases and 29 controls) were excluded based on this criterion. The remaining 2,435 DNAs were retained for the subsequent analyses

Assessment of SNP call rates

A total of 8,706 SNPs (~1.57% overall) failed to provide reliable genotype results due to either no calls or low call rates (<90%). Further quality control analysis was performed on the remaining 546,646 SNPs. An additional 18,473 SNPs with an observed low MAF (<1%) were dropped from the association analysis; thus 528,173 SNPs (95.1%), were maintained in the subsequent analyses.

Table 1. Summary of Completion rate for NHS samples

Sample Completion rate for	528,173 SNPs (retained)	546,646 SNPs (attempted)
Scan 1 study	99.754 %	95.754 %
Scan 1 case	99.756 %	95.704 %
Scan 1 control	99.773 %	95.799 %

The genotyping of the 528,173 retained SNPs on the 2,412 NHS DNAs with high completion rate generated 1.26 billion genotype calls. For this set of SNPs and samples, the percentage of missing data was significantly less than 1%.

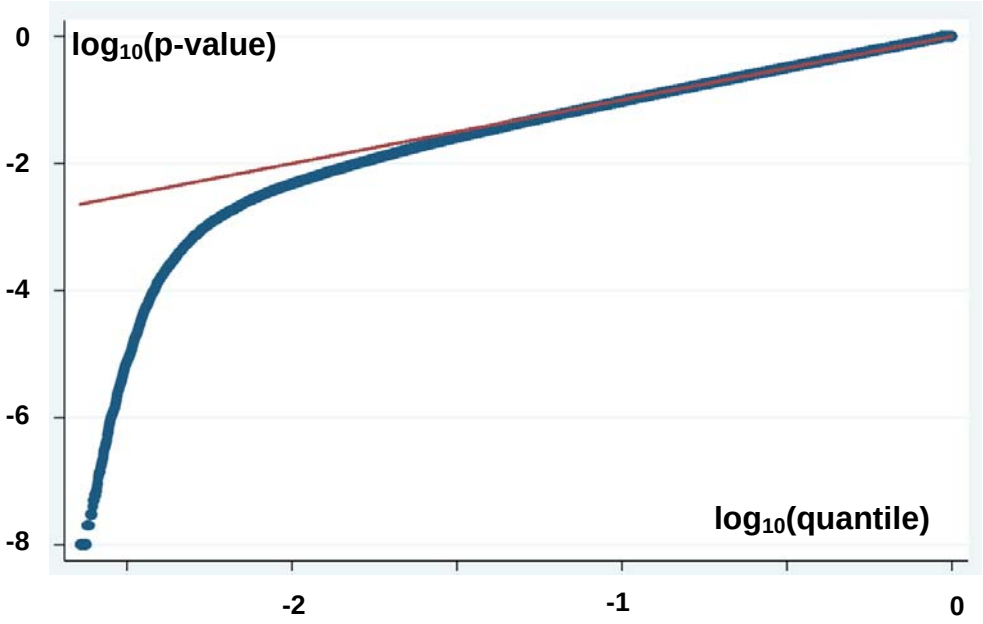
Concordance rate

The genotype concordance rate for SNP assays was evaluated using the 93 pairs of known NHS duplicated DNAs. These pairs of DNAs were separate aliquots from the same DNA preparation and all met quality control criteria requested for the other DNAs, thereby, providing reliable data for comparison. Analysis of the discrepancies within these pairs revealed similar results to the CEPH DNA duplicates reported in the CGEMS prostate cancer genome-wide association scan. An average concordance rate of 99.985% was observed (50,820,003 concordant genotype calls out of 50,827,468 comparisons). No SNPs or samples were excluded from further analysis as a result of this analysis.

Hardy–Weinberg Proportions in control DNA

Genotype data were tested for deviation from Hardy-Weinberg proportions using an exact test[9]. The analysis was conducted in the NHS control group. Significant deviations were observed for 29,318 SNPs (5.55% of all SNPs) at the level of p = 0.05 and for 2,880 SNP (0.55%) at p =0.001 (see figure 1). However, none of these SNPs were excluded from analysis since the tests for association applied to such data are valid in the presence of departure from Hardy-Weinberg proportions, although with potentially reduced power when these deviations are due to systematic genotyping errors with equal effects among cases and controls.

Figure 1. log scale p-value quantile plot for deviation from Hardy-Weinberg proportion



The under the assumption of Hardy-Weinberg proportion for all typed SNPs, the p-value quantile plot is expected to be linear along the first diagonal (red line). The plot shows evidence of a deviation towards lower p-values. particularly for one percentile of the SNPs with lowest p-values. Of 528,169 chromosomal SNPs, 1,200 had p-values lower than 10^{-8} and are not represented.

Final sample selection for association analysis

For all DNAs the frequency of heterozygote loci on the X chromosome was compatible with a female origin. Eighteen DNA samples (5 cases, 13 controls) revealed unclear identity as they could not be mapped back unambiguously to previous genotype results. They were not maintained in the study. Subsequent inspection of the genotype concordance rate between pairs of DNAs did not disclose additional unexpected duplicates. Finally, based on two analyses with two independent sets of 7,050 and 7,061 SNPs with very low linkage disequilibrium (r^2<0.01) using the STRUCTURE[10] program, 4 subjects (3 cases, 1 control) were estimated to be of admixed origin with greater than 15% of either Asian or West African ancestry. These 4 subjects were removed from subsequent analyses. Thus, the genome-wide association scan was performed on a final set of 2,287 unique subjects, of which 1,145 were cases and 1,142 were controls.

Summary of selection of cases and controls for association analysis

	cases	controls
Initially attempted	1,183	1,185
- low completion rate	30	29
- unclear identity	5	13
- admixed origin	3	1
= Used in scan	1,145	1,142

Association Analysis

The primary analysis of the CGEMS breast GWAS study explores the association between a single SNPs and breast cancer susceptibility in a group of 1,145 breast cancer patients and 1,142 controls. This exploration is done one SNP at a time, sequentially for each of the 528,173 SNPs maintained in the study. The analytic approach assumes no structure to the risk across the 3 possible genotypes at each locus. This approach maintains power to detect recessive or over-dominant alleles at the cost of a small decrease in power relative to a Cochrane-Armitage trend test for the detection of alleles with multiplicative risk effect. By maximizing genome coverage with a large number of SNPs and adopting an 'agnostic' approach to the analysis which does not take gene function or prior information on breast cancer or other phenotypes into consideration, we increase the opportunity to pursue different working hypotheses and different regions of interest now and in the future.

Analytic procedure

Analysis with single selection of cases and controls

Each participant is classified in a unique group according to her phenotype at the end of the follow-up period, namely a case diagnosed with invasive breast cancer or a matched control.

Genotypes
In order to maintain high power to detect SNPs that are involved in non-multiplicative models (such as complete recessivity or over-dominance), we provide analyses of the data based on observed genotypes, considering each of the three possible genotype states separately. Accordingly, the analysis uses a statistical test with two degrees of freedom (two phenotypes cross-tabulated with three genotypes) when the rare homozygote genotype is observed more than 15 times (a somewhat arbitrary parameter imposed for numerical reasons). Otherwise, the rare homozygote genotype count is collapsed with the heterozygote genotype count, and a one degree of freedom test is used (such aggregation of genotypes was performed for 64,589 SNPs).

Single SNP statistics
In order to expedite public access to the data, the first-pass analysis of the CGEMS data aims at detecting association of single SNPs with breast cancer susceptibility. Multi-SNP approaches, such as haplotype association, have not yet been performed.

Population stratification
The pooled case and control DNAs were analyzed using a set of 14,111 SNPs with low pair-wise linkage disequilibrium ($r^2 < 0.01$) using the procedure described by Price *et al.*[11]. Testing for significance using the Tracy-Widom statistics[12] identified 4 significant principal components at the level of $p < 0.05$. Inspection of the distribution of the DNAs in the space defined by these components revealed little difference between cases and controls. Nevertheless, statistically borderline significant differences in this distribution for local groups observed in the space defined by the first three components, led to retain these components in the statistical analysis. No clear difference in the distributions was observed with the 4^{th} component, so it was not retained in the analysis.

Statistical tests.

We performed two sets of analyses. For each test, analysis included

- 528,173 **SNPs**,
- 1,145 ***cases*** diagnosed with breast cancer,
- 1,142 ***controls*** that were not diagnosed with breast cancer at the time of matching.

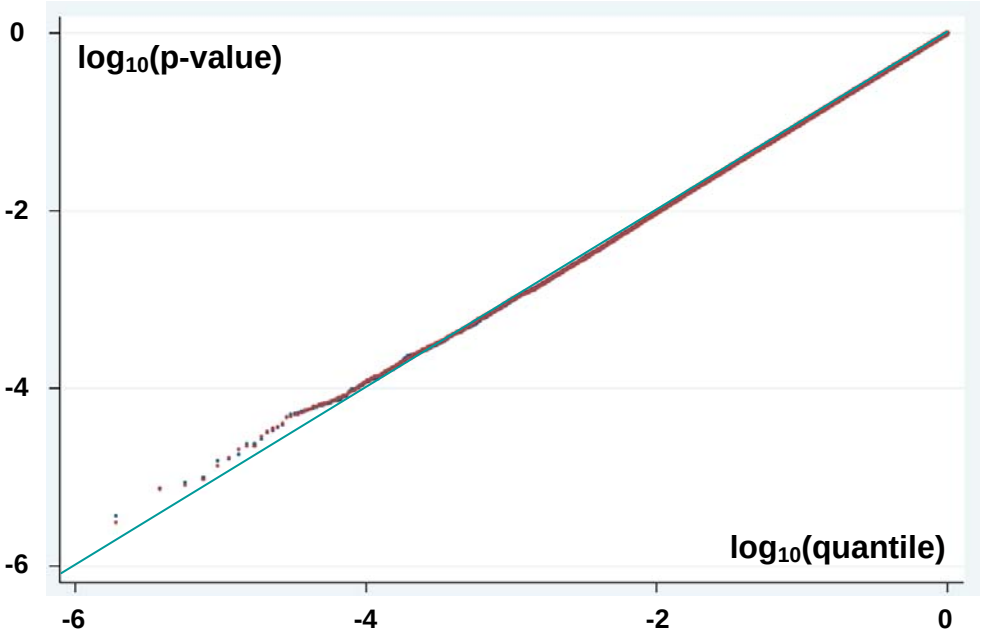
The characteristics of the two tests are:

1. ***Unadjusted*** score test
    - ***3-by-2 contingency table*** of genotypes by phenotypes was constructed.
    - ***No adjustment*** for covariates.
    - The ***p-value*** from the standard test of independence was computed from the ***3-by-2 contingency table***, based on a chi-squared test with up to 2 degrees of freedom.

2. ***Adjusted*** score test

- ***Dichotomous unconditional logistic regression*** was performed.
  - Two phenotypes (either case or control);
  - The regression variable was performed on two-indicator variables for heterozygote and rare homozygote genotype states.
- ***Adjustment*** for
  - 4 indicator variables for age group at randomization (ages categories <55, 55-59, 60-64, 65-69, 70-74, and >74, with range 55-59 as the reference),
  - an indicator for known hormone replacement therapy status at blood draw,
  - an indicator for unknown hormone replacement therapy status at diagnosis; this variable is included because it was a matching factor
  - three sets of coefficients to adjust for estimated population stratification corresponding to the top three eigenvectors identified by the principal component analysis.
- The ***p-value*** was obtained from a score test with up to 2 degrees of freedom.

Figure 2. log scale p-value quantile plot for association tests statistics



All SNPs are represented. Red dots: association test with adjustment for covariates; blue dots association test without adjustment for covariates. The two plots are almost indistinguishable, except for a small number of SNPs with low p-values. The green line represents the expected (uniform) distribution under the null hypothesis.

Interpreting the results

In examining the results one should keep in mind the following points:

1. Markers were selected on genomic criteria, not on functional basis. In the absence of complementary information, each of the SNPs has a low *a priori* probability of being associated with disease. Observation of a low p-value in these tables is not sufficient evidence to demonstrate an association for the marker; additional studies are required to confirm the association. For this analysis, we expected to observe roughly $\alpha \times 5 \times 10^5$ p-values lower than a specified α when there is one statistical test for each of 5×10^5 SNPs by chance alone; thus for $\alpha = 10^{-3}$ or $\alpha = 10^{-5}$, we expected to observe 500 and 5 SNPs, respectively, meeting the criterion, even if none of the 5 x 10^5 SNPs are not associated with breast cancer risk. In the pre-computed analysis presented we observed 529 and 531 depending on which of the two tests was selected for $\alpha = 10^{-3}$. For $\alpha = 10^{-5}$, we observed 4 SNPs for either tests. Nevertheless, the observation of a low p-value for a SNP in this GWAS alone does suggest that the associated gene or chromosomal region has an increased likelihood of harboring a breast cancer susceptibility locus but follow-up analysis is required and is planned in the forthcoming phases of CGEMS (http://cgems.cancer.gov/). The log scale p-value quantile plot is shown for both tests on figure 2.

2. Many pairs of SNP markers may have substantial correlation between them. In fact, correlation may extend across several markers on the same chromosomal region. Before interpreting the observation of clustering of SNPs with low p-values in a small chromosomal region as a strong signal of the presence of susceptibility loci in the region, one must consider that the clustering may be a consequence of linkage disequilibrium among nearby SNPs.

3. The two tests used for each SNP are strongly correlated, so is probably best to choose one. We recommend using the adjusted score test for exploratory purposes.

Citation of data used:

Please cite the website for publications related to data available on this website (http://cgems.cancer.gov/) and reference the full name of the study, Cancer Genetic Markers of Susceptibility.

Reference List

1. Barrett, J.C. & Cardon, L.R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38**, 659-662 (2006).

2. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

3. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-796 (2003).

4. Rockhill B, et al. Physical activity and breast cancer risk in a cohort of young women, J Natl Cancer Inst 90:1155-1160, (1998)

5. Haque, K.A. *et al.* Performance of high-throughput DNA quantification methods. *BMC. Biotechnol.* **3**, 20 (2003).

6. Gunderson, K.L. *et al.* Whole-genome genotyping. *Methods Enzymol.* **410**, 359-376 (2006).

7. Gunderson, K.L. *et al.* Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. *Pharmacogenomics.* **7**, 641-648 (2006).

8. Steemers, F.J. & Gunderson,K.L. Illumina, Inc. *Pharmacogenomics.* **6**, 777-782 (2005).

9. Wigginton, J.E., Cutler,D.J. & Abecasis,G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887-893 (2005).

10. Falush D, Stephens M & Pritchard JK. Inference of population structure using multilocus genotype data : linked loci and correlated allele frequencies. *Genetics* 164:1567-1587 (2003)

11. Price AL et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 8:904-909 (2006)

12. Patterson N, Price AL & Reich D. Population structure and Eigenanalysis. *PLoS Genet* 2:2074-2093 (2006)