Putative linkage signals identified for breast cancer in African American families

Heather M. Ochs-Balcom¹, Xiangqing Sun², Yanwen Chen²,³, Jill Barnholtz-Sloan²,³, Deborah O. Erwin⁴, Lina Jandorf⁵, Lara Sucheston-Campbell⁴, Robert C. Elston²,³

1. Department of Epidemiology and Environmental Health, School of Public Health and Health Professions, University at Buffalo, Buffalo, New York, 14214, USA
2. Department of Epidemiology and Biostatistics, Case Western Reserve University School of Medicine, Cleveland, Ohio, 44106, USA
3. Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio, 44106, USA
4. Department of Cancer Prevention and Population Sciences, Roswell Park Cancer Institute, Buffalo, New York, 14263, USA
5. Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, 10029, USA

Running title: African American breast cancer linkage study

Keywords: breast cancer, family, linkage study, African American

Financial support: Dr. Ochs-Balcom was supported by a grant from Susan G. Komen for the Cure and the School of Public Health and Health Professions at the University at Buffalo.

Correspondence:
Heather Ochs-Balcom, PhD
University at Buffalo
270 Farber Hall
Buffalo, NY 14214-8001
716-829-5338
hmochs2@buffalo.edu

The authors disclose no potential conflicts of interest.

Word count: 2906
Total number of figures: 3
Total number of tables: 1
Abstract

Background: Genome-wide association studies have identified polymorphisms associated with breast cancer subtypes and across multiple population subgroups; however, few studies to date have applied linkage analysis to other population groups.

Methods: We performed the first genome-wide breast cancer linkage analysis in 106 African American families (comprising 179 affected and 79 unaffected members) not known to be segregating BRCA mutations to search for novel breast cancer loci. We performed regression-based model-free multi-point linkage analyses of the sibling pairs using SIBPAL, and two-level Haseman-Elston linkage analyses of affected relative pairs using RELPAL.

Results: We identified \(-\log_{10} p\)-values that exceed 4 on chromosomes 3q and 12q, as well as a region near BRCA1 on chromosome 17 (\(-\log_{10} p\)-values in the range of 3.0-3.2) using both sibling-based and relative-based methods, the latter observation may suggest that undetected BRCA1 mutations or other mutations nearby such as HOXB13 may be segregating in our sample.

Conclusions: In summary, these results suggest novel putative regions harboring risk alleles in African Americans that deserve further study.

Impact: We hope that our study will spur further family-based investigation into specific mechanisms for breast cancer disparities.
Introduction

Breast cancer is a complex phenotype that is known to display disparities according to race; in particular, African Americans are known to have higher pre-menopausal breast cancer incidence, and overall higher breast cancer mortality, compared to European Americans (1). In spite of linkage analyses published decades ago for European American pedigrees (2-5) and numerous genome-wide association studies for the breast cancer phenotype (6-15) and more homogeneous breast cancer subtypes (16-21), and in different population groups (22-24), family-based linkage approaches have not been attempted specifically in African Americans. Linkage analysis is a powerful tool for identifying the presence of novel loci for both Mendelian and complex diseases and phenotypes; several successes have been reported for colon (25), prostate (26) and lung cancers (27).

It is recognized that difficulties exist in recruitment of minorities for genetic research and existing family-based data are limited. The Jewels in Our Genes study began in 2009 for the sole purpose of conducting linkage analysis in African American pedigrees to search for novel breast cancer loci. Using this resource, we conducted the first breast cancer genome-wide linkage analysis in African Americans.

Materials and Methods

Recruitment and Study Approval. Recruitment was focused primarily on sibling pairs that were both concordant and discordant for invasive breast cancer and as well as affected relative pairs (sibling, avuncular, grandparental, first cousin). In addition to recruiting women locally from Buffalo, participants were recruited via a partnership with the National Witness Project®, nationally at conferences and via the Love Army of Women, and partnerships with investigators from existing epidemiologic studies (28). The Jewels in Our Genes study was approved by the University at Buffalo Health Sciences Institutional Review Board (IRB). Our collaborators also obtained IRB approvals to make further contact with previous participants, when necessary.
Women who were interested and eligible were mailed out a packet containing the informed consent and HIPAA forms, study questionnaire, medical records consent and an Oragene saliva collection kit (DNA Genotek). We obtained medical records to adjudicate date of diagnosis, invasive breast cancer diagnosis and tumor characteristics where available.

**Genotyping and Quality Control.** Pedigrees were selected for genotyping (for the primary purpose of linkage analysis) according to relationship type and informativeness. DNA was isolated according to standard protocol from Oragene DNA kits (DNA Genotek) at Roswell Park Cancer Institute (Buffalo, New York). Normalization and analytical quality control of DNA samples and automated preparation of Affymetrix Axiom® Genome-Wide AFR (Admixed population of West African and European ancestry) SNP microarrays (Affymetrix), was performed at Rutgers Environmental and Occupational Health Sciences Institute (Piscataway, New Jersey). Twenty individuals with a dish QC<0.82 were dropped from the analysis, as well as one sample with a dish QC of 0.828 and a call rate of 96.97%, for a total of 21 (6.4%) individuals who were excluded; this resulted in loss of 13 pedigrees. Average heterozygosity was calculated by individual for all SNPs; all values ranged from 0.199-0.245, therefore no individuals were dropped from the analysis. Sex checks were all consistent with reports.

**Preparation of SNPs for Linkage and Pedigree QC.** We limited SNPs to those conforming to Hardy Weinberg proportions (p>0.05) with a call rate ≥99%, and as much as possible with minor allele frequency (MAF) ≥0.40. After thinning the SNPs to at least 0.33 cM apart, we added SNPs with smaller MAF, whenever possible, so that SNPs were not more than 0.67 cM apart. We thus prepared a panel of 8261 SNPs for linkage analysis, 97% of which had MAF ≥0.40 (see Figure 1). We then ran S.A.G.E. RELTEST and MARKERINFO (29) to look for relationship errors and marker inconsistencies; we corrected pedigree relationships as appropriate.
Preparation of Genetic Map File. We generated a genetic map file by using the genetic map for African Americans that was built by Hinch et al (30) based on about 2.1 million crossovers in 30,000 unrelated African Americans.

IBD Estimation. We estimated founder marker allele frequencies from all the pedigree members by maximum likelihood for all the SNPs to be used for linkage using FREQ (29). Multipoint exact identity-by-descent (IBD) sharing for every pair of relatives was estimated by GENIBD (29) by calculating the likelihood of each inheritance vector at multiple markers to generate IBD distributions.

Ancestry Estimation. To obtain individual-level ancestry estimates, we selected 16,697 ancestry informative markers (AIMs) that had allele frequency differences >0.5 between the CEU and YRI populations; among them, 1440 were in linkage equilibrium ($r^2<0.004$). We evaluated linkage disequilibrium within a window of ten SNPs, shifting the window five SNPs at each step. In order to assess what number of ancestral populations is appropriate in our sample, we selected 111 unrelated individuals from the pedigrees and ran STRUCTURE for 1, 2, 3, 4 and 5 groups (31) with the 1440 AIMs in linkage equilibrium, and assessed where -2ln likelihood of these models plateaued; based on previous analyses of African Americans we expected that two groups would be sufficient (i.e. European and African).

Global ancestry for each individual was estimated using SNPweights (32) which infers ancestry proportions using genome-wide SNP weights that are pre-computed from external reference panels (Hapmap 3 unrelated ancestry populations). The SNP weights were used to predict principal components (PCs) for the target admixed samples and from them ancestry proportions were estimated for both related and unrelated individuals. The global African ancestry proportion for each genotyped individual was estimated using the genome-wide
131,274 SNPs that are common to both our data and the SNPs used by SNPweights, and these estimates were used to adjust for ancestry in linkage analyses.

**Model-Free Linkage Analysis.** We performed sibpair model-free linkage analysis using SIBPAL, both the original Haseman-Elston regression and W4 regression using the best linear unbiased predictor (BLUP) of the sibship mean (29, 33). P-values <0.05 of the statistic were evaluated by up to 1,000,000 permutation replicates. Using SIBPAL, we performed linkage analysis without the inclusion of any covariates, and with age or DOB and global ancestry as pair-specific covariates, each taken to be the sibpair difference, the sibpair sum, or both the sibpair sum and the difference. We considered permutation \(-\log_{10} p\text{-values} \geq 4\) (equivalent to LOD≥3) as suggestive for linkage significance for our all of our model-free linkage analyses.

We also applied regression-based model-free two-level Haseman-Elston linkage analyses for the breast cancer phenotype using RELPAL (34, 35), which also models trait data from relative pairs as a function of marker allele sharing IBD, but allows for adjusting for covariates at the individual level. This analysis included 180 affected relative pairs. At the first level, individual level global ancestry and age or date of birth were used as covariates. All individuals are used at the first level; all pairs of relatives are used at the second level linkage analysis. We estimated empirical p-values with up to 1,000,000 permutations. Linkage significance was evaluated using the IBD variance, the robust sandwich variance, and the alternative variance.

**Results**

Of the 127 African American families recruited to the study (n=341 women), 281 women (197 affected, 84 unaffected) from 106 families were selected for genotyping based on informativeness. Twenty-seven pedigrees had one affected, 68 pedigrees had 2, and 10 pedigrees had 3 affected individuals.
Our analyses suggest that our data best fits two ancestry groups (Supplementary Figure 1). The average ± SD African proportion in our sample was 0.78±0.13.

When we modeled linkage without including any covariates, we obtained -log_{10} p-values of 4.37 and 3.93 for the original Haseman-Elston method (Figure 2) and W4 with BLUP means, respectively, on chromosome 3. The chromosome 3 peak was also suggestive of significance (-log_{10} p-value=4.08) when using W4 regression with BLUP means and including as covariates date of birth, using the sibpair sum and the difference, and the sibpair difference in global ancestry (Supplementary Figure 2). The other peak we identified using SIBPAL was on chromosome 16. This peak was less consistent across the many models we examined and reached -log_{10} p-value of 3.99 at 83.5 cM when date of birth, the sibpair sum and difference, and the sibpair difference in global ancestry were included in the regression model as covariates (Supplementary Figure 3). Table 1 shows the estimates of mean allele sharing for siblings at the chromosome 3 and 16 peaks that were statistically significant in SIBPAL. The most significant marker on chromosome 3q displayed excess allele sharing in concordantly affected and concordantly unaffected siblings, and less than expected allele sharing in discordant pairs, providing evidence for linkage. The chromosome 16 peak showed a pattern of less than expected allele sharing in both concordantly affected sibling pairs and discordantly affected sibling pairs, which represents conflicting evidence for linkage.

RELPAL estimates for a region on chromosome 12 (109.9 cM) achieved a maximum -log_{10} p-value of 4.29 (Figure 3). The region on chromosome 12 reached 4 for IBD variance and alternative variance methods only; the robust sandwich variance estimate here reached -log_{10} p-value=3.12.

With regard to concordance of SIBPAL and RELPAL results at the chr 3 and 12 peaks, the strongest signal we observed on chromosome 12 using SIBPAL (at the strongest RELPAL peak) was -log_{10} p-value=2.09. The chromosome 3 peak identified as suggestive for linkage in
SIBPAL analyses did not reach our threshold for suggestive significance in RELPAL (all -log_{10} p-values<1.47).

Interestingly, our analyses were suggestive of linkage to chromosome 17 when covariates (date of birth and global ancestry) were included in the model using both sibling-based and affected relative pair-based analyses; original SIBPAL Haseman-Elston estimates ranged from 3.10-3.16, and RELPAL ranged from 1.38 to 3.24 (Supplementary Figure 4).

Discussion

We examined what genes are located under the peaks, considering boundaries where -log_{10} p-values were above 3.0 (approximately equivalent to a 1 LOD drop): chromosome 3 (194.6-200.7 cM), chromosome 12 (107.2-117.8 cM), and chromosome 16 (79.6-87.1 cM). Here, we converted genetic position to physical position using Ensembl release 75, which imports SNP information from release 138 of dbSNP. Using these base pair regions, we identified genes under each linkage peak from hg38 refGene via the UCSC Genome Browser. The chr 3q and 12q peaks do not include genes directly implicated in common cancer phenotypes to date; however SOX2 on 3q codes for a transcription factor involved in embryonic development.

Under our chromosome 16q peak lies CDH1 that codes for E-cadherin, a calcium-dependent adhesion molecule that regulates differentiation, invasion and metastasis. CDH1 was identified as associated with familial gastric cancer via linkage studies (36, 37). This gene is also associated with familial breast (38) and prostate cancer (39). CDH1 is recognized as a high-risk breast cancer gene due to its high penetrance and lifetime risk for mutation carriers, in the range of 40-50% (40). In our sample of 106 families, 10% reported a family history of gastric/stomach cancer and 3% reported a family history of esophageal cancer, suggesting that our sample is segregating a similar hereditary gastric-breast cancer syndrome.
We examined whether published breast cancer GWAS have revealed any common variants that lie under or near our linkage peaks. The Framingham breast cancer GWAS study identified a hit (rs10513754) on chr 3 located in the \textit{LINC00578} gene which codes for a non-protein coding RNA; however this SNP did not reach genome-wide significance.\textsuperscript{(8)} A recent breast cancer GWAS study in African American unrelateds reported association with rs13074711 on 3q26.31, nearby to \textit{TNFSF10},\textsuperscript{(24)} which is near our identified 1 LOD-drop equivalent on 3q26-27. \textit{TNFSF10} codes for a cytokine that is a member of the tumor necrosis factor ligand family, and is implicated in apoptosis. Kim \textit{et al} 2012 reported a marginally significant association with 3q26.32 rs3806685 in a Korean study \textsuperscript{(14)}.

Similarly, Antoniou \textit{et al} 2012 reported associations of breast cancer with polymorphisms on 12q24 (rs1292011) for \textit{BRCA1} and/or \textit{BRCA2} mutation carriers \textsuperscript{(20)}, which is near the region we identified on 12q22-23. This potential overlap on chromosome 12, in conjunction with our suggestive findings of unidentified \textit{BRCA1} mutations in our sample, indicates that these two regions deserve further joint consideration in the family-based context. Investigators recently analyzed whether two GWAS hits on 12q (12q22: rs17356907 \textsuperscript{(41)} and 12q24: rs1292011 \textsuperscript{(42)}) identified in samples of European ancestry replicated in African Americans \textsuperscript{(22)}. For both of these SNPs, the genetic associations originally identified in samples of European ancestry were attenuated in the African American analyses. Collectively, these findings may suggest that rare variants, possibly in combination with other rare variants, may be driving our linkage peak.

While we excluded African American women who were tested and positive for \textit{BRCA1} mutations, we cannot rule out that there are other \textit{BRCA1} mutations segregating in our sample. Nonetheless, this suggestive finding gives us some confidence that the other linkage signals we identified where we adjusted for covariates are less likely to be false positives.

In our sample of families, we were able to adjudicate 72\% (141/197) of the medical records; the reason for missing clinical information was largely due to time limits for facilities in maintaining records. The mean age at diagnosis in our cases included in the linkage study is
51.6 (SD=12.0). Among the women for whom we obtained medical records, 31% were ER negative, 50% were PR negative, and 72% were HER2/neu negative. Taking all three into account resulted in 16% of our sample being triple negative; this approximates the proportion of invasive breast cancers in the U.S. that are triple negative. Interestingly, in our sample of 106 families, only three pedigrees showed full concordance of triple negative status. Further family-based recruitment studies along with full adjudication of tumor phenotype will allow for investigation of differences in linkage signals when taking into account tumor heterogeneity. It is important to note that linkage signals may result from causal alleles up to as far away as five megabases on each side of the peak; therefore we cannot rule out that our signal indicates the presence of causal variants, but not under a 1-LOD drop of the peaks we identified. We also note that there are other candidates on chromosome 17 where \textit{BRCA1} is located, and this includes \textit{HOXB13} (43, 44). This chromosome 17 region deserves additional follow-up to refine our signal.

One limitation of our study is sample size. Future studies developed like ours in partnership with community groups and existing epidemiologic studies with a well-considered recruitment protocol should be conducted to increase the availability of family samples for study. This will not only allow for further refinement and perhaps the identification of novel linkage peaks, but also for replication of GWAS studies to determine to what degree the common GWAS-identified polymorphisms also contribute to familial breast cancer, defined here loosely, as our study is largely limited to families with two affected women.

Since our study was not able to consider genetic heterogeneity due to the limited sample size, larger family-based linkage studies in African Americans should be conducted with the application of more recent developments in linkage methods for complex disease to overcome issues related to disease heterogeneity (45, 46), locus heterogeneity and incomplete penetrance, which can decrease power to detect linkage. Additional family-based recruitment
and linkage analyses in a sample with more homogeneous subsets of families will increase the ability to detect linkage. Furthermore, because the genetic architecture of breast cancer is in itself complicated by pleiotropy, allelic heterogeneity and the polygenic nature of cancer, novel methods are needed to take into account the complexity and architecture of breast cancer in admixed samples. One breast cancer admixture study in African Americans produced no specific regions associated with breast cancer; however there were differences in global ancestry according to ER and PR status and stage of disease (47).

In conclusion, our study is the first to localize novel breast cancer susceptibility loci in pedigrees of African ancestry. We identified three linkage peaks at two regions on chromosomes 3 and 16 and one region on chromosome 12. The new peaks, which may harbor susceptibility loci unique to African Americans, represent loci that deserve additional scrutiny. Based on the overlap we report herein of our putative linkage peaks and GWAS findings, our data support the need to appreciate family-based designs and the linkage approach for complex disease mapping of both rare and common variants. Overall, these preliminary data suggest that there are novel putative regions harboring risk alleles for breast cancer in African American families that deserve further study in larger samples.

Acknowledgements

We extend a special thanks to Veronica Meadows-Ray and her family. Mrs. Meadows-Ray, who insists that her family’s breast cancer is genetic, spurred the development of the Jewels in Our Genes study. We wish to thank the National Witness Project and its Steering Committee for helping us with this work, the Buffalo Witness Project, Rosa Bordonaro, Mary Crawford, Anne Weaver, Youjin Wang, Laurie Grieshober, and others who assisted the Jewels team for their tireless efforts. Others who were instrumental in study recruitment include Greg Ciupak and Drs. Christine Ambrosone and Elisa Bandera (Women’s Circle of Health Study), Dr.
Susan Kadlubar (Spit for the Cure study), Drs. Cheryl Thompson and Li Li (Breast Density Study), Jennie Ellison (GATE/TACT study), and Teri Deans-McFarlane (Witness Project of Harlem). Thank you to the Dr. Susan Love Research Foundation’s Army of Women Program. We also wish to thank all of the families who participated in our study.

This paper is dedicated to Ms. Mattye Willis, who tirelessly helped to recruit families to the study and inspired her community to be champions of scientific research via her role in the National Witness Project and as a special consultant for the Jewels in Our Genes study before her passing in June 2014.
References


35. Wang T, Elston RC. A quantitative linkage score for an association study following a linkage analysis. BMC Genet 2006;7:5.


Table 1. Chromosome regions with the most statistically significant linkage: mean tests and Haseman-Elston regression (in sibling pairs only).

<table>
<thead>
<tr>
<th>Chromosome Location (cM)</th>
<th>Concordantly affected sibling pairs (n=27)</th>
<th>Discordantly affected sibling pairs (n=84)</th>
<th>Concordantly unaffected sibling pairs (n=19)</th>
<th>H-E P valuea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Allele sharing</td>
<td>SE</td>
<td>P value</td>
<td>Allele sharing</td>
</tr>
<tr>
<td>3 (196)</td>
<td>0.587</td>
<td>0.060</td>
<td>0.079</td>
<td>0.415</td>
</tr>
<tr>
<td>16 (83)</td>
<td>0.486</td>
<td>0.068</td>
<td>0.580</td>
<td>0.475</td>
</tr>
</tbody>
</table>

Note: the three mean tests are not adjusted for any covariates. All tests are one-sided as implemented in SIBPAL.

aBased on 1 million replicate permutations.
bHaseman-Elston regression with W4 regression (dependent variable is a weighted combination of sibpair mean-corrected squared trait sum and sibpair squared trait difference), and adjusting for date of birth (pairwise sum and absolute difference) and African ancestry proportion (pairwise absolute difference).
cOriginal Haseman-Elston regression (dependent variable is sibpair squared trait difference), and adjusted for date of birth (pairwise sum and absolute difference) and African ancestry proportion (pairwise absolute difference).
Figure 1. Selection of SNPs for linkage.

Figure 2. Genome-wide model-free multi-point linkage by SIBPAL, P-values based on up to 1,000,000 permutations (Haseman-Elston regression), without adjustment for any covariates.

Figure 3. Model-free multi-point linkage by RELPAL using all affected relative pairs and the IBD variance, adjusting for date of birth and global ancestry at the individual level. P-values evaluated on the basis of up to 1,000,000 permutations.
(1) 528,786 SNPs with call rate > 0.99 and MAF ≥ 0.05

SNP thinning by chromosomes and HWE

(2) 479,452 SNPs on autosomal chromosomes with MAF ≥ 0.05 and P_{HWE} ≥ 0.05

SNP thinning by MAF and interval

(3) 7,990 SNPs with MAF ≥ 0.4 and intervals > 0.333 cM

Inserting most informative SNPs from set (2) so that largest interval < 0.667 cM

(4) 8,261 SNPs for linkage analysis, with average MAF 0.445 (SD = 0.044), and average interval 0.427 cM (SD = 0.094 cM)
Figure 2
Putative linkage signals identified for breast cancer in African American families

Heather M. Ochs-Balcom, Xiangqing Sun, Yanwen Chen, et al.

Cancer Epidemiol Biomarkers Prev Published OnlineFirst December 4, 2014.

Updated version
Access the most recent version of this article at:
doi:10.1158/1055-9965.EPI-14-1131

Supplementary Material
Access the most recent supplemental material at:
http://cebp.aacrjournals.org/content/suppl/2014/12/05/1055-9965.EPI-14-1131.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.