**Retrospective analysis of *The Two Sister Study* using haplotype-based association tests to identify loci associated with early-onset breast cancer**

James R. Gilbert, Ph.D. 1, James J. Cray, Ph.D. 2, Joseph E. Losee, M.D. 1, Gregory M. Cooper, Ph.D. 1,3,4.

1. Department of Plastic Surgery, University of Pittsburgh/Children’s Hospital of Pittsburgh, Pittsburgh, PA 15201.
2. Division of Anatomy, The Ohio State University College of Medicine, Columbus, OH 43210.
3. Department of Oral Biology, University of Pittsburgh/Children’s Hospital of Pittsburgh, Pittsburgh, PA 15201.
4. Department of Bioengineering, University of Pittsburgh/Children’s Hospital of Pittsburgh, Pittsburgh, PA 15201.

*Email addresses*:

James.Gilbert2@chp.edu

James.Cray@osumc.edu

joseph.Losee@chp.edu

Greg.Cooper@chp.edu

*Running Title*:

*Funding*: The work described within this study was funded through the Children’s Fund of Children’s Hospital of Pittsburgh of UPMC and through the Ross H. Musgrave Endowment (J.E.L).

*Conflicts of Interest*: There are no conflicts of interest to disclose.

*Corresponding Author*:

Dr. Gregory M. Cooper

Department of Plastic Surgery

3533 Rangos Research Building

530 45th St, Pittsburgh, PA 15201

412/692-5384 (office)

greg.cooper@chp.edu

*Keywords*: young-onset; early-onset; cancer; familial; breast cancer

**SUPPLEMENTAL INFORMATION**

**SUPPLEMENTAL FIGURE LEGENDS**

**Figure S1. LOESS curve based upon age-at-diagnosis in *The Two Sister Study* breast cancer population.** The optimal cutoff for distinguishing between younger and older patients affected by breast cancer is calculated to be 45 years-of-age at the time of diagnosis.Martingale residuals plotted against age-at-diagnosis suggest that age-at-diagnosis is a discontinuous variable. The assumption of linearity is not fulfilled.

**Figure S2. Quantile-quantile plot.** An initial haplotype-based analysis was performed using a static 10 kb window to identify regions of interest. A QQ plot was constructed by plotting -log[observed p] versus -log[expected p]. The dashed line depicts an ideal line with a slope of 1. Note that individual data points closely follow the theoretical ideal with a slight tail becoming evident at the outermost extremes of the plot.

**Figure S3. Haplomaps representing candidate regions identified by haplotype trend regression with full scan permuted p values ≤ 0.05.** Images include chromosomal position using GRCh38 coordinates. Affected genes are defined beneath each haplomaps and solid black lines represent non-coding introns. Coding exons and their relative position within the displayed haplomaps are represented by solid black blocks. No gene is associated with the block displayed in **Fig S3g** or **Fig S3n**.

**Figure S4. Regression analysis evaluating age-at-diagnosis and candidate gene expression.**  Gene expression data and age-at-diagnosis was obtained from The Cancer Genome Atlas. Regression analysis indicated that expression of *IL2* and *CNNM1* significantly correlated with age-at-diagnosis.

**FIGURES**

**Figure S1. Statistical testing warrants dichotomization of *The Two Sister Study* breast cancer population by age-at-diagnosis.**

****

**Figure S2. Using age-at-diagnosis to dichotomize breast cancer patients yields normally distributed data.**

****

**Figure S3. Mapping of the most significant haploblocks defined by haplotype trend regression provides supporting evidence of underlying haplotype structure within candidate regions.**

****

**Figure S3, continued.**

****

**Figure S4. Regression analysis comparing age-at-diagnosis with gene expression data derived from The Cancer Genome Atlas indicates expression of IL2 and CNNM1 correlate with breast cancer age-of-onset.**

****

**Figure S4 continued**

****