Supplementary Materials

A high-risk haplotype for premature menopause in childhood-cancer survivors exposed to gonadotoxic therapy

Authors: Russell J. Brooke PhD [1], Cindy Im MPH [2], Carmen L. Wilson PhD [1], Matthew J. Krasin MD [1], Qi Liu MSc [2], Zhenghong Li MSc [1], Yadav Sapkota PhD [1], WonJong Moon PhD [1], Lindsay M. Morton PhD [3], Gang Wu PhD [1], Zhaoming Wang PhD [1], Wenan Chen PhD [1], Rebecca M. Howell PhD [4], Gregory T. Armstrong MD [1], Smita Bhatia MD [5], Sogol Mostoufi-Moab MD [6], Kristy Seidel MS [7], Stephen J. Chanock MD [3], Jinghui Zhang PhD [1], Daniel M. Green MD [1], Charles A. Sklar MD [8], Melissa M. Hudson MD [1], Leslie L. Robison PhD [1], Wassim Chemaitilly MD [1], Yutaka Yasui PhD [1].

Affiliations:
(1) St. Jude Children's Research Hospital, Memphis, TN; (2) University of Alberta, Edmonton, AB; (3) National Cancer Institute, National Institutes of Health, Bethesda, MD; (4) The University of Texas MD Anderson Cancer Center, Houston, TX; (5) University of Alabama at Birmingham, Birmingham, AL; (6) The Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; (7) Fred Hutchinson Cancer Research Center, Seattle, WA; (8) Memorial Sloan-Kettering Cancer Center, New York, NY.

Supplementary Methods

Description of the St. Jude Lifetime Cohort study

The eligibility criteria for inclusion in the St. Jude Lifetime ("SJLIFE") Cohort Study included being treated at St. Jude Children's Research Hospital ("SJCRH") for a malignancy, survival greater than or equal to 10 years from diagnosis (recently expanded to five years), and recruitment age greater than or equal to ≥ 18 years of age (recently this age criterion was removed)(31). SJCRH survivors are followed by the SJCRH Cancer Registry and invited to participate in SJLIFE after leaving the After Completion of Therapy ("ACT") Clinic, but attendance of survivors in the ACT Clinic is not a requirement for recruitment into SJLIFE. SJCRH has made long-term commitments to the SJLIFE study and the pool of potential recruits for the study will increase over time as more survivors become eligible for the study. Over time the characteristics of the cohort, including demographics, diagnoses, and treatment, will change reflecting changes in treatment protocols and childhood cancer patients treated at St. Jude(31). The age of the SJLIFE participants in the analysis ranges from 19 to 60 years with a median of 32 years.

The major difference in study design of SJLIFE in comparison to Childhood Cancer Survivor Study ("CCSS") is its clinical assessment of late effects outcomes, compared the self-report ascertainment of CCSS. SJLIFE is a single institution study, while CCSS is a consortium of 31 institutions. CCSS contains a subset of SJLIFE participants: they were excluded from the replication analysis.

Genotyping quality control

Quality control of SJLIFE genotype data was performed using PLINK version 1.90 and excluded SNPs with minor allele frequency ("MAF") <0.01 in the study population, >5% genotype missingness, and SNPs not in Hardy-Weinberg Equilibrium (p-value<10^{-6} in individuals with European ancestry)(14), see Supplementary Figure Genotyping quality control diagram. Exclusion criteria of individuals were: >5% genotype missingness (N=0); cryptic relatedness (N=0); excess per-sample heterozygosity (+/–3 standard deviations from the mean)(N=0); sex discordance between genetically predicted and clinical record (N=11), and ancestry groups with less than five individuals (N=4).

Conditional analysis

We performed a conditional analysis using the 13 SNPs to determine if there were multiple signals in the group of 13 SNPs on Chromosome 4q32.1. Using the clinical model from the genome wide association study, we iteratively chose and added the SNP with the lowest p-value to the clinical model to evaluate the additive impact of SNPs on the clinical model until there were no significant SNPs remaining (cutoff of p-value 0.05). Two SNPs were identified with nominal significance.

Participant ancestry and outliers

To investigate the impact of ancestry and outliers, we repeated the analysis limiting to survivors who are regarded to have European ancestry defined by a cutoff of the STRUCTURE estimated ancestry value 0.5 adjusting for the same treatment covariates as in the main analysis. The European-only analysis showed consistent results as the combined analysis with high-risk

haplotype in ovarian radiation exposed survivors having an odds ratio of 17.33 (95% CI 3.76–99.68, p-value < 0.001). A second analysis in Europeans was performed with a more conservative cutoff of > 0.8 and the high-risk haplotype in ovarian radiation exposed survivors had results similar to the European-only analysis using the 0.5 cutoff (OR 13.97, 95% CI 2.94–81.46, p-value=0.002) adjusting for the same treatment covariates as in the main analysis. To visualize and confirm the ancestry of survivors in our study population against the 1000 Genome populations, we plotted the first two principle components, stratified by cases and controls, superimposed on top of the ancestry groups from the 1000 Genome project for the entire study population, survivors with European ancestry defined by an estimated ancestry value > 0.5, and survivors with European ancestry defined more conservatively with an estimated ancestry value > 0.8, see **Supplementary Figure 4**, **5**, and **6**, respectively.

Analysis of imputed genotypes

We imputed genotypes from the Affymetrix 6.0 array up to the 1000 Genomes Phase 3 Version 5 mixed reference panel using the University of Michigan Imputation Server(15). Common autosomal SNPs (MAF ≥ 0.01 in study sample) were imputed and included in the imputed analysis if they had an imputation quality score ("INFO") greater or equal to 0.4. There were 11,343,365 imputed SNPs with 10,993,255 having an imputed quality score greater than or equal to 0.4 used in the imputed analysis. We then performed a genome-wide association study of imputed genotypes (dosage scores) adjusting for the same non-genetic co-variates in the clinical model as we did for the analysis of directly genotyped SNPs, see **Method**s section Study Design and Participants. Of the imputed autosomal SNPs analyzed, none reached genome wide significance (p-value < 5.0×10^{-8}) or refined the observed signal motivating our focus on

genotyped data and specifically the region with the largest cluster of genotyped SNPs with p-values < 10^{-5}.

Haplotype analysis

Current standard methodologies for SNP data involve measuring genotypes of SNPs without distinguishing between the maternal or paternal chromosomal origin of alleles. The standard analysis evaluates each single SNP one at a time for its genotype's association with the phenotype, without considering other SNPs. When considering more than one SNP in proximity, we need to account for the chromosome on which a set of bases (SNP alleles) reside because transcription reads from a (either maternal or paternal) chromosome. A haplotype is a DNA sequence (not necessarily adjacent to each other) on the same chromosome. It is, therefore, meaningful to examine a haplotype when investigating multiple SNPs that tend to be inherited together. When we create a haplotype from multiple SNPs we need to obtain phased data which, by statistical estimation, separates proximal SNP alleles into two chromosomes. This allows investigation of the association between the haplotype and a phenotype of interest.

To determine whether the observed genetic signal could be better captured using multiple SNPs to form a haplotype, we calculated the log likelihood for all single SNPs, all two SNP haplotypes, all three SNP haplotypes, and all four SNP haplotypes using phased data obtained with PHASE(16). The log likelihood decreases with the addition of each additional SNP with the four SNP haplotype having the best performing model, which we used to define our risk haplotype.

- 1 SNP (rs4323056:A): -64.0
- 2 SNP (rs7669884:C AND rs9999820:G): -61.7

- 3 SNP (rs7669884:C AND rs4323056:A AND rs9999820:G): -61.1
- 4 SNP (rs4323056:A AND rs13114936:G AND rs4402990:C AND rs9999820:G): -60.8

Treatment associations

Previously reported treatment variables that significantly increased the risk of premature menopause ("PM") in survivors, including ovarian-radiotherapy ("RT") exposure (RT >10 Gray versus no RT, OR 109.59, 95% CI 28.15–426.70) and alkylating agents (upper tertile alkylating agent score versus no exposure, OR 5.78 (95% CI 2.90–11.55)(6), were included in our clinical base model. In our analysis, the clinical base model without the homozygous risk haplotype found that the radiation exposure indicator variable (yes/no) had an odds ratio of 11.28 (95% CI 3.84–34.66, p-value $1.2x10^{-5}$) and radiation exposure dosimetry (one Gray increase) having an OR of 1.07 (95% CI 1.00–1.15, p-value 0.04) for the prevalence of PM. The association of cyclophosphamide equivalent dose indicator variable (CED > $8g/m^2$, yes/no) with the PM prevalence became weaker with an odds ratio of 2.66 (95% CI 1.08–6.90, p-value 0.04). In the model with an indicator variable for the homozygous high-risk haplotype, radiation exposure remains significantly associated with PM prevalence with the radiation exposure indicator variable (yes/no) having an OR of 8.79 (95% CI 1.87–47.89, p-value 0.007) and radiation exposure dosimetry (one Gray increase) having an OR of 1.09 (95% CI 1.01–1.18, p-value 0.03). The association of cyclophosphamide equivalent dose indicator variable (CED > $8g/m^2$, yes/no) with PM prevalence became weaker with an OR of 2.87 (95% CI 0.92–9.90, p-value 0.08).

Bioinformatics analysis

Data tracks for the predicted chromatin state (ChromHMM) and histone modification mark peaks from ENCODE ChIP-seq experiments (chromatin immunoprecipitation combined with massive parallel sequencing) associated with enhancer (H3K4me1), promoter (H3K4me3), and Polycomb-repressed (H3K27me3; H3K9me3) states were considered for relevant tissue types. Epigenetic data across multiple bioinformatics resources were compiled to characterize the SNPs in the expanded genetic signal ("expanded GS"). Significant associations between SNPs in the expanded GS and cis-gene expressions from the Genotype-Tissue Expression Project ("GTEx") were assessed(26). HaploReg was used to identify SNPs that overlapped with: (a) enhancer-related ChIP-seq histone modification mark peaks (H3K4me1, H3K27ac); (b) DNase I hypersensitivity site peaks; and (c) transcription factor ("TF") binding sites motifs with significant alterations between SNP alleles. ANNOVAR was employed to identify SNPs with evidence of bound transcription factors (ENCODE ChIP-seq data for 161 TFs) and conservation across 46 vertebrate species (ENCODE 46-way PhastCons data)(32).

Temporal trends associated with premature menopause in St. Jude Lifetime Cohort study

We performed two supplementary analyses, adjusting for the same ancestry and treatment exposures as in the main analysis.

Using the Cox proportional hazards model for survivors with ovarian radiation exposure, the homozygous high-risk haplotype was associated with premature menopause with an adjusted hazard ratio of 9.10 (95% CI 3.58–23.12, p-value=3.5×10^{-6}) with the time at risk of study participants starting at eligibility for inclusion in SJLIFE (verified the proportional hazards assumption with scaled Shoenfeld residuals graphically and testing).

In a matched case-control analysis where cases were matched based on age at clinical assessment (+/- two years) and ancestry, the conditional logistic regression model showed that, in survivors with ovarian radiation exposure, females with the homozygous high-risk haplotype had an increased odds of premature menopause prevalence (OR 14.78, 95% CI 4.25–51.34, p-value $2.3x10^{-5}$).

Note that since the clinical assessment cannot identify the exact age/time at premature menopause, our primary analysis assessed, not the incidence of premature menopause, but the prevalence of having had premature menopause by the age at clinical assessment.

Replication dataset AUC, sensitivity, and specificity

In the replication dataset, adding the homozygous high-risk haplotype to the non-genetic model of premature menopause increased the AUC from 0.66 to 0.71 in survivors with ovarian radiation exposure, with a sensitivity of 0.29 (95% CI 0.15–0.46) and a specificity of 0.89 (95% CI 0.84–0.92).

References

1 Armstrong GT, Chen Y, Yasui Y, et al. Reduction in late mortality among 5-year survivors of childhood cancer. *N Engl J Med* 2016;374:833–842.

2 Howlader N, Noone AM, Krapcho M, et al. *SEER cancer statistics review, 1975-2014, National Cancer Institute*. Bethesda, MD, https://seer.cancer.gov/csr/1975_2014/, based on November 2016 SEER data submission, posted to the SEER web site, April 2017.

3 Hudson MM, Ness KK, Gurney JG, et al. Clinical ascertainment of health outcomes among adults treated for childhood cancer. *JAMA* 2013 Jun 12;309(22):2371–2381. doi: 10.1001/jama.2013.6296.

4 Bhakta N, Liu Q, Ness KK, et al. The cumulative burden of surviving childhood cancer: an initial report from the St. Jude Lifetime Cohort study. *Lancet* (in press).

5 Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. *Nat Rev Cancer* 2014 Jan;14(1):61–70. doi: 10.1038/nrc3634.

6 Sklar CA, Mertens AC, Mitby P, et al. Premature menopause in survivors of childhood cancer: a report from the childhood cancer survivor study. *J Natl Cancer Inst* 2006 Jul 5;98(13):890–896.

7 Chemaitilly W, Mertens AC, Mitby P, et al. Acute ovarian failure in the childhood cancer survivor study. *J Clin Endocrinol Metab* 2006 May;91(5):1723–1728.

8 Anderson RA, Mitchell RT, Kelsey TW, et al. Cancer treatment and gonadal function: experimental and established strategies for fertility preservation in children and young adults. *Lancet Diabetes Endocrinol* 2015 Jul;3(7):556–567. doi: 10.1016/S2213-8587(15)00039-X.

9 Hudson MM, Ehrhardt MJ, Bhakta N, et al. Approach for classification and severity grading of long-term and late-onset health events among childhood cancer survivors in the St.

Jude Lifetime Cohort. *Cancer Epidemiol Biomarkers Prev* 2017 May;26(5):666–674. doi: 10.1158/1055-9965.EPI-16-0812.

10 Chemaitilly W, Li Z, Krasin MJ, et al. Premature ovarian insufficiency in childhood cancer survivors: a report from the St. Jude Lifetime Cohort. *J Clin Endocrinol Metab* 2017 Mar 24. doi: 10.1210/jc.2016–3723.

11 Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015 Feb 25;4:7. doi: 10.1186/s13742-015-0047-8.

12 Green DM, Nolan VG, Goodman PJ, at al. The cyclophosphamide equivalent dose as an approach for quantifying alkylating agent exposure: a report from the Childhood Cancer Survivor Study. *Pediatr Blood Cancer* 2014 Jan;61(1):53–67. doi: 10.1002/pbc.24679.

13 Hubisz MJ, Falush D, Stephens M, et al. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* 2009 Sep;9(5):1322–1332. doi: 10.1111/j.1755-0998.2009.02591.x.

14 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015 Oct 1;526(7571):68–74. doi: 10.1038/nature15393.

15 Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016 Oct;48(10):1284–1287. doi: 10.1038/ng.3656.

16 Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003 Nov;73(5):1162–1169.

17 Good, Ph.: Permutation Tests. *A practical guide to resampling methods for testing hypotheses*. Springer Series in Statistics, Springer-Verlag, Berlin — Heidelberg — New York: 1995, x, 228 pp., DM 74,00; õS 577.20; sFr 74.–. ISBN 3-540-94097-9

18 Wallace WH, Smith AG, Kelsey TW, et al. Fertility preservation for girls and young women with cancer: population-based validation of criteria for ovarian tissue cryopreservation. *Lancet Oncol* 2014 Sep;15(10):1129–1136.

19 Robison LL, Armstrong GT, Boice JD, et al. The Childhood Cancer Survivor Study: a National Cancer Institute-supported resource for outcome and intervention research. *J Clin Oncol* 2009 May 10;27(14):2308–2318. doi: 10.1200/JCO.2009.22.3339.

20 Mostoufi-Moab S, Seidel K, Leisenring WM, et al. Endocrine Abnormalities in Aging Survivors of Childhood Cancer: A report From the Childhood Cancer Survivor Study. *J Clin Oncol* 2016 Sep 20;34(27):3240–3247. doi: 10.1200/JCO.2016.66.6545.

21 Morton LM, Sampson JN, Armstrong GT, et al. Genome-wide association study to identify susceptibility loci that modify radiation-related risk for breast cancer after childhood cancer. *J Natl Cancer Inst* 2017;109(11): djx058

22 Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* 2016 Jan 4;44(D1):D877–881. doi: 10.1093/nar/gkv1340.

23 Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015 Feb 19;518(7539):317–330. doi: 10.1038/nature14248.

24 R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, accessed 30 July 2016.

25 Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005 Jan 15;21(2):263–265. PMID: 15297300

26 GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015 May 8;348(6235):648–660.

27 Fan HY, Liu Z, Shimada M, et al. MAPK3/1 (ERK1/2) in ovarian granulosa cells are essential for female fertility. *Science* 2009 May 15;324(5929):938–941. doi: 10.1126/science.1171396.

28 Kuo LE, Kitlinska JB, Tilan JU, et al. Neuropeptide Y acts directly in the periphery on fat tissue and mediates stress-induced obesity and metabolic syndrome. *Nat Med* 2007 Jul;13(7):803–811.

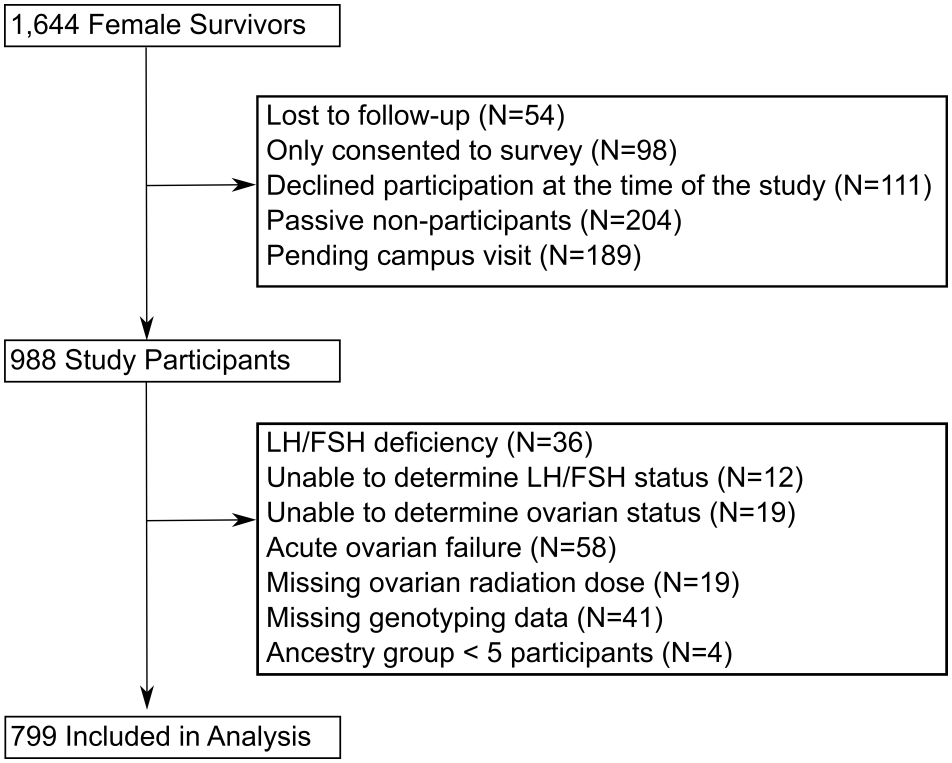
29 Xu M, Hill JW, Levine JE. Attenuation of luteinizing hormone surges in neuropeptide Y knockout mice. *Neuroendocrinology* 2000 Nov;72(5):263–271.

30 Day FR, Ruth KS, Thompson DJ et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat Genet* 2015 Nov;47(11):1294-1303.

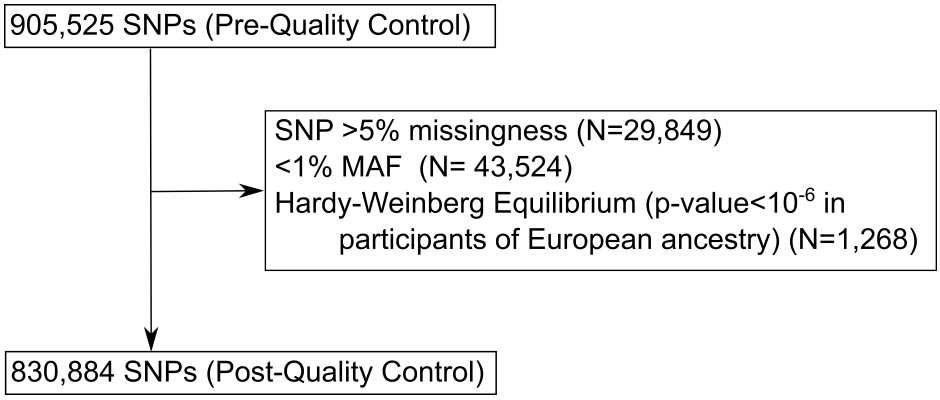
31 Hudson MM, Ness KK, Nolan VG, Armstrong GT, Green DM, Morris EB, Spunt SL, Metzger ML, Krull KR, Klosky JL, Srivastava DK, Robison LL. Prospective medical assessment of adults surviving childhood cancer: study design, cohort characteristics, and feasibility of the St. Jude Lifetime Cohort study. *Pediatr Blood Cancer*. 2011 May;56(5):825-36. doi: 10.1002/pbc.22875.

32 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010 Sep 1;38(16):e164. doi: 10.1093/nar/gkq603.

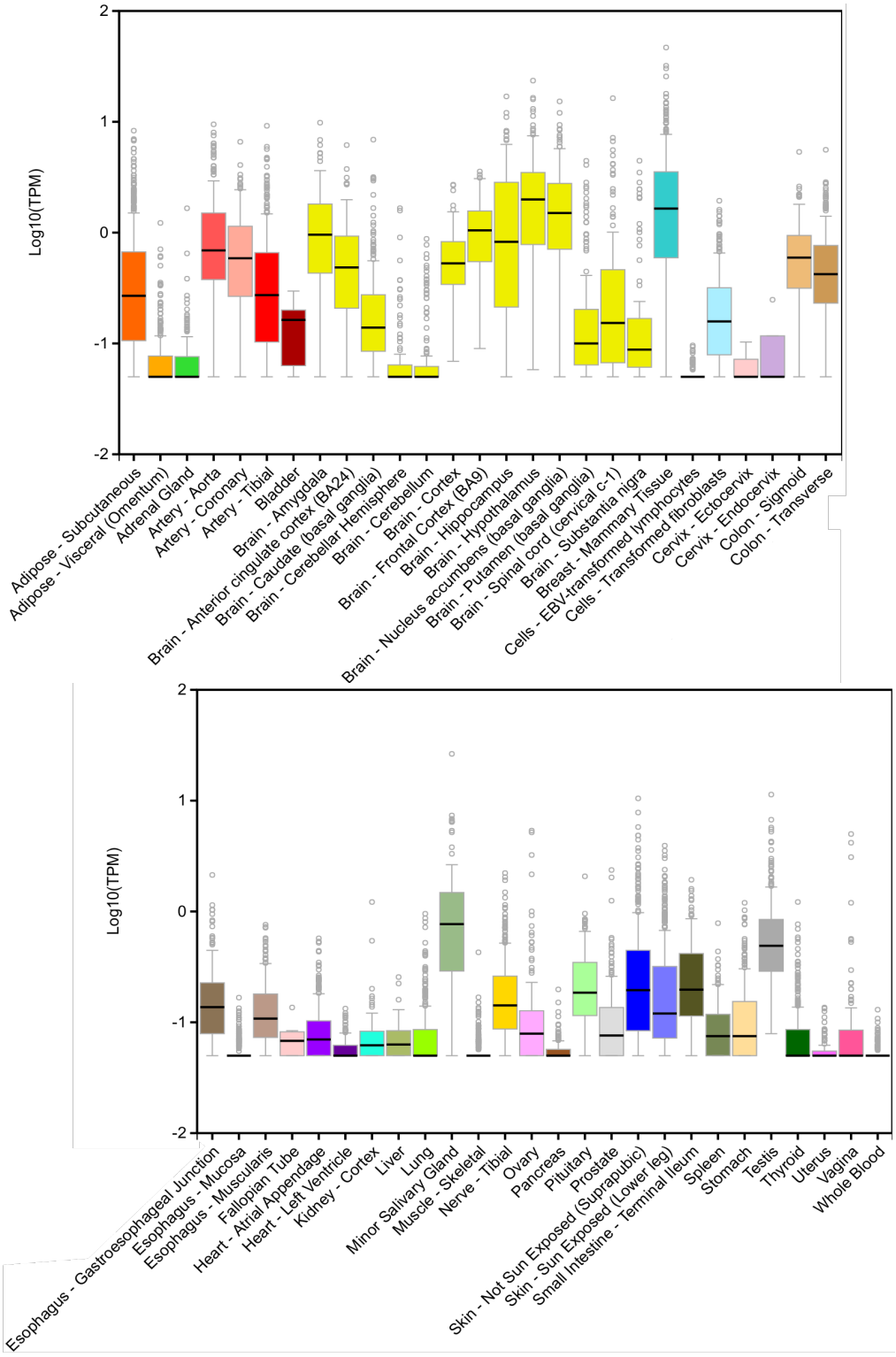
Supplementary Figures



Supplementary Figure 1. Consort diagram for St. Jude Lifetime ("SJLIFE") Cohort Study.

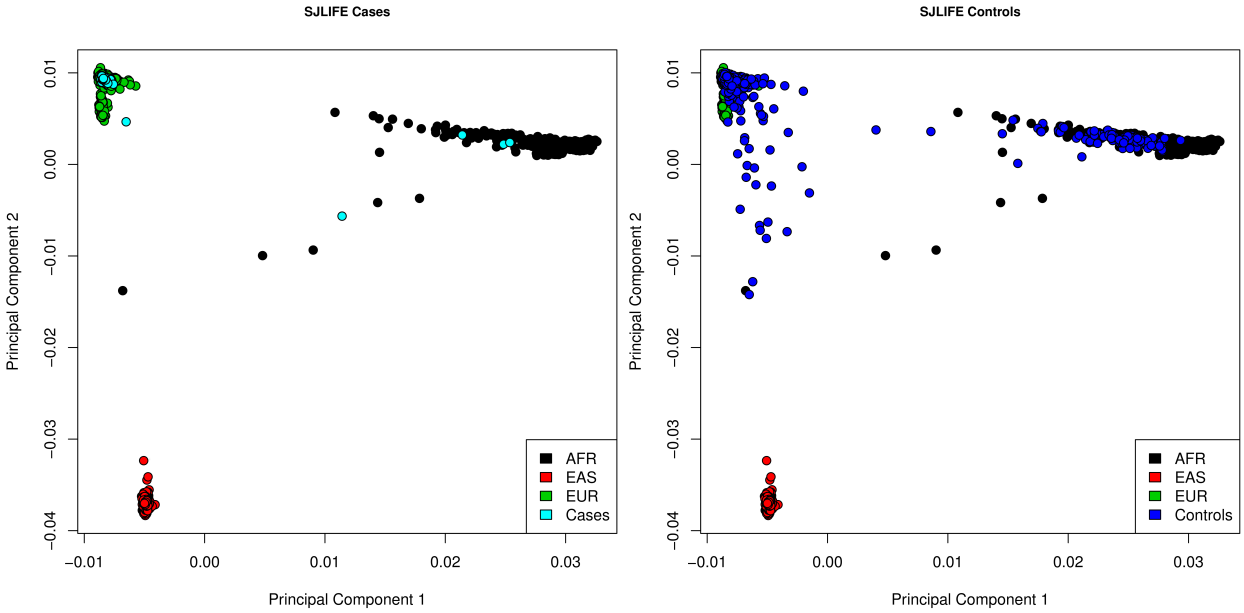


Supplementary Figure 2. Genetic quality control exclusion diagram.

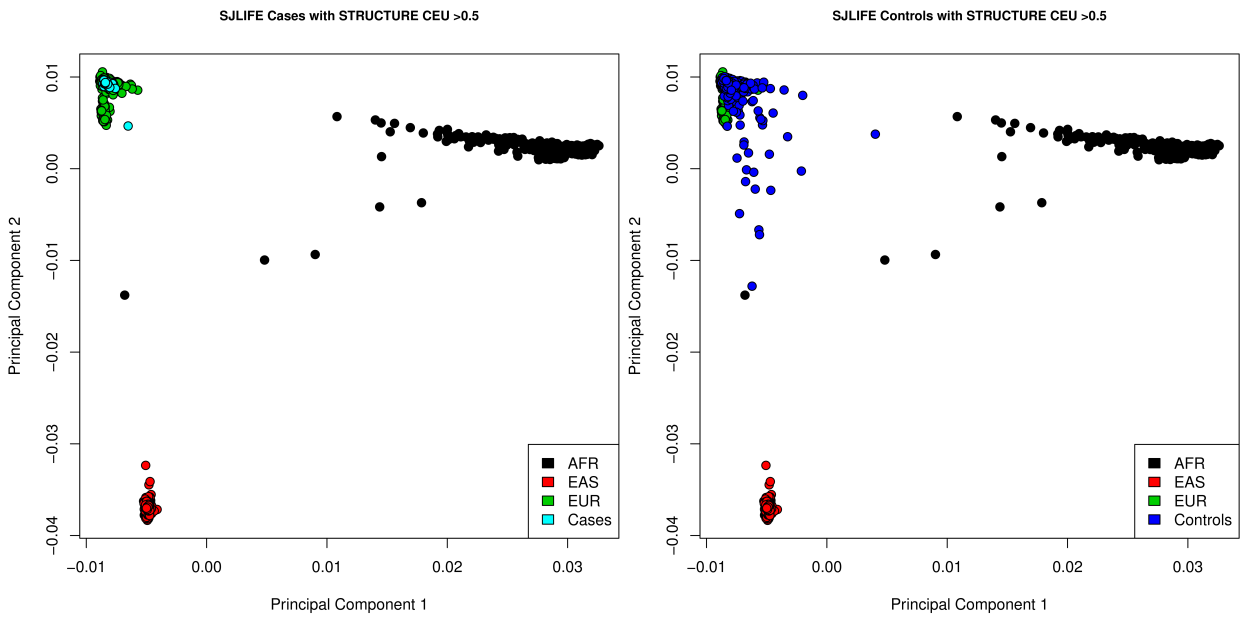


Supplementary Figure 3. GTEx Project normalized gene expression plot showing median and 25th/75th percentile of $\log_{10}$ Reads Per Kilobase of transcript per Million mapped reads for *NPY2R* among tissues with N ≥ 70 samples indicating

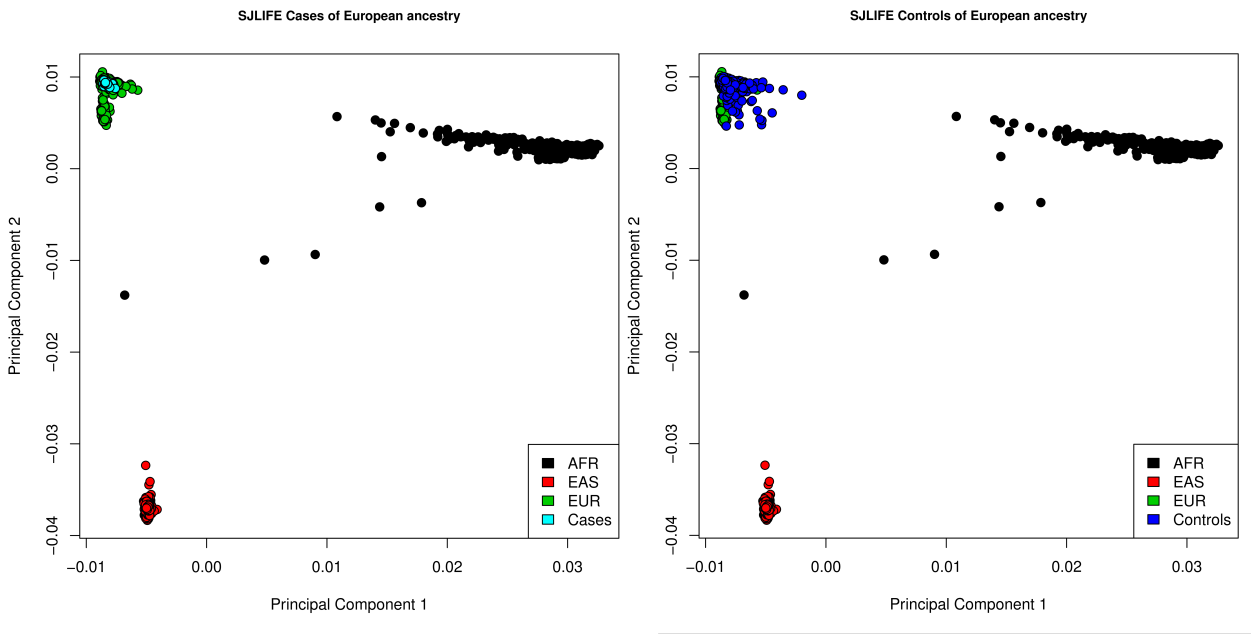
the greatest expression in the hypothalamus (top: 27 types, bottom: 26 tissue types)(26).



Supplementary Figure 4. Plot of the first two principle components from principle component analysis for cases and controls in SJLIFE superimposed on the 1000 Genomes EUR/AFR/EAS reference populations.



Supplementary Figure 5. Plot of the first two principle components from principle component analysis for cases and controls of European descent (using STRUCTURE CEU variable > 0.5) in SJLIFE superimposed on the 1000 Genomes EUR/AFR/EAS reference populations.



Supplementary Figure 6. Plot of the first two principle components from principle component analysis for cases and controls of European descent (using STRUCTURE CEU variable > 0.8) in SJLIFE superimposed on the 1000 Genomes EUR/AFR/EAS reference populations.

Supplementary Table 1. Conditional analysis top SNPs iteratively adding genotyped SNPs to the clinical model until no additional significant SNPs remain with a p-value cutoff of 0.05

One SNP		Two SNPs		Three SNPs	
SNP	**P-value***	**SNP**	**P-value***	**SNP**	**P-value***
rs9999820	3.25×10^{-7}	rs9999820 +		rs9999820 +	
rs4323056	3.49×10^{-7}	rs13114936	0.04	rs13114936 +	
rs6810505	9.46×10^{-7}	rs6810505	0.07	rs4323056	0.53
rs12643129	9.84×10^{-7}	rs2880418	0.07	rs4402990	0.53
rs2880418	1.45×10^{-6}	rs4323056	0.08	rs4456917	0.53
rs13114936	2.03×10^{-6}	rs13121931	0.08	rs11099988	0.53
rs7669884	4.01×10^{-6}	rs12643129	0.09	rs4428241	0.53
rs13121931	5.08×10^{-6}	rs7669884	0.09	rs6810505	0.59
rs11735253	5.68×10^{-6}	rs4402990	0.14	rs12643129	0.64
rs4402990	8.24×10^{-6}	rs4456917	0.14	rs7669884	0.70
rs4456917	8.31×10^{-6}	rs11099988	0.14	rs13121931	0.78
rs11099988	8.32×10^{-6}	rs4428241	0.14	rs11735253	0.88
rs4428241	8.34×10^{-6}	rs11735253	0.43	rs2880418	0.95

*Two-sided likelihood ratio test.

Supplementary Table 2. The Edinburg Criteria for prioritizing fertility saving procedures (35 years of age with >50% risk of premature menopause) was applied using the clinical model with and without the high-risk haplotype to predict prevalence of premature menopause in the SJLIFE study population

	Premature menopause predicted risk	
Clinical model	**Risk ≥50% (Number of observed PM cases in SJLIFE thus far)**	**Risk <50% (Number of observed PM cases in SJLIFE thus far)**
With high-risk haplotype	15 (9)	784 (21)
Without high-risk haplotype	1 (1)	798 (29)