SUPPLEMENTARY MATERIAL

Supplementary Methods

Study population

Participants were identified from the ongoing large prospective cohort studies Nurses' Health Study (NHS) and NHSII, both of which have been followed biennially by mailed questionnaire to update information on exposure status and to ascertain newly diagnosed diseases, including cancers. All women reporting incident diagnoses of breast cancer were asked for permission to review their medical records; for cases for which pathology reports were obtained, cases were confirmed by medical record review (>99%). Details of the selection of breast cancer patients have been described previously[1,2]. Briefly, invasive breast cancer cases with sufficient RNA from formalin-fixed paraffin-embedded (FFPE) tumor blocks for transcriptomic profiling were included in this study. Archived FFPE breast tumor blocks were obtained from the cohort tumor tissue repository; tumor tissue block collection has been described previously[3]. Tumor estrogen receptor (ER) expression status was obtained from tissue microarrays[1,2,4].

Assessment of early-life body size and other covariates

Information on early-life body size and covariates was obtained from NHS and NHSII questionnaires. In 1988 (NHS) and 1989 (NHSII), women were asked to select the figure from a validated 9-figure drawing (Supplementary Figure 1) that best corresponded to their body size at ages 5, 10, and 20, respectively[5,6]. Body size level 1 represents the leanest and level 9 the most obese. Women in each cohort complete biennial questionnaires that provide detailed information on demographic, important breast cancer risk factors, lifestyle, and medical history. Covariate data included in this analysis, such as first degree family history of breast cancer, alcohol consumption and physical activity were obtained from the NHS or NHSII questionnaire at baseline and subsequent biennial questionnaires; for body mass index (BMI), menopausal

status and menopausal hormone use, the information from the most recent questionnaire prior to diagnosis was used.

Gene expression microarray and quality control analysis

RNA extraction and transcriptomic profiling have been described in detail previously[1,2,4]. Briefly, RNA was extracted from multiple 1 or 1.5mm cores taken from tumor or tumor-adjacent histologically-normal tissues from FFPE blocks using the Qiagen AllPrep RNA isolation kit (Qiagen, Valencia, CA). Since FFPE samples are known to have variable yields, tissue from all the cores from the same patient were placed into one microtube to maximize RNA yield. Tumor-adjacent histologically-normal tissue was generally greater than 1 cm from the tumor edge, though a minimum of 2 mm between tumor and tumor-adjacent was permitted. Gene expression profiling was done in two batches during 2012-2014 and 2015-2018 using two types of microarray chips: Glue Grant Human Transcriptome Arrays (HTA) 3.0 prerelease version (Affymetrix, Santa Clara)[2,4] and HTA 2.0[1]. Gene expression data were normalized and summarized into log2 values using Robust Multiarray Average (Affymetrix Power Tools v1.18.0). Sample quality was evaluated using Affymetrix Power Tools probeset summarization-based metrics, including the area under the curve (AUC); samples with AUC <0.55 were excluded from the analysis. We further excluded samples that failed the non-outlier analysis by arrayQualityMetrics v3.24.0[7]. A total of 835 tumors and 663 tumor-adjacent tissue samples were included in this analysis.

We further assessed biological concordance (i.e. probe expression concordance with protein markers measured by immunohistochemistry [IHC]) for select probes. We confirmed the correlation between probes for *ESR1, PGR, and ERBB2* with IHC markers, ER, progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), in tumors to confirm biological reproducibility of the data[2].

Supplementary Table 1. Tumor characteristics by early-life body size.

Participant characteristics	Shape 1 (N=91)	Shape 1.5-2 (N=261)	Shape 2.5-3 (N=259)	Shape 3.5-4 (N=164)	Shape ≥4.5 (N=60)
Tumor stage					
I	58 (64%)	141 (54%)	159 (61%)	106 (65%)	38 (63%)
II	23 (25%)	92 (35%)	77 (30%)	45 (28%)	20 (33%)
III	10 (11%)	26 (10%)	22 (9%)	9 (6%)	2 (3%)
III	--	2 (1%)	1 (0%)	3 (2%)	--
Tumor grade					
1	22 (24%)	58 (23%)	57 (22%)	52 (32%)	10 (16%)
2	53 (58%)	130 (52%)	128 (50%)	71 (43%)	40 (68%)
3	14 (16%)	56 (22%)	65 (25%)	35 (21%)	10 (16%)
4	2 (2%)	8 (3%)	6 (2%)	5 (3%)	--

Supplementary Table 2. Results of differential gene expression analysis by early-life body size in ER+ tumor, ER+ tumor-adjacent, ER- tumor, and ER- tumor-adjacent, respectively. Complete lists of results of all the 13,343 genes are shown in a separate excel sheet.

Supplementary Table 3. Significantly[1] up or down regulated gene sets by early-life body size in all tumor combined or all tumor-adjacent tissue.

All tumor combined (N=835)				
Pathway name	Number of genes	Direction	p-value	FDR
HALLMARK_MYC_TARGETS_V1	187	Down	1.23E-18	6.17E-17
HALLMARK_E2F_TARGETS	147	Down	6.10E-11	1.52E-09
HALLMARK_MTORC1_SIGNALING	165	Down	5.05E-08	8.41E-07
HALLMARK_G2M_CHECKPOINT	149	Down	1.51E-07	1.89E-06
HALLMARK_ALLOGRAFT_REJECTION	150	Down	1.58E-06	1.58E-05
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	105	Down	2.38E-06	1.99E-05
HALLMARK_INTERFERON_ALPHA_RESPONSE	76	Down	4.93E-06	3.52E-05
HALLMARK_OXIDATIVE_PHOSPHORYLATION	185	Down	1.40E-05	8.74E-05
HALLMARK_INTERFERON_GAMMA_RESPONSE	153	Down	2.27E-05	1.26E-04
HALLMARK_PROTEIN_SECRETION	83	Down	3.75E-04	0.002
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	174	Up	0.001	0.005
HALLMARK_PI3K_AKT_MTOR_SIGNALING	92	Down	0.001	0.005
HALLMARK_DNA_REPAIR	129	Down	0.005	0.020
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	42	Down	0.008	0.029
HALLMARK_MYOGENESIS	178	Up	0.011	0.038
HALLMARK_MITOTIC_SPINDLE	166	Down	0.015	0.047
All tumor-adjacent (N=663)				
Pathway name	Number of genes	Direction	p-value	FDR
HALLMARK_INTERFERON_GAMMA_RESPONSE	153	Down	2.44E-05	0.001

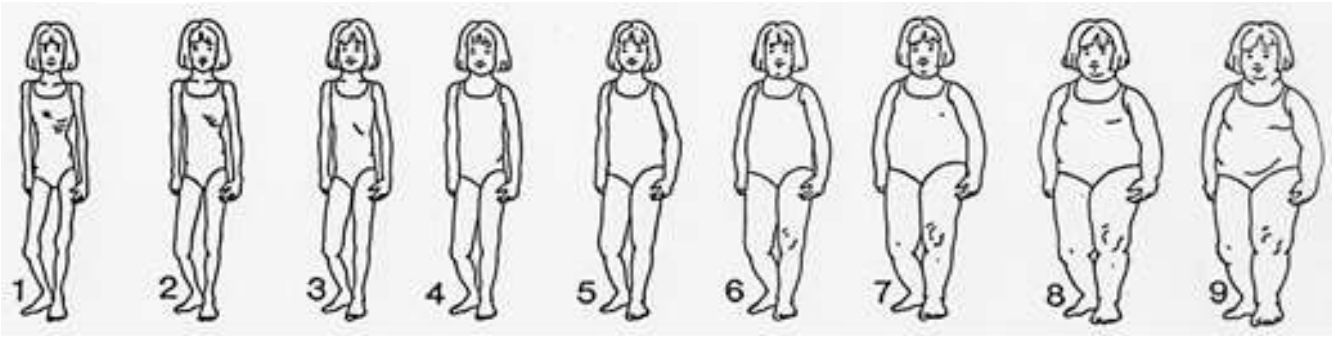
[1]Only gene sets with FDR < 0.05 were presented.
[2]Number of genes that contributed to the enrichment of the gene set in this dataset.

Supplementary Table 4. Immunohistochemical analysis of Ki67 and cytokeratin in breast tumors by early-life body size.

	IHC Results by Somatotype								
	Shape 1		Shape 1.5-2		Shape 2.5-3		Shape 3.5+		P
	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	0.79
KI67*	40	13.6 (16.2)	111	13.6 (13.9)	94	11.7 (10.8)	77	13.1 (16.2)	
	N	%	N	%	N	%	N	%	
CK5/6†									0.12
0	64	83.1	178	82.8	168	80.8	154	89.5	
1	13	16.9	37	17.2	40	19.2	18	10.5	
CK5/14‡									0.21
0	6	54.5	34	68.0	28	60.9	26	81.3	
1	5	45.5	16	32.0	18	39.1	6	18.7	
CK7/18‡									0.73
0	3	3.9	6	2.7	9	4.2	9	4.8	
1	73	96.1	215	97.3	206	95.8	179	95.2	

* Ki67 expression defined as continuous.
† CK5/6 expression status defined as: 0 = no staining and 1 = any positivity
‡ CK5/14 and CK7/18 expression status defined as: 0 = no staining and 1 = weak/moderate/strong staining

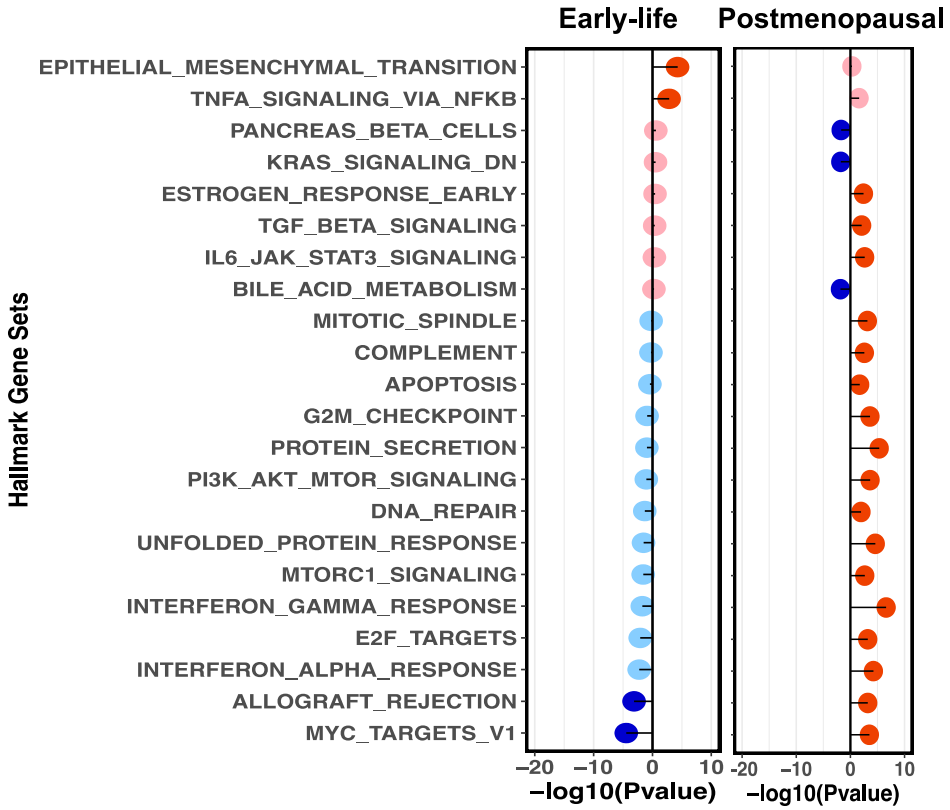
Supplementary Figure 1. Figure drawing (9-level pictogram) used to assess body shape at ages 10 and 20 years in the Nurses' Health Study and Nurses' Health Study II.



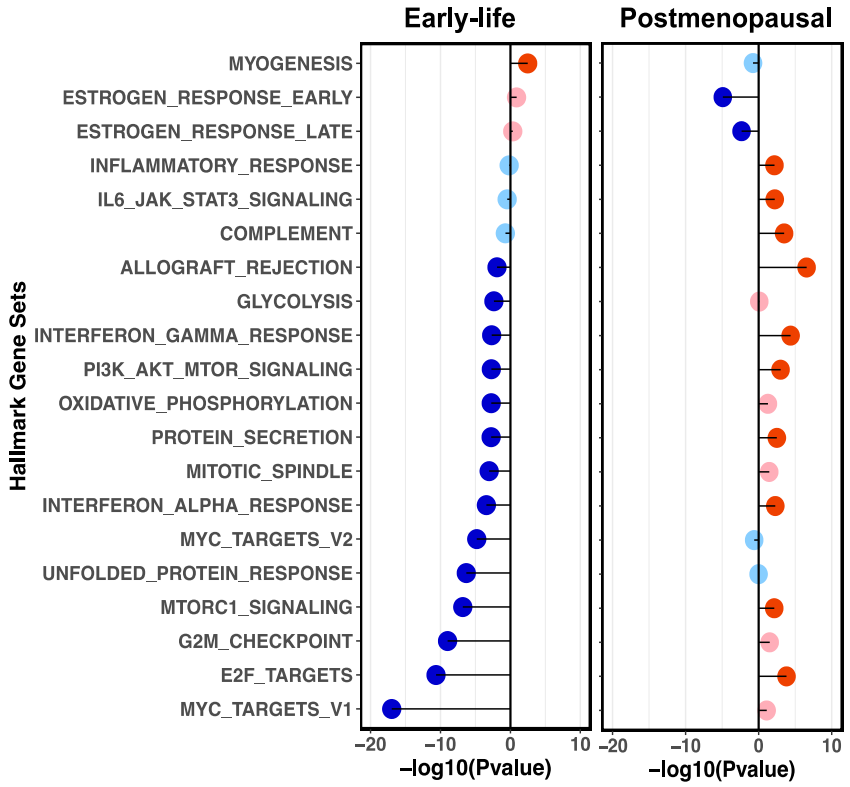
Supplementary Figure 2. Comparison of significantly[1] up- or down- regulated Hallmark gene sets found in analysis of early-life body size vs. postmenopausal BMI at breast cancer diagnosis in A). ER+ tumors and B). ER- tumors, respectively.

- Up-regulated (FDR <0.05)
- UP-regulated (FDR >0.05)
- Down-regulated (FDR <0.05)
- Down-regulated (FDR >0.05)

A) ER+ Tumor



B) ER- Tumor



[1]Only gene sets that were significantly (FDR <0.05) up- or down-regulated in analysis of early-life body size and/or postmenopausal BMI at diagnosis are presented (i.e., each gene set in the figure was found significantly up- or down- regulated in either early-life body size or postmenopausal BMI at diagnosis, or in both).
Up-regulated gene sets are denoted by -log10(*P*value) >0 and down-regulated gene sets are denoted with -log10(*P*value)<0.

References

1. Kensler KH, Sankar VN, Wang J, et al. PAM50 Molecular Intrinsic Subtypes in the Nurses' Health Study Cohorts. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* 2019;28(4):798-806.
2. Wang J, Heng YJ, Eliassen AH, et al. Alcohol consumption and breast tumor gene expression. *Breast cancer research : BCR.* 2017;19(1):108.
3. Tamimi RM, Baer HJ, Marotti J, et al. Comparison of molecular phenotypes of ductal carcinoma in situ and invasive breast cancer. *Breast cancer research : BCR.* 2008;10(4):R67.
4. Heng YJ, Wang J, Ahearn TU, et al. Molecular mechanisms linking high body mass index to breast cancer etiology in post-menopausal breast tumor and tumor-adjacent tissues. *Breast cancer research and treatment.* 2019;173(3):667-677.
5. Sorensen TI, Stunkard AJ. Does obesity run in families because of genes? An adoption study using silhouettes as a measure of obesity. *Acta psychiatrica Scandinavica Supplementum.* 1993;370:67-72.
6. Stunkard AJ, Sorensen T, Schulsinger F. Use of the Danish Adoption Register for the study of obesity and thinness. *Research publications - Association for Research in Nervous and Mental Disease.* 1983;60:115-120.
7. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics (Oxford, England).* 2009;25(3):415-416.