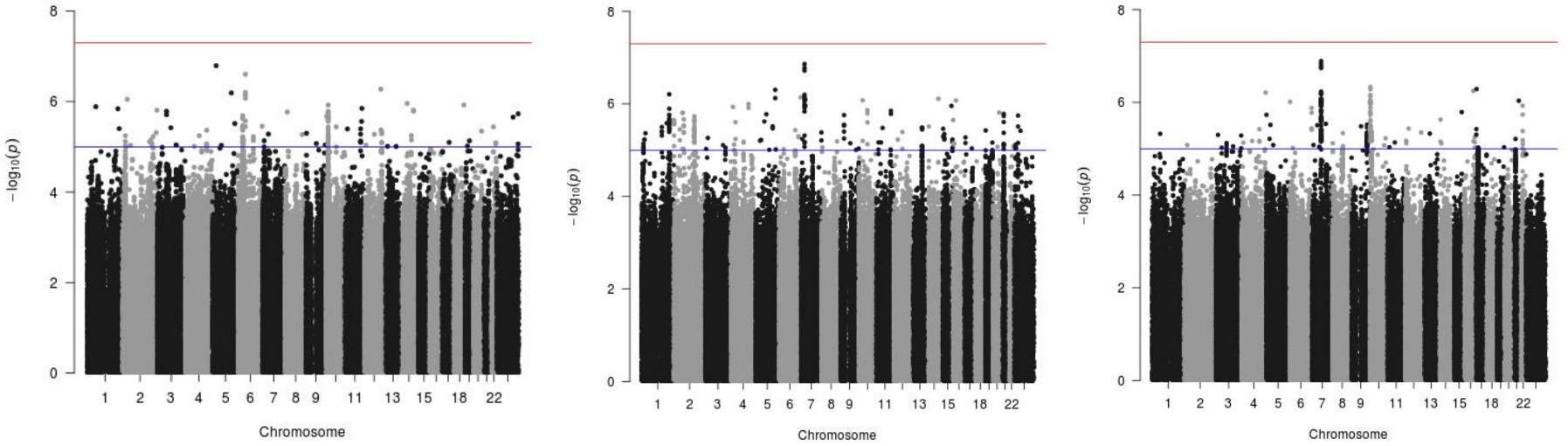


Figure 1. Association plot for the meta-analysis of the twelve datasets for breast cancer-specific mortality analyses (censored at 15 years) for A) all breast tumours (censored at 15 years), B) ER-negative tumours and C) ER-positive tumours. The y axis shows the $-\log_{10}P$ values of each variant analysed, and the x axis shows their chromosome position. The red horizontal line represents $P=5x10^{-8}$.

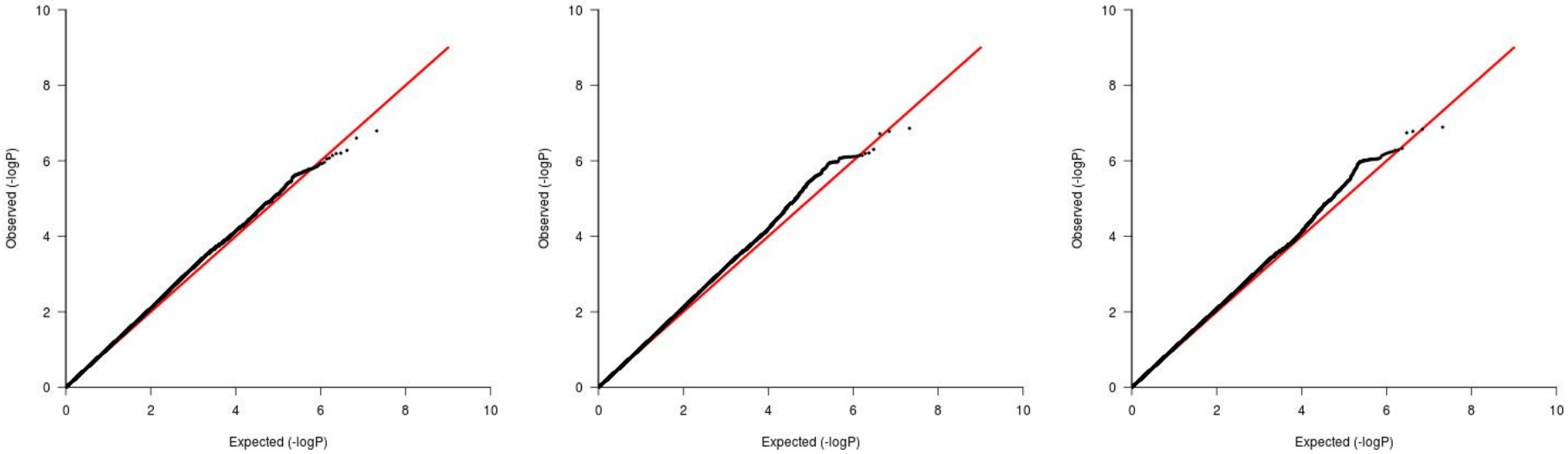


Figure 2. Q-Q plots for the meta-analysis of the twelve datasets for breast cancer-specific mortality analyses (censored at 15 years) for A) all breast cancer tumours (censored at 15 years), B) ER-negative tumours and C) ER-positive tumours. The y axis represents the observed $-\log_{10}P$ value, and the x axis represents the expected $-\log_{10}P$ value. The red line represents the expected distribution under the null hypothesis of no association. Analyses were not corrected for LD-structure.

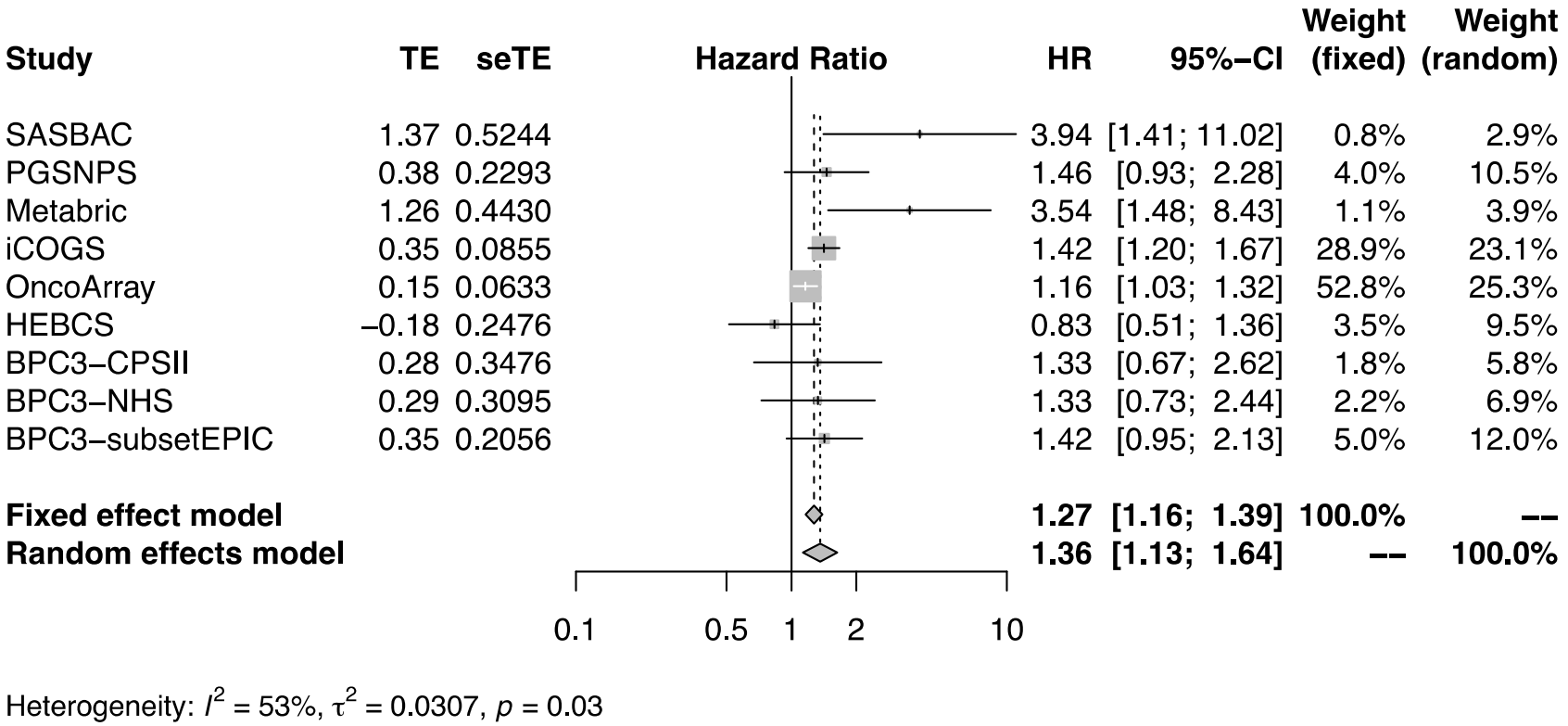


Figure 3. Forest plot showing the association between the ER-negative variant rs67918676 and breast cancer specific-mortality in ER-negative tumours for the datasets used in the meta-analysis. The size of the square reflects the size of the study (see also Supplementary Table 3).

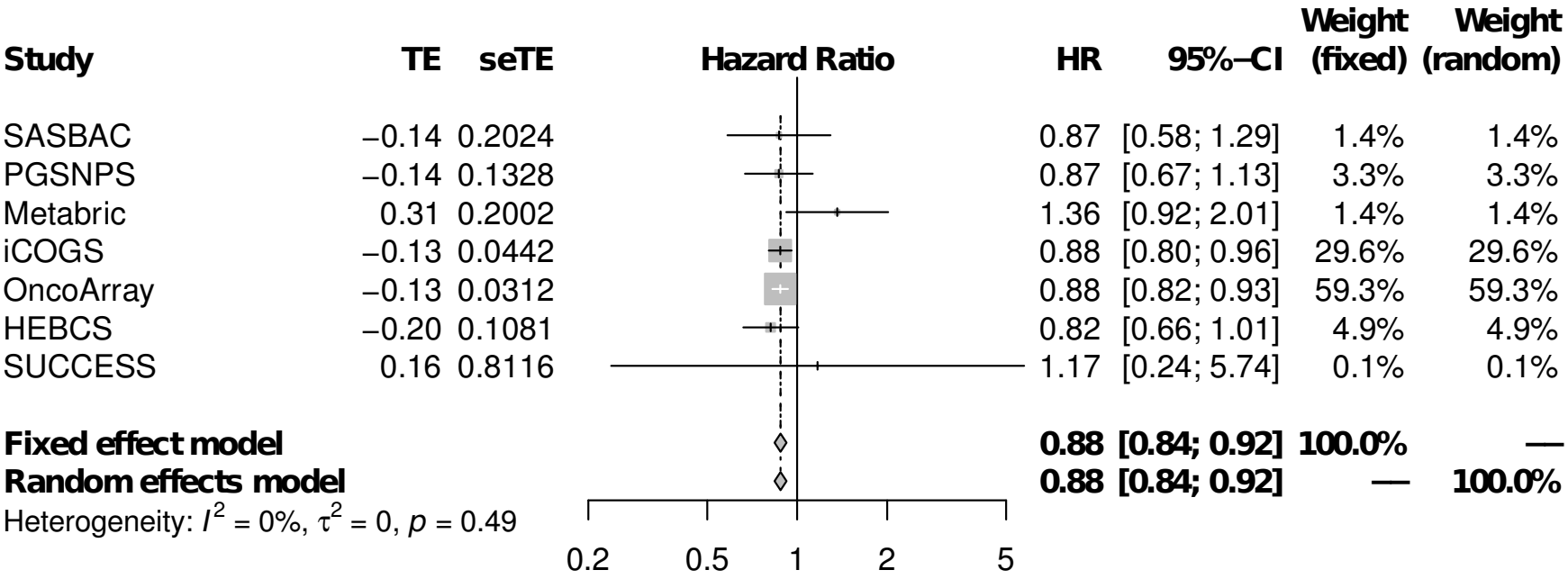


Figure 4. Forest plot showing the association between the ER-positive variant rs4717568 and breast cancer-specific mortality in ER-positive tumours for the datasets used in the meta-analysis. The size of the square reflects the size of the study (see also Supplementary Table 3).

Supplementary Data

List of Tables

List of Figures

Supplementary Table 1. Description of the twelve datasets by genotype platform

Study	Genotyping platform	Number of principal components*	Age at diagnosis in years	Sample size	Country	Description of the study
OncoArray	Illumina OncoArray-500K BeadChip	10	18-98	54,798	International	Breast cancer patients from 62 studies of the Breast Cancer Association Consortium (BCAC), genotyped as part of the OncoArray Consortium[1]
COGS	Illumina iSelect	9	16-96	29,959	International	Breast cancer patients from 34 studies of the BCAC, genotyped as part of the Collaborative Oncological Gene-Environment Study (COGS)[2]
CGEMS	Illumina Hap550K	0	44-83	1,145	USA	Breast cancer patients from Nurses Health Study genotyped as part of CGEMS project[3]
SASBAC	Illumina HumanHap300 and HumanHap240S	0	50-75	787	Sweden	Breast cancer patients from Swedish Case-control study, part of BCAC[4]
UK2	Illumina 670k	3	17-88	2,763	UK	Consist of National study of breast cancer of age < 41 years, and Subset of samples from national familial breast cancer study[5]
Metabric	Affymetrix SNP 6.0	1	26-96	369	UK	UK samples from international breast cancer genomics project[6]
PG-SNPs	Affymetrix SNP 6.0	2	22-77	1,797	UK	UK samples from breast cancer chemotherapy treatment response study[7–10]
HEBCS	Illumina 550K	0	26-87	742	Finland	Helsinki Breast Cancer Study[11–14]
SUCCESS-A	Illumina HumanOmniExpress-12v1 FFPE	0	19-85	3,229	Germany	A Genome-Wide Association Study in Breast Cancer Patients From the Prospectively Randomized SUCCESS Trial[15]
BPC3-CPSII	Illumina 660W	0	51-89	293	USA	The National Cancer Institute Breast and Prostate Cancer Cohort Consortium: American Cancer Society Cancer Prevention Study-II[16]
BPC3-EPIC	Illumina 660W	0	27-75	476	Europe	The National Cancer Institute Breast and Prostate Cancer Cohort Consortium: European Prospective Investigation of Cancer[16]
BPC3-NHS2	Illumina 550K	0	44-83	233	USA	The National Cancer Institute Breast and Prostate Cancer Cohort Consortium: Nurses' Health Studies II[16]

*Number of principal components included in the survival analyses to correct for population structure[17]

Supplementary Table 3. Follow-up and ER-status description for the twelve datasets

Dataset	All tumours		ER-positive		ER-negative	
	N (breast cancer deaths)	Person-years	N (breast cancer deaths)	Person-years	N (breast cancer deaths)	Person-years
OncoArray	54,798 (3,632)	346,059	38,685 (2,159)	233,298	8,424 (1,048)	52,507
COGS	29,959 (2,643)	196,439	20,249 (1,441)	132,676	4,775 (614)	29,790
CGEMS	1,145 (93)	7,711	--	--	--	--
SASBAC	787 (89)	4,133	483 (53)	2,539	108 (15)	551
UK2	2,763 (305)	29,664	--	--	--	--
Metabric	369 (88)	1,582	291 (61)	1,280	63 (25)	225
PG-SNPs	1,797 (211)	5,884	1,192 (122)	3,957	591 (87)	1,906
HEBCS	742 (321)	5,652	492 (197)	4,202	196 (106)	1,214
SUCCESS-A	3,299 (175)	13,145	2264 (83)	9,289	1,013 (90)	3,806
BPC3-CPSII	293 (30)	2,544	--	--	293 (30)	2,544
BPC3-EPIC	476 (74)	2,226	--	--	476 (74)	2,226
BPC3-NHS2	233 (36)	2,732	--	--	233 (36)	2,732
Total	**96,661 (7,697)**	**622,404**	**64,171 (4,116)**	**424,377**	**16,172 (2,125)**	**133,365**

ER=estrogen receptor

Supplementary Table 5A: Association of survival with gene expression of nearby genes at the chr7q21.1 locus

Study Group	***HOXA9*** (209905_at)	***HOXA10-AS*** (231365_at)	***HOXA10*** (213150_at)	***HOXA11*** (213823_at)	***HOXA11-AS*** (230666_at)	***HOXA13*** (231786_at)	***HOTTIP*** (244553_at)	***EVX1*** (207914_x_at)	***HIBADH*** (234452_at)	***TAX1BP1*** (213786_at)
All	1.12 (1.00-1.25), p=0.059	0.90 (0.77-1.06), p=0.2	0.88 (0.79-0.99), p=0.032	0.88 (0.79-0.99), p=0.032	0.81 (0.69-0.95), p=0.0097	1.18 (1.01-1.39), p=0.042	0.79 (0.67-0.92), p=0.0036	0.76 (0.68-0.85), p=2.4x10E-6	0.93 (0.79-1.09), p=0.36	1.21 (1.08-1.36), p=0.00098
ER-negative (GE-based)	1.38 (1.11-1.72), **p=0.0042**	1.16 (0.88-1.53), p=0.29	1.04 (0.83-1.29), p=0.74	0.94 (0.75-1.17), p=0.58	0.93 (0.70-1.22), p=0.58	1.10 (0.84-1.45), p=0.5	0.67 (0.51-0.89), **p=0.0047**	0.68 (0.54-0.85), **p=0.00065**	1.00 (0.76-1.32), p=0.99	1.41 (1.13-1.76), **p=0.0023**
ER-positive (GE-based)	1.04 (0.91-1.18), p=0.6	0.77 (0.63-0.94), p=0.0091	0.84 (0.74-0.96), p=0.012	0.87 (0.76-1.00), p=0.042	0.73 (0.60-0.89), p=0.0018	1.17 (0.96-1.43), p=0.12	0.82 (0.67-1.00), p=0.046	0.80 (0.70-0.92), p=0.0012	0.89 (0.73-1.09), p=0.25	1.15 (1.00-1.31), p=0.045

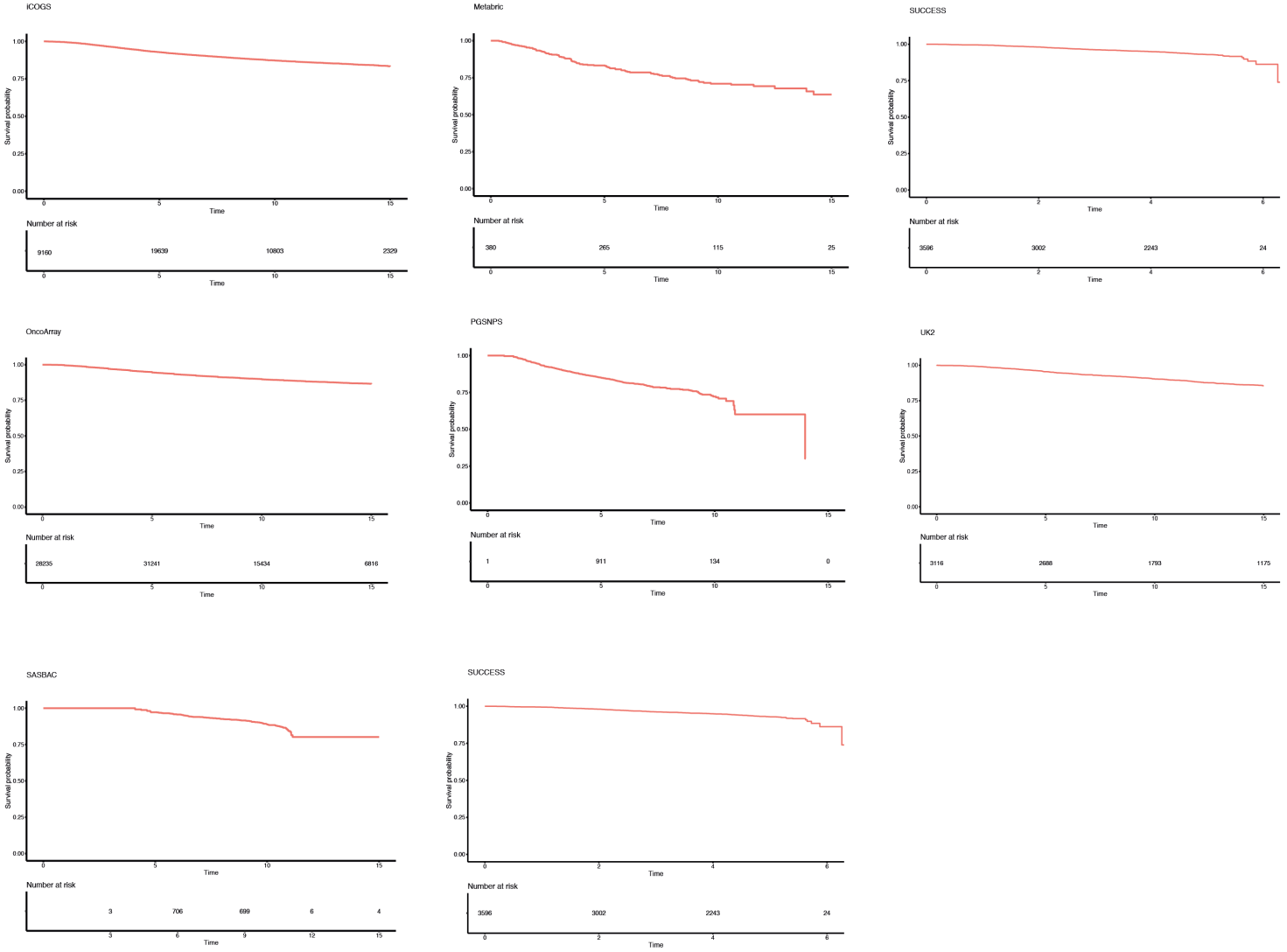
Supplementary Table 5B: Association of survival with gene expression of nearby genes at the chr7q11.22 locus

Study Group	***AUTS2*** (243364_at)	***GALNT17*** (227434_at)
Overall	0.89 (0.76-1.04), p=0.13	0.94 (0.81-1.10), p=0.45
ER-negative (GE-based)	0.81 (0.62-1.06), p=0.12	1.17 (0.94-1.52), p=0.24
ER-positive (GE-based)	0.87 (0.65-1.16), p=0.35	0.84 (0.70-1.02), p=0.08

Legend: Genes with available probes within a 500 MBp window centred at the identified set of highly correlated variants and were tested for the association of their mRNA/ncRNA expression in breast tumours with recurrence-free survival using KMplotter (kmplot.com/analysis). The headers of the columns indicate the genes tested and the gene expression probes used. Association analyses were performed for all breast cancer, ER-negative breast cancer and ER-positive breast cancer, with ER status based on gene expression. Hazard ratios are shown with 95% confidence intervals. A. Results for chr7q21.1 locus, p-values<0.005 for ER-negative breast cancer are indicated in bold. B. Results for chr7q11.22 locus in ER-positive breast cancer. ER=estrogen receptor; GE=gene expression.

Supplementary Figure 1. Kaplan-Meier Curves for the cohorts used in the meta-analysis*. The survival probabilities are obtained from Kaplan-

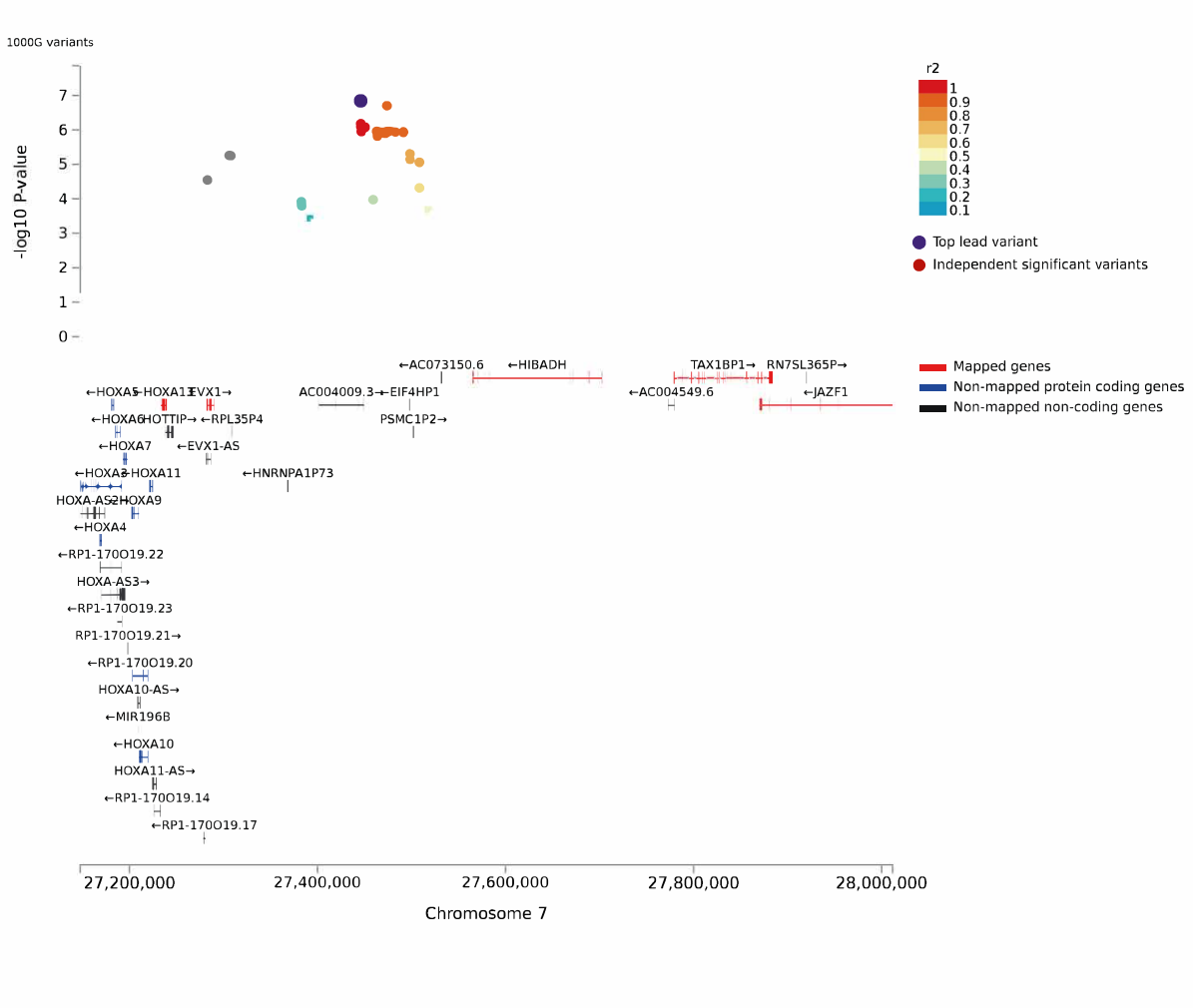
Meier estimates.



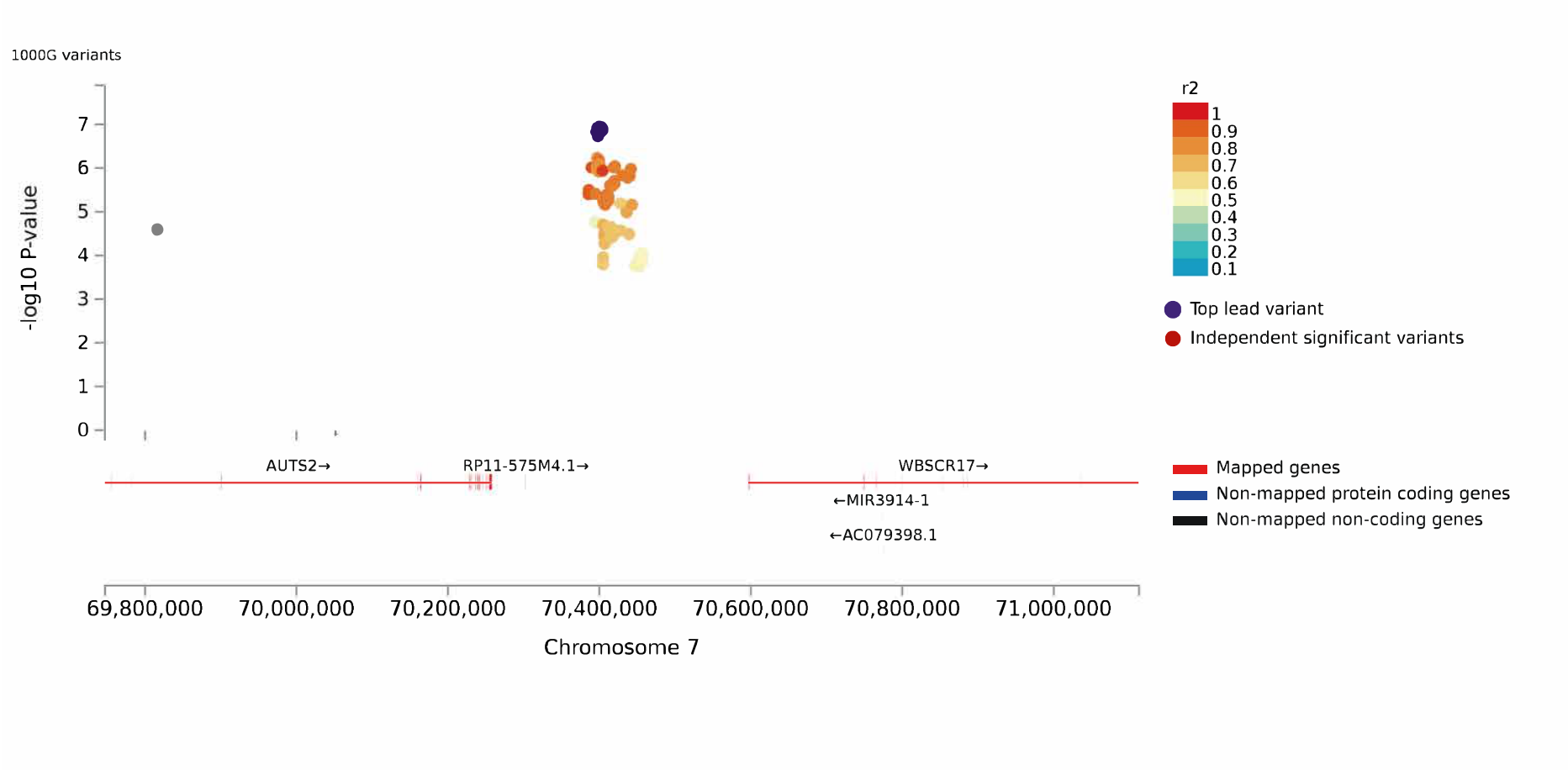
* Not have raw survival data for the CGEMS, HEBCS and BPC3 studies.

Supplementary Figure 2. Linkage disequilibrium matrix for the ER-negative variants associated with breast cancer-specific mortality in ER-

negative tumours.



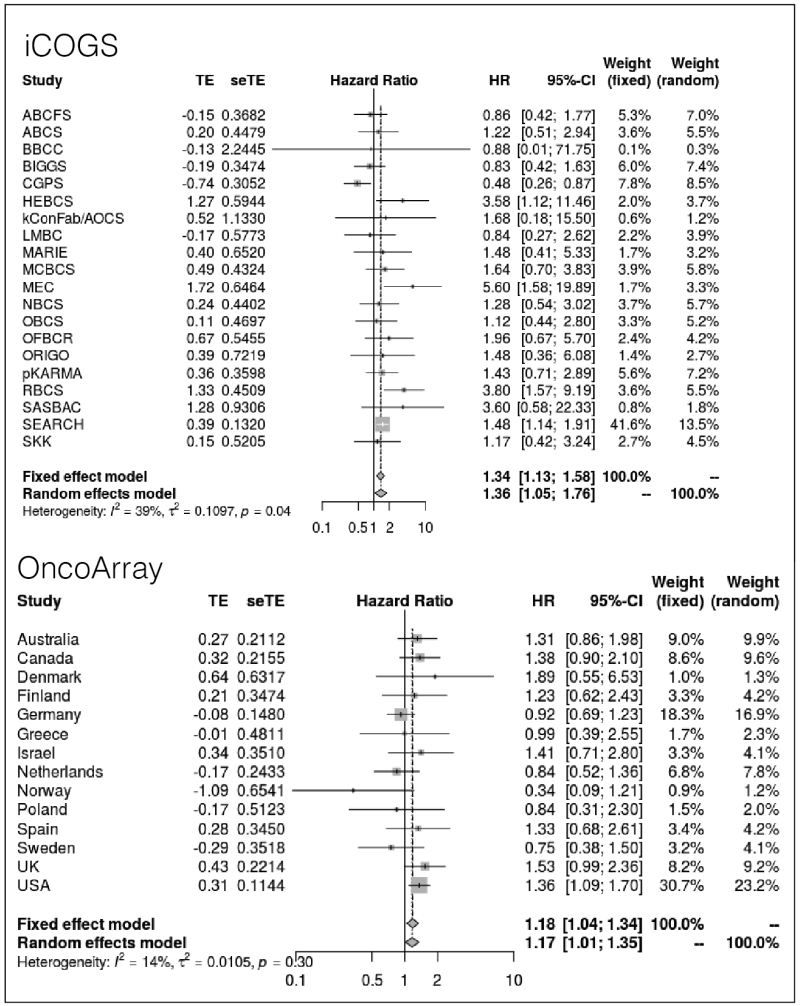
Supplementary Figure 3. Linkage disequilibrium matrix for the ER-positive variants associated with breast cancer-specific mortality in ER-positive tumours



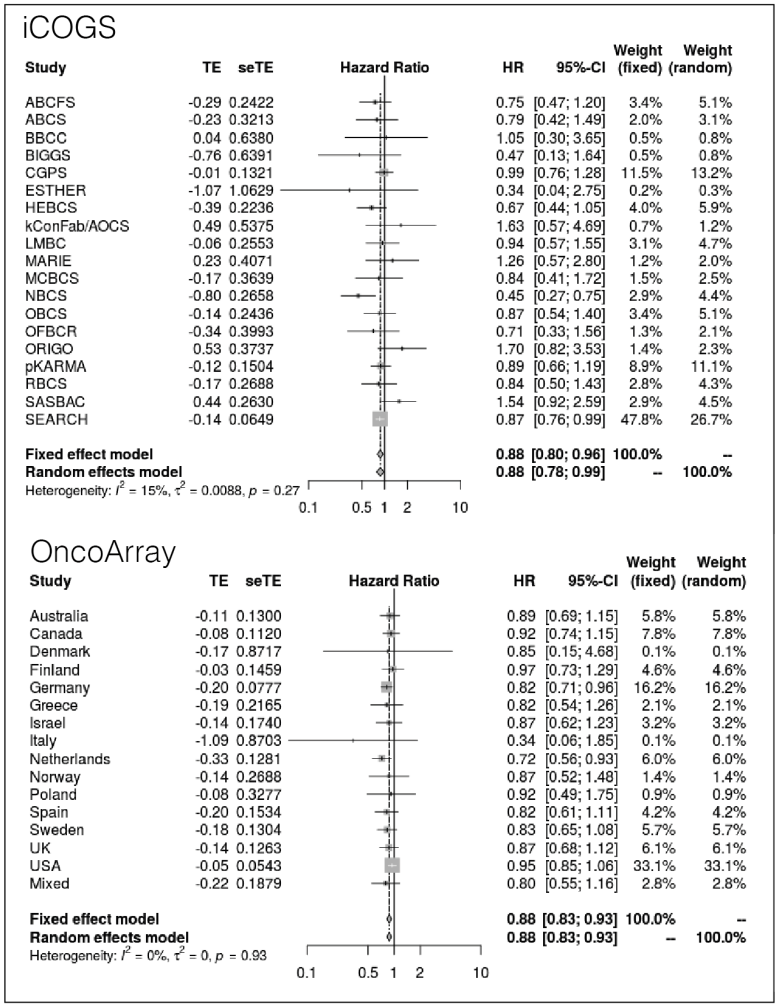
Supplementary Figure 4. Forest plots showing the association between breast cancer specific-mortality and the most significant variant in ER-

negative and ER-positive disease for iCOGS and OncoArray separately: by study in iCOGS and by country in OncoArray. (a) ER-negative most significant variant: rs67918676. (b) ER-positive most significant variant: rs4717568. The size of the square reflects the size of the study.

a) ER-negative most significant variant: rs67918676



b) ER-positive most significant variant: rs4717568



Supplementary methods

Derivation of the Standard Errors (SE) for maximum-likelihood estimates using Likelihood Ratio Test Statistics.

The likelihood ratio test (LRT) and the Wald test are identical for normally distributed data (where the log-likelihood is quadratic) and also asymptotically equivalent (i.e. approximately equivalent in large samples; see for example Ch.9 of Cox and Hinkley's *Theoretical Statistics* (1974)[67]). In finite samples the LRT is generally preferable – in particular it has better properties for association tests based on rare counts, and we have therefore used LRT and maximum-likelihood estimates of all datasets.

In the meta-analysis the test statistic is of the form:

$$\frac{\sum Z_j \sqrt{w_j}}{\sqrt{\sum w_j}}$$

and the summary estimate is of the form:

$$\frac{\sum w_j \theta_j}{\sum w_i}$$

where θ_j are the study-specific estimates, Z_j are the study-specific signed test statistics, and w_j are study-specific weights. The meta-analysis is valid whatever weights are used but the most efficient test uses inverse variance weights, and hence requires approximate standard errors for the parameter estimates. To derive approximate standard-errors for the maximum likelihood estimates, we can use the asymptotic equivalence of the LRT and the Wald test.

The likelihood ratio statistic (LRT) is of the form:

$$W = 2\{l(\hat{\theta}; Y) - l(\theta_0; Y)\}$$

where θ_0 is the null value of the parameter (0 in this case), $\hat{\theta}$ is the maximum likelihood estimate (MLE) and $l(\theta_0; Y)$ is the log-likelihood given the data *Y*.

Expanding in a Taylor's series (Cox and Hinkley's pp. 313):

$$W = 2(\hat{\theta} - \theta_0) U(\hat{\theta}) - (\hat{\theta} - \theta_0)^2 U' + \text{¿} \quad (1)$$

where U refers to the first derivative of the log-likelihood and U ' the second derivative, since by definition $U(\hat{\theta})=0$ at the maximum-likelihood,

+¿ $-\theta_0$

where $\theta^{¿}$.

¿

¿

Since, asymptotically, $\hat{\theta}$ is consistent, +¿ $\theta^{¿}$ ¿ U' ¿

The variance of $\hat{\theta}$ is, asymptotically, $1/i(\theta)=-1/E(U^{'}(\theta);\theta)$, that is the inverse of the information matrix (Cox and Hinkley pp. 294). Hence this can be estimated by $-1/U^{'}(\hat{\theta})$ (minus the second derivative of the log-likelihood at the MLE).

Combining with (1), the variance (and hence standard error) of the maximum likelihood estimate $\hat{\theta}$, can thus be estimated, using the likelihood ratio test statistic W, by:

$$var(\hat{\theta})=(\hat{\theta}-\theta_0)^2/W$$

Table 1. Results of the variants with BFDP<15% in the meta-analysis of the twelve studies of breast cancer-specific mortality.

Subgroup	Variant	Chr	Position	Alt	Ref	Eaf_Ref	HR	LCL	UCL	P-value	BFDP
ER-negative	rs67918676:27445956:A:AT	7	27445956	AT	A	0.12	1.27	1.16	1.39	$1.38x10^{-7}$	0.11
ER-negative	rs192185001:27448012:A:AT	7	27448012	AT	A	0.12	1.27	1.16	1.39	$1.66x10^{-7}$	0.13
ER-negative	rs145963877:27473909:CAG:C	7	27473909	C	CAG	0.11	1.28	1.17	1.41	$1.91x10^{-7}$	0.15
ER-positive	rs4717568:70400700:T:C	7	70400700	C	T	0.62	0.88	0.8	0.92	$1.28x10^{-7}$	0.07
ER-positive	rs1917618:70396442:T:A	7	70396442	A	T	0.62	0.88	0.84	0.93	$1.46x10^{-7}$	0.08
ER-positive	rs1546774:70398441:T:G	7	70398441	G	T	0.62	0.88	0.84	0.93	$1.66x10^{-7}$	0.09
ER-positive	rs1546773:70398437:T:C	7	70398437	C	T	0.62	0.88	0.84	0.93	$1.81x10^{-7}$	0.10
All	rs370332736:50395136:AACTT:A	6	50395136	A	AACTT	0.09	1.16	1.10	1.24	$2.48x10^{-7}$	0.13