

## **Association of breast cancer risk with genetic variants showing differential allelic expression: Identification of a novel breast cancer susceptibility locus at 4q21**

### **Supplementary Material**

**S1 Table: List of selected genes and genetic variants associated with differential allelic expression.**

For Table S1, please see the attached Excel file

**S2 Table: Associations for the 313 genotyped SNPs with overall, ER-positive and ER-negative breast cancer risk.**

For Table S2, please see the attached Excel file

**S3 Table: Associations for imputed and genotyped SNPs in the 4q21 locus (4q21: 84,132,874-84,631,193) for overall, ER-positive and ER-negative breast cancer risk.**

For Table S3, please see the attached Excel file

**S4 Table: Regulome DB analysis of SNPs in the 4q21 locus (4q21: 84,132,874-84,631,193) with  $r^2 > 0.8$  with top associated SNP rs11099601.** The scoring scheme refers to the following available datatypes: **1a** = eQTL + transcription factor (TF) binding + matched TF motif + matched DNase Footprint + DNase peak; **1b** = eQTL + TF binding + any motif + DNase Footprint + DNase peak; **1c** = eQTL + TF binding + matched TF motif + DNase peak; **1d** = eQTL + TF binding + any motif + DNase peak; **1e** = eQTL + TF binding + matched TF motif; **1f** = eQTL + TF binding / DNase peak; **2a** = TF binding + matched TF motif + matched DNase Footprint + DNase peak; **2b** = TF binding + any motif + DNase Footprint + DNase peak; **2c** =

TF binding + matched TF motif + DNase peak; **3a** = TF binding + any motif + DNase peak; **3b** = TF binding + matched TF motif; **4** = TF binding + DNase peak; **5** = TF binding or DNase peak; **6** = other.

For Table S4, please see the attached Excel file

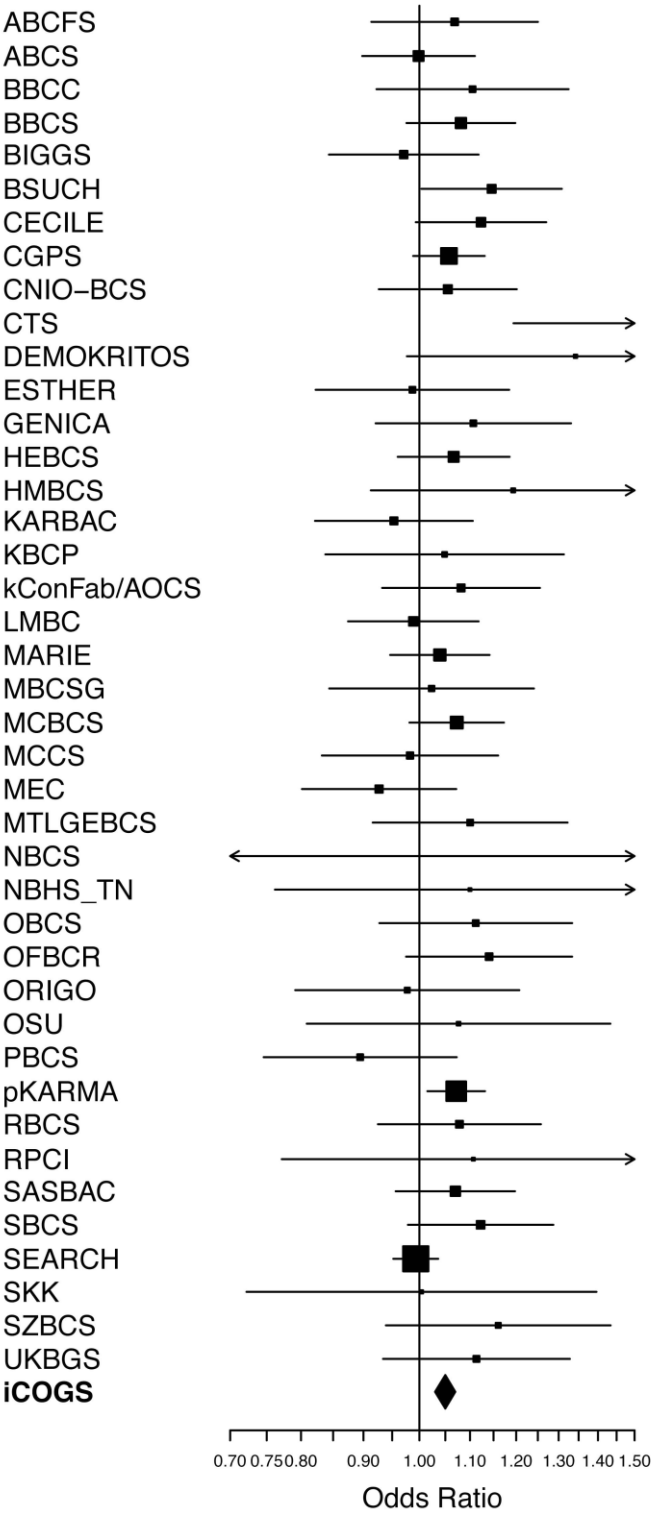
**S5 Table: Description of the BCAC studies with subjects of European origin contributing to iCOGs.**

For Table S5, please see the attached Excel file

**S6 Table: Data sources for *in silico* analyses of the 4q21 breast cancer susceptibility loci.**

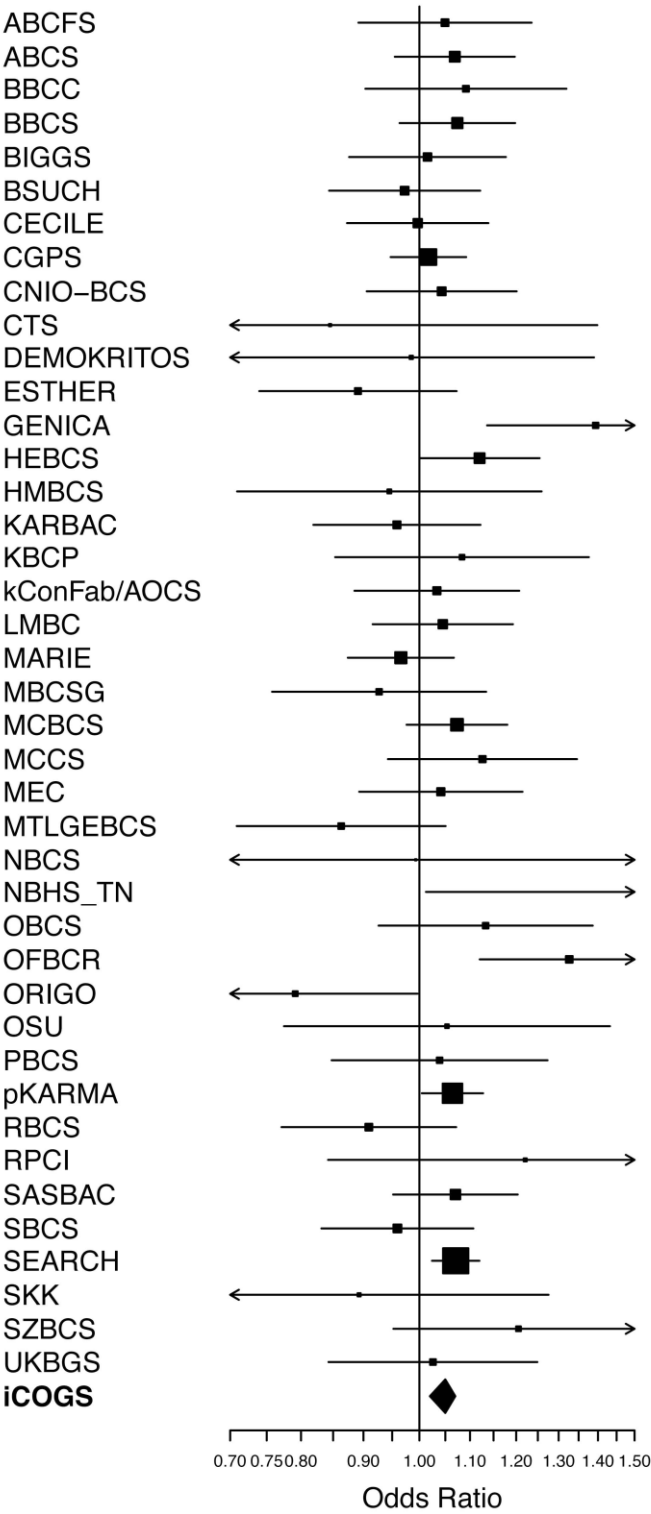
For Table S6, please see the attached Excel file

rs11099601-iCOGS



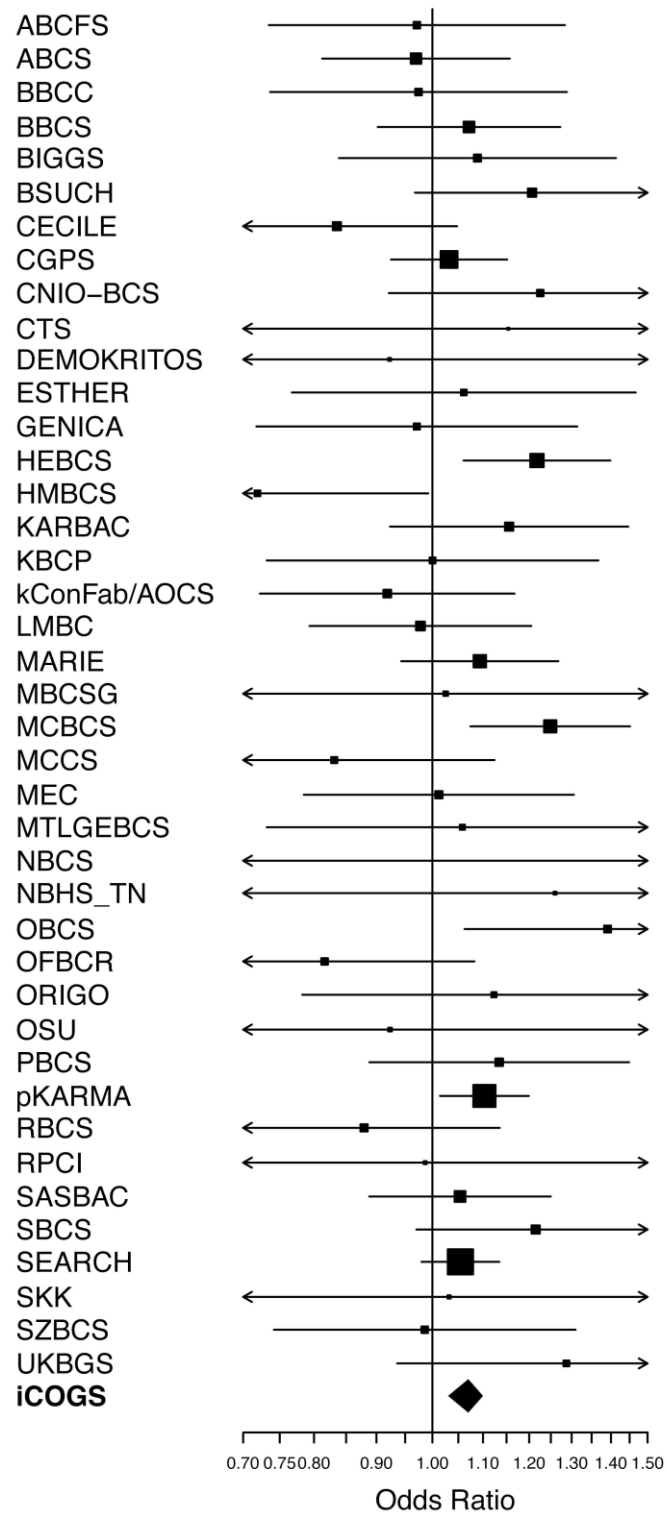
p-het=0.56 I<sup>2</sup>=0

rs656040-iCOGS

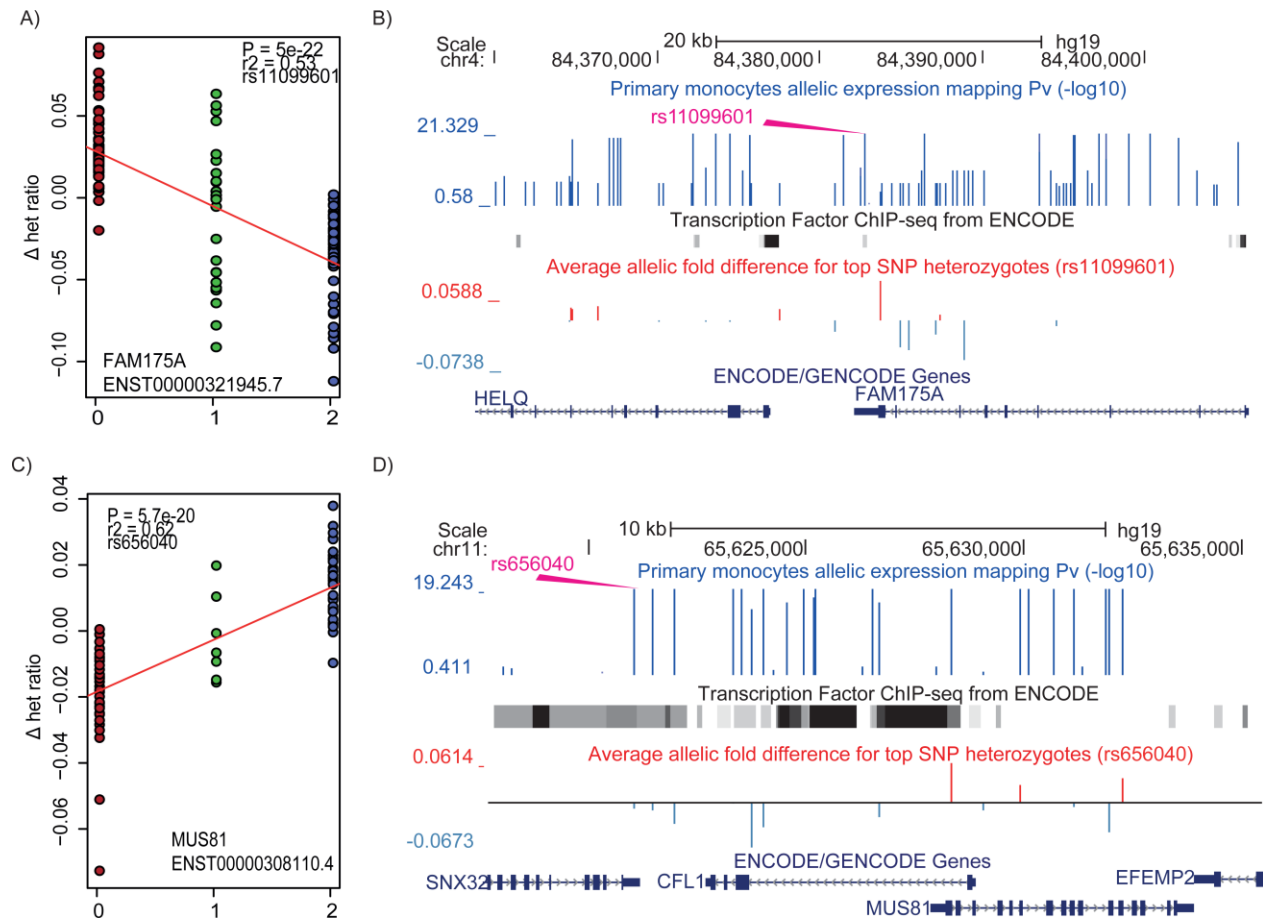


p-het=0.11 I<sup>2</sup>=22.19

# rs738200-iCOGS



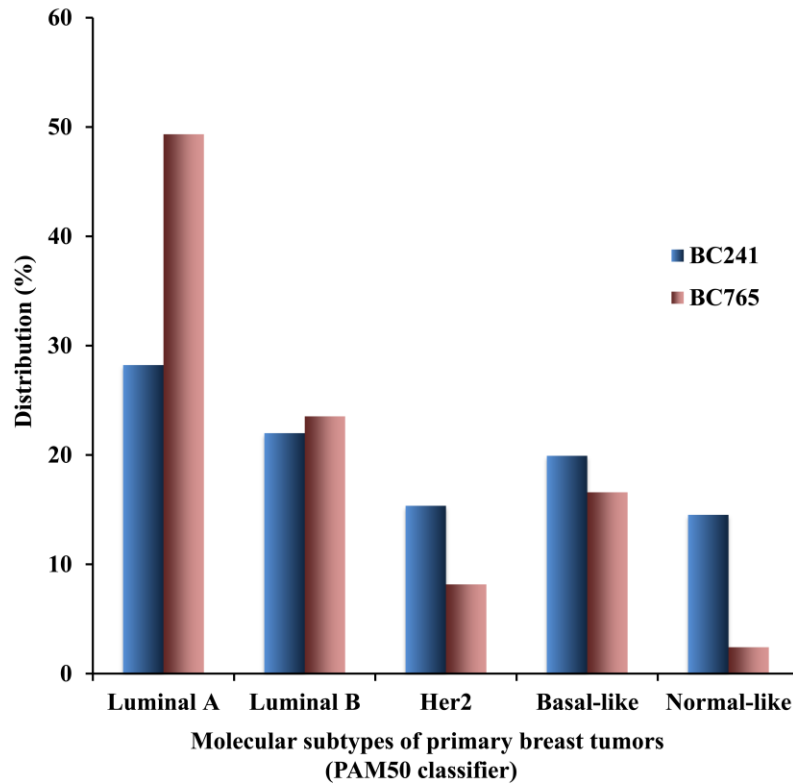
**S1 Figure: Forest plots for the three most significant SNPs (overall  $P$ -value  $<10^{-4}$ ).** Squares indicate the estimated per-allele OR for the minor allele in Europeans. The horizontal lines indicate 95% confidence limits. The area of the square is inversely proportional to the variance of the estimate. The diamond indicates the estimated per-allele OR from the combined analysis.



**S2 Figure: Differential allelic expression mapping of *FAM175A* and *MUS81* loci.** (A) (C)

The most significant *cis*-regulatory variants mapped by regression analysis in the primary monocyte population for *FAM175A* and *MUS81* are rs11099601 ( $P=5 \times 10^{-22}$ ) (A) and rs656040 ( $P=5.7 \times 10^{-20}$ ) (C). Screenshot of the rs11099601 (B) and rs656040 (D) regions from the UCSC genome browser. Tracks display from top to bottom the  $P$ -values ( $-\log_{10}$ ) of the allelic expression mapping in primary monocytes for each SNP, transcription factor binding

(ENCODE ChIP-seq data) and average allelic expression across all individuals heterozygous for rs11099601 (B) and rs656040 (D).



**S3 Figure: Distribution of the 5 molecular subtypes (PAM50 classifier) of breast primary tumors in the two breast cancer samples sets used for eQTL analysis – BC241 and BC765.**

The distribution of Luminal A, Luminal B, Human epidermal growth factor receptor 2-enriched (Her2), Basal-like and Normal-like subtypes is represented as the percentage of the total number of samples in each sample set.