# Supplementary Materials for

## Over-expression of *EPS15* is a favorable prognostic factor in breast cancer

Xiaofeng Dai[1,2,#], Zhaoqi Liu[3,#], Shihua Zhang[3,*]

[1]School of Biotechnology, National Engineering Laboratory for Cereal Fermentation Technology, Jiang-Nan University, Wuxi 214122, China
[2]Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
[3]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Contact: zsh@amss.ac.cn

### Gene expression data

**GSE24450 dataset** The GSE24450 data set consists of 183 primary breast tumor samples, among which 151 were collected as a part of the unselected series at the department of Oncology of the Helsinki University Central Hospital (HUCH) in 1997, 1998 and 2000 [1, 2] and at the department of Surgery from 2001 to 2004 [3]. The remaining 32 patients belong to an ongoing collection of additional familial breast cancer series from the department of Clinical Genetics at HUCH. Total RNA was extracted from the 183 primary breast tumors, and the samples were processed and hybridized to Illumina HumanHT-12_V3 Expression BeadChips, containing 24660 Entrez Gene entities, according to the manufacturer recommendations (http://www.illumina.com). Gene expression profiling was carried out at SCIBLU Genomics Centre, Lund University, Sweden.

Microarray raw data were imported into R [4] and processed by the methods included in the BioConductor facilities [5, 6]. Briefly, after quality control [7], the data were normalized using the quantile method [8] and the gene expression matrix was obtained by averaging the probes mapped to the same Entrez Gene IDs [9].

**GSE4922 dataset** The GSE4922 (GPL97) data set was retrieved from GEO [10], which is comprised of 249 samples including 89 events with relapse or breast cancer specific death [11]. Tissue samples were collected in Uppsala County, Sweden, from January 1, 1987, to December 31, 1989 [11]. RNA was extracted using the RNeasy mini protocol (Qiagen, Hilden, Germany), and the tumor samples were profiled on the Affymetrix U133A genechips at the Genome Institute of Singapore [11]. The data were normalized using the global mean method, natural-log-transformed and scaled by adjusting the mean signal to a target value of log 500 [11]. The maximum follow-up time is 153 months. The information provided on the disease free survival (DFS), was analyzed within the course of this study.

**GSE25307 dataset** The GSE25307 data set was retrieved from GEO [10], which is comprised of 577 samples including 228 events with overall death [12]. Tissue

samples were collected from the Southern Sweden Breast Cancer Group tissue bank at the Department of Oncology, Skåne University Hospital (Lund, Sweden), the Helsinki University Central Hospital (Helsinki, Finland) and Landspitali University Hospital (Reykjavik, Iceland) [12]. The RNA was extracted from these tumor samples, which were profiled using oligonucleotide microarrays (GEO platform GPL5345) at the SCIBLU Genomics Centre at Lund University (Lund, Sweden) [12]. The data were normalized using block-based Lowess [13]. The maximum follow-up time is 382 months provided with the information on overall survival (OS), which was analyzed during the course of this study.

**TCGA dataset** The level 3 primary solid breast tumor mRNA expression data was retrieved from TCGA (**http://cancergenome.nih.gov**) on 21st November 2011. The data includes 514 samples, among which 512 have recorded information on OS including 53 death events. The mRNA data has been produced using Agilent 244K Custom Gene Expression G4502A-07-3 platform, lowess normalized followed by log2-transformation of the ratio between two channels. The maximum follow-up time is 226.5 months, which is truncated at 10 years here since death after 10 years is less likely to be caused by breast cancer. OS was recorded in this dataset, which was analyzed in this study.

**METABRIC dataset** The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset contains detailed clinical annotations, patient overall survival time, expression profiles, CNV profiles, and SNP genotypes derived from 1981 breast tumors collected from participants of the METABRIC trial [14]. Nearly all oestrogen receptor (ER)-positive and lymph node (LN)-negative patients did not receive chemotherapy, whereas ER-negative and LN-positive patients did. None of the HER2+ patients in this trial received trastuzumab. This dataset was accessed through Synapse (synapse.sagebase.org). The expression profiles contain 49576 probe sets, performed on the Illumina HT 12v3 platform, re-normalized at Sage Bionetworks by the BCC Support Team. We used the 15-year overall survival time in this study. More detailed description on METABRIC data is available at the Breast Cancer Challenge support page (https://sagebionetworks.jira.com/wiki/display/BCC).

**GLEDBLS dataset**

This large-scale breast cancer dataset was manually curated by Győrffy Lab at Hungarian Academy of Sciences and Semmelweis University Budapest [15]. This dataset was downloaded (July 23, 2014) from their online webserver: www.kmplot.com which was developed to assess the relevance of the expression levels of various genes on the clinical outcome both in untreated and treated breast cancer patients. A background database was established using gene expression data and survival information of breast cancer patients downloaded from GEO (Affymetrix microarrays only) and EGA and these multiple datasets were combined to increase the statistical power when performing survival analysis. The relapse free survival was adopted in this study.

## Tables

**Table S1 - Data sets description for computational prediction at the genetic and transcriptional levels.**

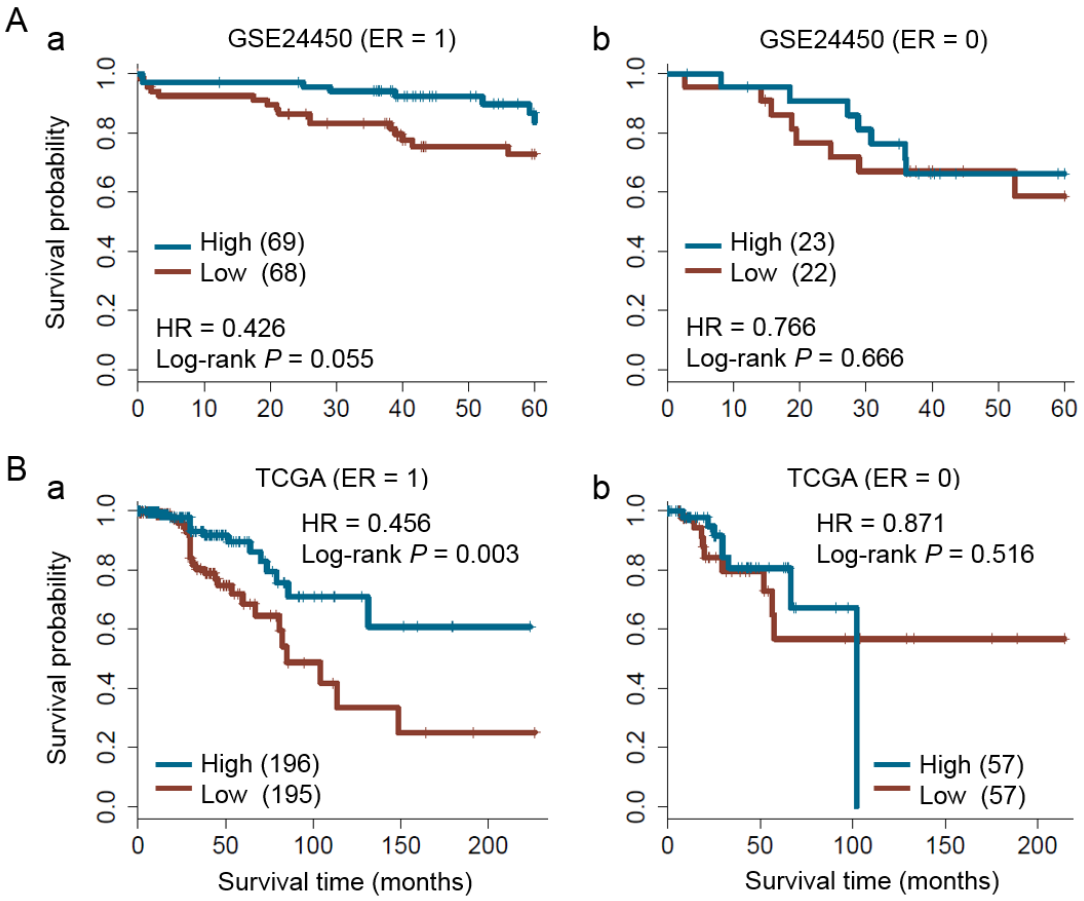
| Analysis | GEX survival | | | | | | eQTL analysis | | |
|---|---|---|---|---|---|---|---|---|---|
| **Data** | GSE24450 | GSE4922 | GSE25307 | TCGA | METABRIC | GLEDBLS | TCGA | | |
| **Data type** | GEX | GEX | GEX | GEX | GEX | GEX | SNP | GEX | CNV |
| **Number** | 183 (39) | 249 (89) | 551 (228) | 502 (65) | 1981 (888) | 3455 (1160) | 502 | 502 | 502 |

**Table S2 - 21 SNPs significantly associated with *EPS15* expression (FDR < 0.1).**

| SNP | Chromosome/ Position | *P* value | FDR | Gene | Regulation type |
|---|---|---|---|---|---|
| **rs10094308** | 8:10406197 | 4.46E-14 | 3.00E-07 | MSRA | *trans*-eQTLs |
| **rs4615335** | 5:110862024 | 3.48E-08 | 6.32E-07 | | *trans*-eQTLs |
| **rs41488049** | 8:3593141 | 4.50E-11 | 6.97E-05 | CSMD1 | *trans*-eQTLs |
| **rs7843727** | 8:16028400 | 1.35E-10 | 0.00013053 | | *trans*-eQTLs |
| **rs9923546** | 16:73307210 | 2.15E-09 | 0.001616799 | | *trans*-eQTLs |
| **rs13393078** | 2:145636412 | 0.000565 | 0.004860672 | | *trans*-eQTLs |
| **rs9969649** | 8:10476769 | 1.70E-08 | 0.006894577 | LOC101929191 | *trans*-eQTLs |
| **rs17125079** | 8:17674978 | 2.01E-08 | 0.00864128 | MTUS1 | *trans*-eQTLs |
| **rs17053428** | 8:25348147 | 4.07E-08 | 0.013713771 | DOCK5 | *trans*-eQTLs |
| **rs6882329** | 5:73934031 | 0.000991 | 0.039546476 | ARHGEF28 | *trans*-eQTLs |
| **rs1565114** | 4:42321725 | 3.04E-07 | 0.044312076 | | *trans*-eQTLs |
| **rs2632689** | 4:42324198 | 4.13E-07 | 0.053241146 | | *trans*-eQTLs |
| **rs9497** | 16:56943662 | 1.62E-07 | 0.05326167 | HERPUD1 | *trans*-eQTLs |
| **rs6041945** | 20:13184534 | 4.42E-07 | 0.054098293 | | *trans*-eQTLs |
| **rs2305433** | 17:48188071 | 4.32E-07 | 0.058079008 | SKAP1 | *trans*-eQTLs |
| **rs12678557** | 8:5291862 | 5.55E-07 | 0.063111147 | | *trans*-eQTLs |
| **rs7008036** | 8:13733245 | 6.04E-07 | 0.068823482 | | *trans*-eQTLs |
| **rs17667604** | 5:169161040 | 6.34E-07 | 0.074513988 | SLIT3 | *trans*-eQTLs |
| **rs12302658** | 12:74728536 | 6.95E-07 | 0.075879943 | | *trans*-eQTLs |
| **rs1161456** | 13:28862278 | 0.000246 | 0.082412878 | | *trans*-eQTLs |
| **rs6102600** | 20:41958395 | 1.10E-06 | 0.0987093 | | *trans*-eQTLs |

## Figures

**Figure S1 - Kaplan-Meier cumulative survival curves showing the association between *EPS15* expression and breast cancer patient survival in ER positive (a. ER = 1) and negative (b. ER = 0) tumors using A) GSE24450 (HEBCS) data, and B) GLEDBLS data, respectively. For each dataset, breast cancer patients are divided into two equal-sized groups using the median of *EPS15* expression. Two groups are denoted as high expression (High) and low expression (Low) respectively and the number in the brackets indicate the size of the group. Log-rank *P* value and hazard ratio are showed in each subplot.**




## References for Supplementary materials


1. Syrjakoski K, Vahteristo P, Eerola H, Tamminen A, Kivinummi K, Sarantaus L, Holli K, Blomqvist C, Kallioniemi OP, Kainu T *et al*: **Population-based study of BRCA1 and BRCA2 mutations in 1035 unselected Finnish breast cancer patients**. *Journal of the National Cancer Institute* 2000, **92**(18):1529-1531.

2. Kilpivaara O, Bartkova J, Eerola H, Syrjakoski K, Vahteristo P, Lukas J, Blomqvist C, Holli K, Heikkila P, Sauter G *et al*: **Correlation of CHEK2 protein expression and c.1100delC mutation status with tumor characteristics among unselected breast cancer patients**. *Int J Cancer* 2005, **113**(4):575-580.

3. Fagerholm R, Hofstetter B, Tommiska J, Aaltonen K, Vrtel R, Syrjakoski K, Kallioniemi A, Kilpivaara O, Mannermaa A, Kosma VM *et al*: **NAD(P)H:quinone oxidoreductase 1 NQO1*2 genotype (P187S) is a strong prognostic and predictive factor in breast cancer**. *Nat Genet* 2008, **40**(7):844-853.

4. Team RDC: **R: A language and environment for statistical computing**: R Foundation for Statistical Computing; 2009.

5. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: Open software**

**development for computational biology and bioinformatics R**. *Genome biology* 2004, **5**(10):R80.

6. Smyth GK: **Limma: linear models for microarray data**. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* New York: Springer; 2005: 397-420.

7. Du P, Kibbe WA, Lin SM: **Lumi, a pipeline for processing illumina microarray**. *Bioinformatics* 2008, **24**(13):1547-1548.

8. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

9. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2005, **33**(Database issue):D54-D58.

10. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**(1):207-210.

11. Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, Lindahl T, Pawitan Y, Hall P, Nordgren H *et al*: **Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer**. *Cancer research* 2006, **66**(21):10292-10301.

12. Jonsson G, Staaf J, Vallon-Christersson J, Ringner M, Gruvberger-Saal SK, Saal LH, Holm K, Hegardt C, Arason A, Fagerholm R *et al*: **The retinoblastoma gene undergoes rearrangements in BRCA1-deficient basal-like breast cancer**. *Cancer research* 2012, **72**(16):4028-4036.

13. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation**. *Nucleic acids research* 2002, **30**(4):e15.

14. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y *et al*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012, **486**(7403): 346-352.

15. Győrffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, Szallasi Z: **An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients.** *Breast cancer research and treatment* 2010, **123**(3): 725-731.