Supplementary Information

Regulators of genetic risk of breast cancer identified by integrative network analysis.

Mauro AA Castro[1], Ines de Santiago[2,3], Thomas M Campbell[2,3], Courtney Vaughn[2,3], Theresa E Hickey[4], Edith Ross[2], Wayne D Tilley[4], Florian Markowetz[2], Bruce AJ Ponder[2,3], Kerstin B Meyer[2,3,*]

[1]Bioinformatics and Systems Biology Lab, Federal University of Paraná (UFPR), Polytechnic Center, Rua Alcides Vieira Arcoverde, 1225 Curitiba - PR 81520-260 - Brazil.
[2]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK.
[3]Department of Oncology, University of Cambridge, Hutchison/MRC Research Centre, Box 197, Hills Rd, Cambridge CB2 0XZ, UK.
[4]Dame Roma Mitchell Cancer Research Laboratories, School of Medicine, The University of Adelaide, Adelaide, SA 5000.

*Email: kerstin.meyer@cruk.cam.ac.uk.

Supplementary Note

Supplementary Figures

Supplementary Tables

Supplementary Note

Effects of copy number on network structure

In order to assess the impact of copy number variation on the structure of the network, we compared the correlation network obtained from the unadjusted gene expression values of cohort I to the correlation network obtained from gene expression values that had been adjusted for copy number variation using linear regression. Supplementary Figure 1 shows that there is very high correlation between the two networks, suggesting that in this data set CNV does not have a strong effect on the network structure.

Genes contributing to significant mapping tallies in the AVS to regulon association

The EVSE analysis tests the association between an AVS and a regulon, with p-values derived from the comparison of a mapping tally generated with a set of GWAS SNPs compared to mapping tallies generated with multiple random SNP lists. In this analysis some risk loci (listed by their tagging SNP) contribute to the mapping tallies of a large number of different regulons (see **Figure 1**). We wondered whether in these cases the same or different genes drive the observed associations. For the 36 significantly associated risk regulons, we therefore extracted the SNPs in the AVS that acted as eQTLs and the target genes whose expression correlated with the allelic status at that SNP. These SNP-gene pairs are listed in Supplementary Figure 4a and 4b for cohort I and II of the METABRIC data set. The figure displays the data in a gene centric way. Each gene is listed in a separate row even if a single locus is able to drive the expression of multiple target genes. Where several risk loci act as eQTL for the same target gene, the gene is only listed once. Only genes that are part of significant regulons are shown, and the figure does not represent a complete eQTL map. The figures illustrate that there is a very complex relationship between the AVS and the regulons.

Some risk loci act as eQTL for a single target gene that contributes to the enrichment of only a few regulons (e.g cohort I: rs11075995|FTO - 2 regulons: E2F2 and E2F3), while for other loci the linked target gene can contribute to the risk-association of multiple regulons (cohort I: rs616488|CASZ1 - 23 regulons). We also observe instances where a single locus acts as eQTL for multiple target genes that contribute to many different regulons (eg rs3903072, with 8 and 11 target genes in cohort I and cohort II, respectively). Our analysis demonstrates that the significant associations between the AVS and the regulons are driven by overlapping but distinct gene sets for each regulon. No two regulons have an identical set of contributing genes (Supplementary Figure 4a and 4b).

We noted that in the two cohorts different genes can contribute to the significant association of the AVS with a regulon. For example the SOX10 regulon is linked to 21 genes in cohort II, and to 15 genes in cohort I, but only 8 of the genes contribute in both cohorts. This illustrates that a regulon-based analysis can integrate information from disparate data sets: although different

targets contribute, the regulators that are identified can still be the same.

We also examined whether there was a positive or negative correlation between the expression of a risk-TF and the expression of its target gene (highlighted by black or grey squares respectively, Supplementary Figure 4a and 4b). We observed that different risk-TFs are able to regulate overlapping sets of genes, but the direction of regulation can vary, with risk-TFs apparently falling into two groups. This relationship is explored further in the analysis depicted in **Figure 4**.

Identification and analysis of METABRIC samples with homogeneous genetic background

In order to avoid spurious associations due to ancestry differences between populations, only samples with a homogeneous genetic background were selected. Population structure was detected using principal component analysis (PCA) of the genotypes. First, the principal components of the genotypes of the 11 HapMap populations were computed using the program EIGENSTRAT from the EIGENSOFT package (version 4.2)[57,58]. Then, the METABRIC samples were projected onto these components.

The genotypes of the HapMap populations were downloaded in PED format from the Broad Institute (http://www.broadinstitute.org/debakker/hapmap3r2.html). To convert them into the format required by EIGENSTRAT, the 11 genotype files were recoded from the PED format, which contains both alleles for each SNP (e.g. AA, AB or BB) to allele counts (e.g. 0, 1 or 2 if B is the allele whose frequency is counted). This was done using PLINK[59]. For each SNP, the allele chosen to be counted matched the allele used for the coding of the METABRIC genotypes. However, care had to be taken to choose the allele aligning to the correct strand: In the HapMap data set all alleles are aligned to the forward strand, whereas in the METABRIC data set genotypes of a fraction of the SNPs are reported with respect to the reverse strand.

Then, the genotype files of the HapMap populations and the METABRIC samples were joined, omitting SNPs that were not genotyped in all groups. To reduce the computation time of EIGENSTRAT 250,000 SNPs were randomly selected for the PCA. Most METABRIC samples have a very homogeneous genetic background (Supplementary Figure 5a): while few METABRIC samples co-cluster with HapMap samples of Asian or African ancestry, the large majority co-clusters with samples of European ancestry (CEU and TSI). A threshold to remove samples from the analysis was chosen visually at -0.035 of the second principal component. 1829 METABRIC samples were selected for further analysis. EVSE analyses of the 36 risk-TFs in these genetically homogeneous cohorts generated results that are very similar to those obtained with the entire METABRIC cohorts (Supplementary Figure 5).

Comparison of EVSE to other methods

The EVSE method successfully and reproducibly identifies risk-TFs. To understand how generalizable the approach is we compared the different analysis steps in the EVSE pipeline to alternative methods.

Analysis steps	Method in EVSE	Compared to	Results
Network reconstruction methods	MI (Gaussian Estimator)	Spearman rho and MI (Empirical Estimator)	Supp Fig 6
	ARACNe/DPI	MRNETB, MRNET, CLR and PCOR	Supp Fig 7
Expansion of tagging SNP into AVS	D'>.99, LOD>3	r2	Supp Fig 8
eQTL filtering	MANOVA eQTL	250kb windows only	Supp Fig 9
	MANOVA eQTL	pre-defined eQTLs	Supp Fig 10 Supp Fig 11 Supp Fig 12
	MANOVA eQTL (ER+)	predefined eQTLs (ER+)	Supp Fig 18a

Network reconstruction methods

Correlation in gene expression was used to estimate the TF-target associations. In this study we used a Gaussian estimator to compute Mutual Information (MI) between TFs and potential targets using the *Minet* R package[60]. The results were compared with those obtained with Spearman's correlation and mutual information using an empirical estimator (Supplementary Figure 6). The distributions differ to some extent, but they are highly correlated (P<2.2e-16, F-statistic, both comparisons). The overall associations derived by either mutual information or by correlation are similar.

We constructed the regulatory networks based on mutual information using permutation and bootstrap analyses to infer the TF-target interactions, and the ARACNe algorithm was used to evaluate the resulting adjacency matrix. In order to compare ARACNe with different choices of network construction methods we applied four other algorithms to evaluate the same adjacency matrix: Maximum Relevance Minimum Redundancy (MRNET), Maximum Relevance Minimum Redundancy Backward (MRNETB) (both of which are based on feature selection), Context Likelihood (CLR) and Partial Correlation (PCOR). For detailed information about these algorithms please refer to Meyer *et al.*[60] and Altay *et al.*[61]. These algorithms aim to identify the most relevant interactions and generate regulons of slightly smaller size. To generate similar, comparable regulons in ARACNe a DPI (data processing inequality) threshold of 0.1 was therefore applied. We used the *Minet* R package[60] to execute the MRNET, MRNETB and CLR algorithms, while PCOR was computed with the R package *Corpcor*[62]. The ARACNe algorithm is used as the underlying reference to compare the networks by the Receiver Operating Characteristic (ROC) approach (Supplementary Figure 7a). Also, precision-recall (PR) curves have been used as an alternative to assess the performance of the algorithms (Supplementary Figure 7b). These comparisons demonstrate

some differences in the computed networks.

(a) Validation of the 36 risk-TFs
The regulons for the 36 risk-TFs inferred by the five network construction algorithms were then tested for risk association by the EVSE analysis (Supplementary Figure 7c). Three of the methods (ARACNe, MRNET and MRNETB) identify >90% of the risk-TFs. Partial correlation showed the lowest consistency, but still generated 75% overlap with the reference network. Our data show that whilst the derived networks are not the same the identified risk-TFs can be validated when alternative network construction algorithms are used. It is important to note that all these algorithms remove, at some extent, part of the conditional dependencies among the regulons – including ARACNe (i.e. DPI is set to 0.1), and despite that the results are consistent with those presented in **Figure 1**.

The robustness of the regulatory network construction can also be tested by experimental validation. We therefore examined the density of ESR1 binding sites at all genes in the ESR1 regulon using ChIP-seq data[4]. Supplementary Figure 7d shows the density distribution of estrogen binding sites mapped to genes in the ESR1 regulons inferred by ARACNe, MRNET, MRNETB, CLR and PCOR algorithms. Overall the sensitivity of the different algorithms to identify genes that are bound by ESR1 is similar.

(b) Hypothesis free derivation of risk-TFs using alternative regulon construction methods
The above analysis largely validates the identification of the 36 risk-TFs using alternative methods to define the regulons. However, we also wondered whether additional risk-TFs would be identified if these methods were applied to all regulons in the network (Supplementary Figure 7e-f). This work is discussed at the end of the Supplementary Information (page 8) as it needs to be interpreted in the context of results depicted in **Figure 5** of the main manuscript.

Expansion of tagging SNP into AVS

EVSE uses a D'>0.99 and LOD>3.0 to expand the risk SNP into an AVS. While VSE enrichment was first described using these parameters, genetic linkage relevant to GWAS is frequently measured by r^2. We therefore carried out the EVSE analysis for the 36 risk-TF using AVSs calculated using different r^2 cut-off values. The r^2 statistic was computed between SNPs up to 250 kb apart, obtained from the 1000 Genomes project[63], release 20130502, GRCh37/hg19. LD was computed for the CEU/CEPH population using the software PLINK[59]. Average enrichment scores remained significant (Supplementary Figure 8a). Supplementary Figure 8b and c list the results of the EVSE analysis for each regulon using r^2>0.8 in cohort I and II respectively. Most of risk-regulons continue to generate significant enrichment scores. The sensitivity of the technique falls when the r^2 is increased above 0.8, presumably because the distribution of SNPs in the AVS becomes too sparse to generate sufficient hits. Using r^2>0.8 appears to be somewhat more stringent than the analysis with D'>0.99 and LOD>3. However, it is not possible to compare

our analysis methods to a "gold-standard" that allows us to determine which risk scores are the most relevant. We therefore used the method as it was first described[7]. At many risk loci additional risk variants have been identified in genetic fine mapping studies and the D' measure may have greater sensitivity to detect these variants (reviewed in Fachal and Dunning[22]), which may show relatively low LD, but are often physically close to the strongest risk variant, still mapping to the same haplotype block.

eQTL filtering

The EVSE analysis uses MANOVA to identify genes associated with the GWAS loci, based on a significant association of gene expression of any target gene within a +/-250kb window around the AVS with the SNP status of any of the SNPs in the AVS. We compared this to two different approaches. (1) We selected all genes in a given window size of +/-250 kb). The results showed some overlap with our EVSE analysis (Supplementary Figure 9a and 9b). However, the results were less reproducible between cohorts and only 29% of TFs identified in cohort I replicated in cohort II (Supplementary Figure 9c), and known risk genes such as ESR1 are not identified in this analysis. (2) We replaced the eQTL step in our EVSE analysis by calling eQTLs in the METABRIC dataset using previously described methods[10]. The eQTLs were called on a genome-wide basis using only ER-positive breast cancer cases, and the analysis was not limited by a window size as in the EVSE. As this requires many comparisons, a correction for multiple testing had to be applied so that only 15 of the 72 risk-associated regions were represented by at least one significant eQTL (FDR < 0.05), accounting for only 18 eQTL genes within the query risk regions. Therefore, this analysis led to a sparser distribution of gene associations with GWAS hits and lower mapping tallies (Supplementary Figure 10a and 10b for cohort I and II respectively). Given this distribution the subsequent comparison to random hits becomes less stable. We identified a consensus of 10 TFs across the two cohorts (Supplementary Figure 10c). All of the 10 identified TFs were also identified by EVSE in at least one cohort, and 8 in both cohorts (Supplementary Figure 11). Therefore there was good overlap with the EVSE analysis, but the total number of identified risk-TFs was much smaller and did not include TFs whose association with GWAS loci was subsequently verified using ChIP-seq data (AR and RAR for example). The analysis with 'pre-defined' eQTLs therefore lacked sensitivity. The same analysis was repeated using an eQTL dataset derived from ER-negative breast cancer cases, however this analysis did not reveal any significant results, mainly because only 2 out of the 72 risk-associated regions were represented in the eQTL dataset (Supplementary Figure 12a and 12b). While some variation is to be expected across different analysis platforms, the results with pre-defined eQTLs are supportive of our risk-TF assignment, but this analysis appears to be far less sensitive than the EVSE analysis.

Overall, we find that different methods to derive the networks, the AVS and the eQTL filtering will identify slightly different sets of risk-TFs. Similarly, the use of different data sets (compare cohort I and cohort II, Supplementary Figure 3a and 3b) will result in the identification of different but overlapping sets of risk-TFs. Partially this may to be due to differences in the

underlying data, but may also reflect that our analysis will generate some false positive as well as false negative risk associations. Our comparisons to alternative methods give some insight into which of the risk-associations are the most stable.

Confirmation of risk association based on master regulator analysis (MRA)

The FGFR2 and the estrogen responses were assayed in three ER^+ breast cancer cell lines: MCF-7, T47D and ZR751. In each case cells were synchronised by estrogen starvation and then treated with estrogen (E2) alone or E2 plus FGF10. Microarray gene expression experiments were carried out in triplicate 24 hours after E2 or E2 plus FGF10 treatment. DE gene lists were called using limma[48] (E2 versus vehicle or E2+FGF10 versus E2) and used in MRA. This analysis tests whether a TF-regulon is enriched for a gene signature, thereby identifying the relevant TF as master regulator (MR) of the response that generated the gene signature. MRs were identified using the previously described RTN package[4,16] with networks calculated independently for cohort I and II in the METABRIC data set. To gain confidence in the analysis we only considered TFs that were significant for both cohorts and reproducible in at least two cell lines (Supplementary Table 2). When examining the response to estrogen stimulation resting cells were compared to rapidly proliferating cells and many MRs are expected simply to coordinate the growth response. Proliferation-associated MRs were identified by carrying out a MRA[15] with the meta-PCNA signature (a cell type independent list of proliferation associated genes)[47] as previously described[4] (shown in red in Supplementary Table 2). After estrogen induction, of the 7 MRs not associated with the experimentally induced proliferation, 6 were also identified as risk-TFs from at least one of the METABRIC cohorts (ASCL2, CITED, ESR1, GATA3, PGR and SPDEF: Supplementary Table 2). After FGFR2 signalling we found that of the 8 non-proliferation associated MRs, 5 were risk-associated in at least one cohort: ESR1, GATA3, SPDEF, XBP1 and CSDA. The list of identified TFs is very similar to the list of MRs of FGFR2 that we previously identified in MCF-7 cells only, which was based on multiple distinct experimental systems to elicit the FGFR2 response in a single cell line.

Identification of risk-TFs using alternative network construction methods

As discussed above an adjacency matrix between TFs and all potential target genes was generated based on mutual information using bootstrap and permutation analysis. Regulons were then calculated using three alternative network construction methods and the EVSE analysis applied to each of these. We chose to compare three methods: MRNETB, CLR and ARACNe. As above, to generate regulons of comparable size we ran ARACNe with a DPI (data processing inequality) threshold of 0.1. This step aims to remove indirect targets from the regulons. The analyses were carried out independently for cohort I and cohort II. Supplementary Figure 7e shows the overlap of the results obtained for each of the three

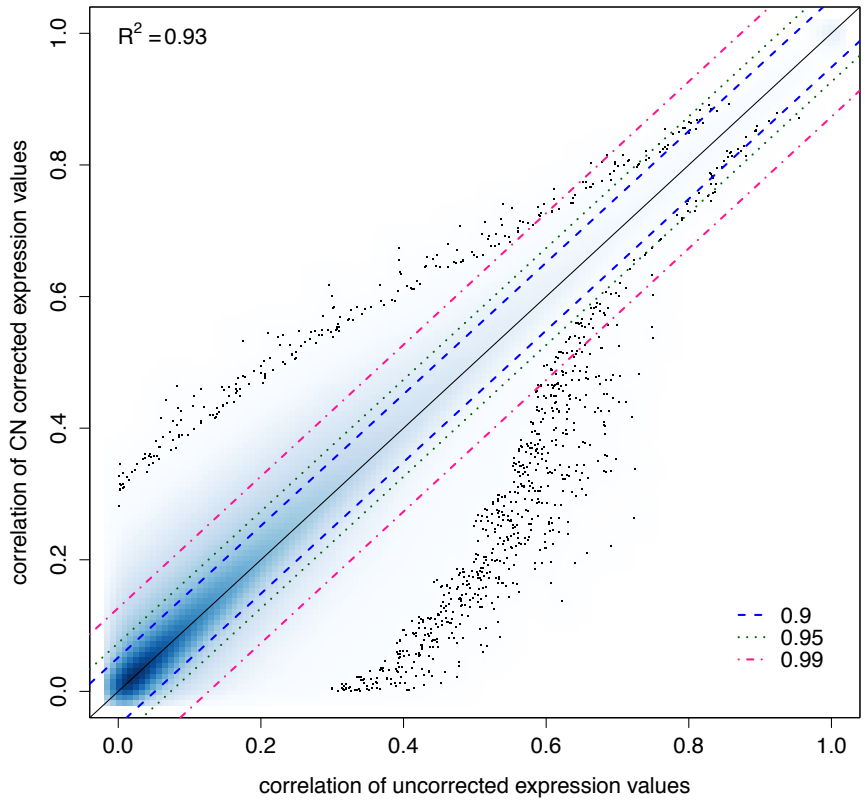
approaches. In each case it was apparent that more regulons were identified for cohort I. In contrast, our main analysis (ARACNe, DPI threshold of 1) identified roughly the same number of risk associated TFs for each of the METABRIC cohorts, giving an overlap of 36 risk-TFs (Supplementary Figure 7f).

We then defined a second set of risk-TFs that were significantly enriched for risk-associated genes in at least 4 of the 6 EVSE analyses (using MRNETB, CLR, ARACNe, each in cohort I and II) (listed in Supplementary Table 4). This threshold allows for some false negative results in the analyses. When this new set of risk-TFs is mapped onto the tree and leaf diagram (see **Figure 5**) cluster 1 and 2 (defined in Figure 5) are again strongly highlighted (Supplementary Figure 18b). Whilst 4 of the original risk-TFs in cluster 2 no longer pass the threshold, 5 new risk-associated TFs (SRF, ELF5, CREB3L2, KLF11 and TCFL1) map to cluster 2. For cluster 1, 6 risk-TFs did not reach the threshold of 4 out of 6 positive associations, but 2 new ones were identified, again generating a strong clustering within the region previously defined as cluster 1. All factors we had previously validated using ChIP-seq data (XBP-1, FOXA1, ESR1, GATA3, RAR and AR) remain associated. Additional new risk-TFs not mapping to the two clusters were found. However, no obvious new clusters were identified (Supplementary Figure 18b).

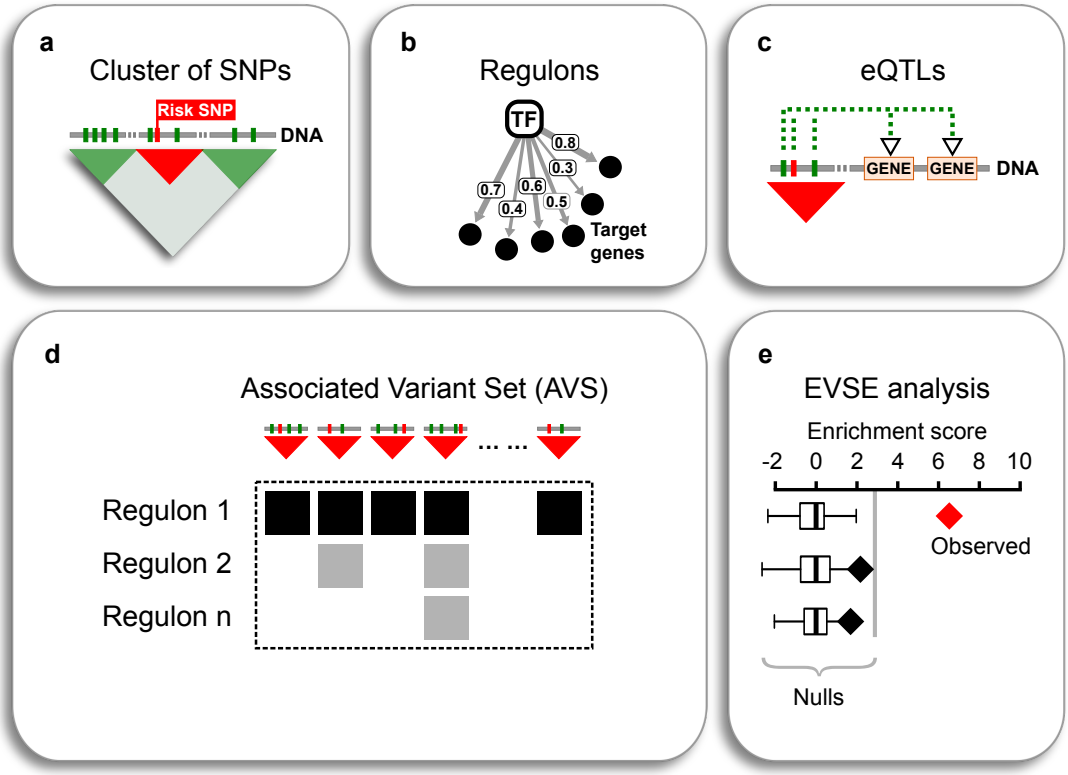
Source code

The source code developed in this study is publicly available from the Bioconductor[55] in the R packages *RTN*[16] and *RedeR*[56]:

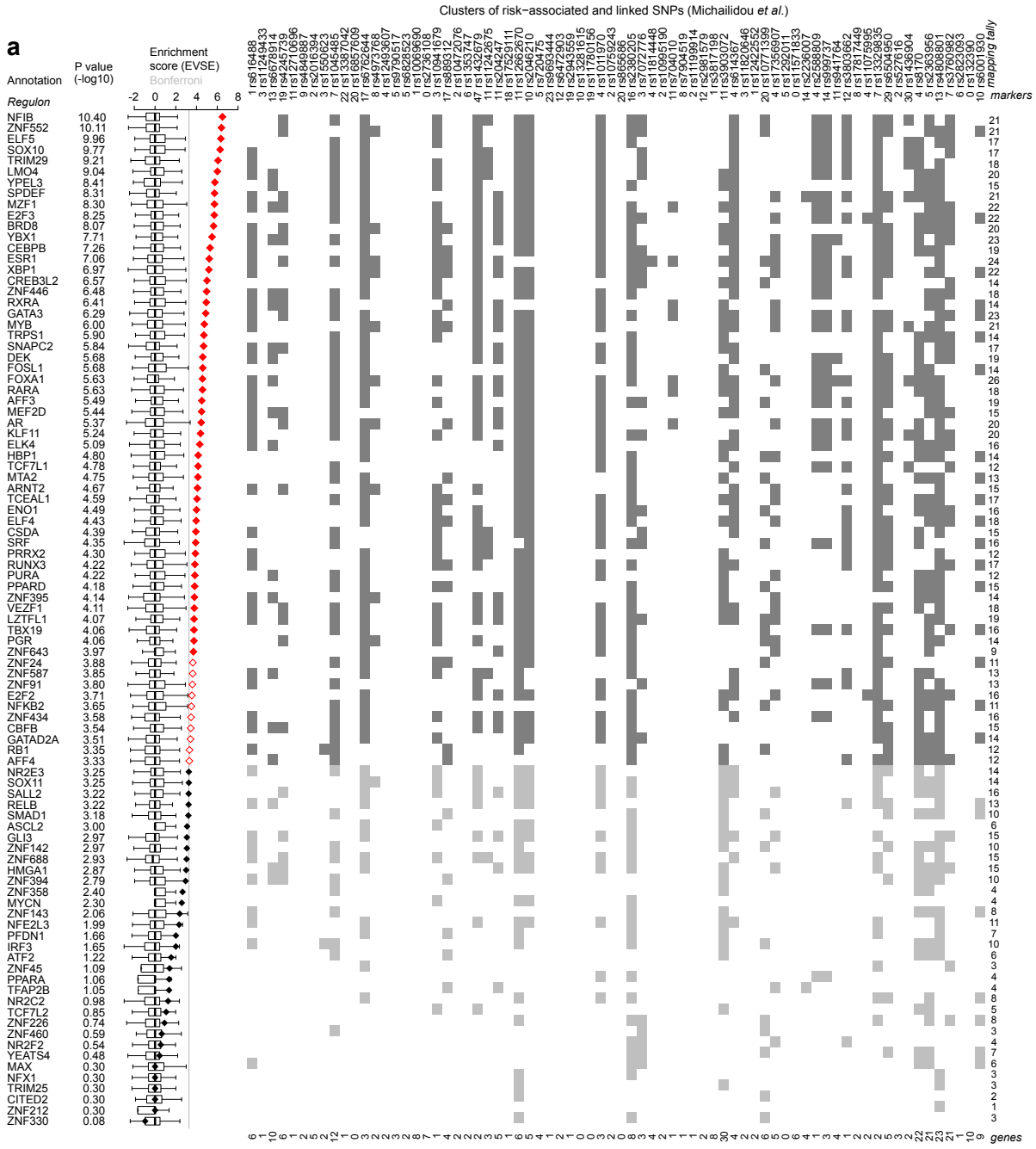
- R package *RTN*
  http://bioconductor.org/packages/RTN/
- R package *RedeR*
  http://bioconductor.org/packages/RedeR/

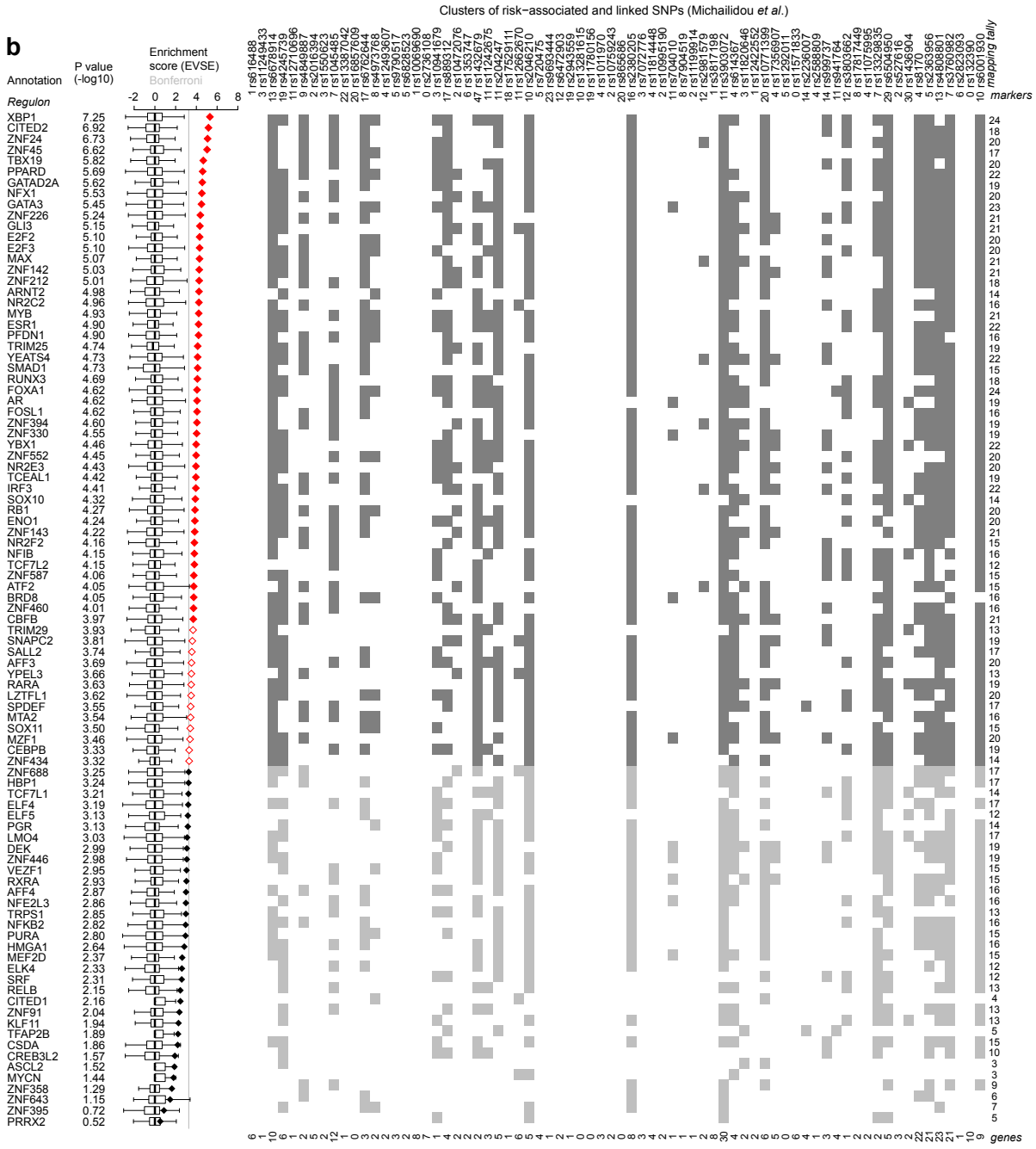


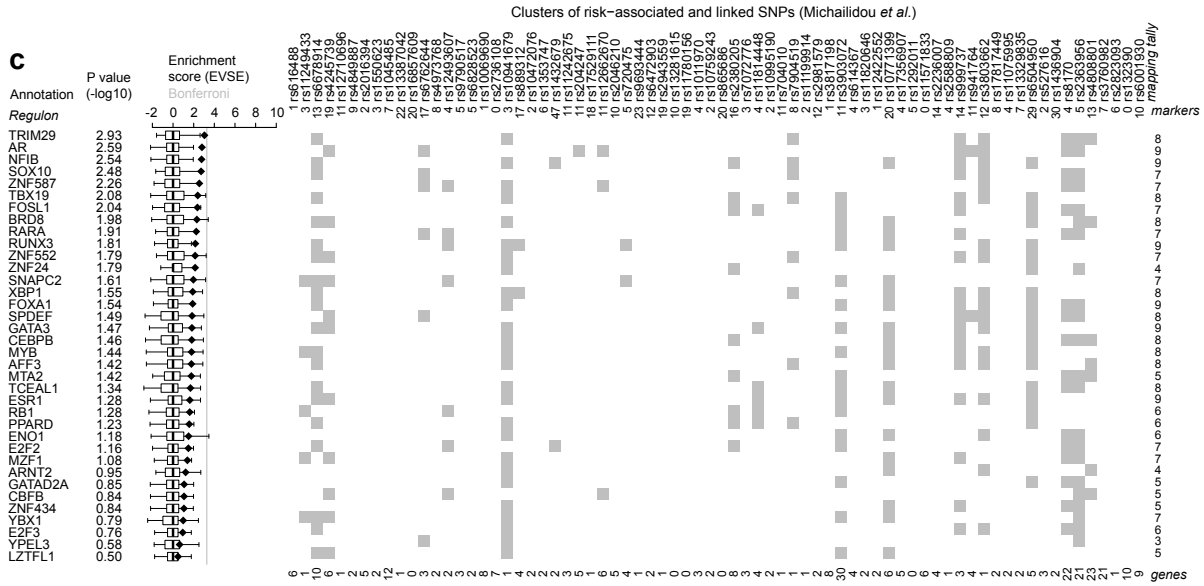
Supplementary Figure 1: **Correlation of edge weights of network inferred from unadjusted gene expression data with edge weights of network inferred from copy number adjusted gene expression data.** This figure shows a smoothed colour density representation of the scatterplot. 1000 points lying in the regions of lowest densities are shown as outliers. The marked regions around the diagonal contain 90, 95 and 99% of the points, respectively.



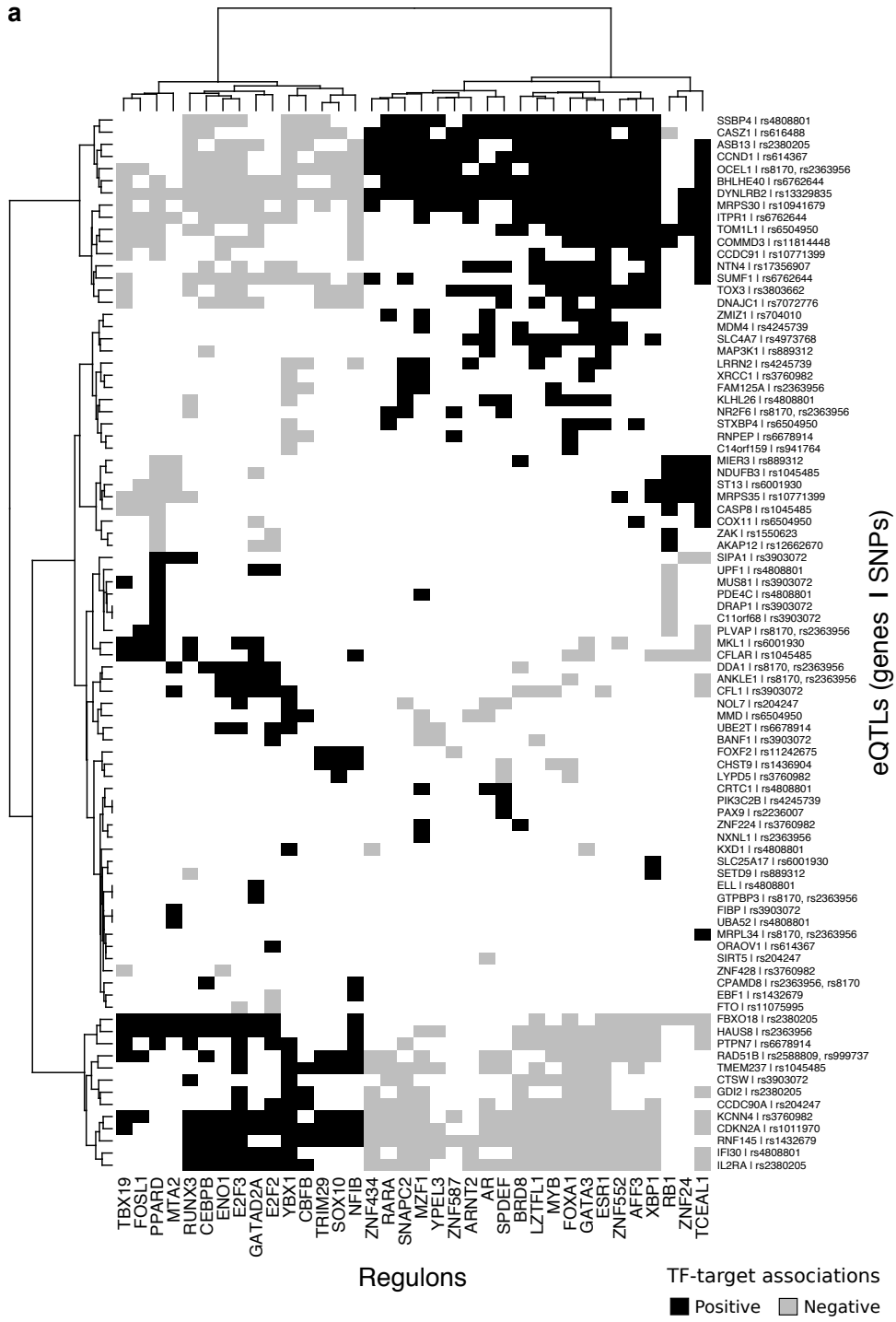
Supplementary Figure 2: **Cartoon of the analysis pipeline highlighting the different datasets considered by the EVSE analysis.** (**a**) GWAS data set: Each GWAS hit is expanded into an associated variant set. The AVS is depicted as a red triangle. It contains multiple SNPs in linkage with the tagging risk SNP. (**b**) Gene expression data set: The network and the regulons are computed from gene expression data on the basis of mutual information assigning a set of target genes (black circles) to any given TF. Example mutual information values are shown. (**c**) Gene expression and genotyping data set: A multivariate eQTL analysis is carried out between the AVS and any gene in a given regulon that is present within a +/- 250kb window of the AVS. (**d**) Each time an association is detected between an AVS and a regulon the locus is counted towards a mapping tally, visualised by black/grey boxes. (**e**) Lastly, the statistical significance is assessed by comparing the observed mapping tally to a null distribution based on random permutations of the AVS (that is, matched random variant sets). The enrichment score is obtained by subtracting the mean of the null distribution from the mapping tally and dividing by the standard deviation. The normalised null distribution is then be used to calculate the empirical p-values.

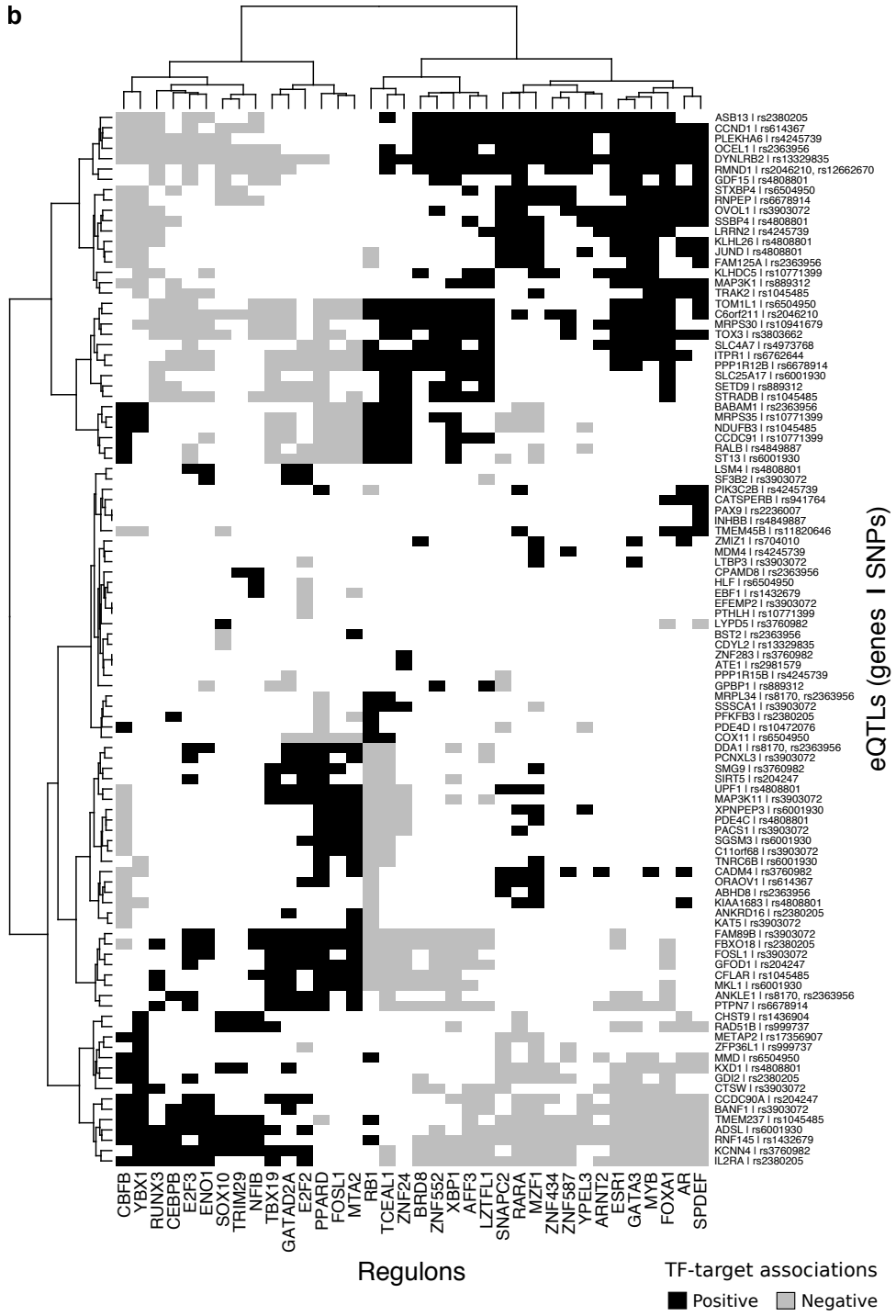

a
Clusters of risk−associated and linked SNPs (Michailidou et al.)
Annotation
P value (-log10)
Enrichment score (EVSE)
Bonferroni
Regulon
markers
genes
mapping tally


b
Clusters of risk−associated and linked SNPs (Michailidou et al.)
Annotation
P value (-log10)
Enrichment score (EVSE)
Bonferroni
Regulon
markers
genes
mapping tally

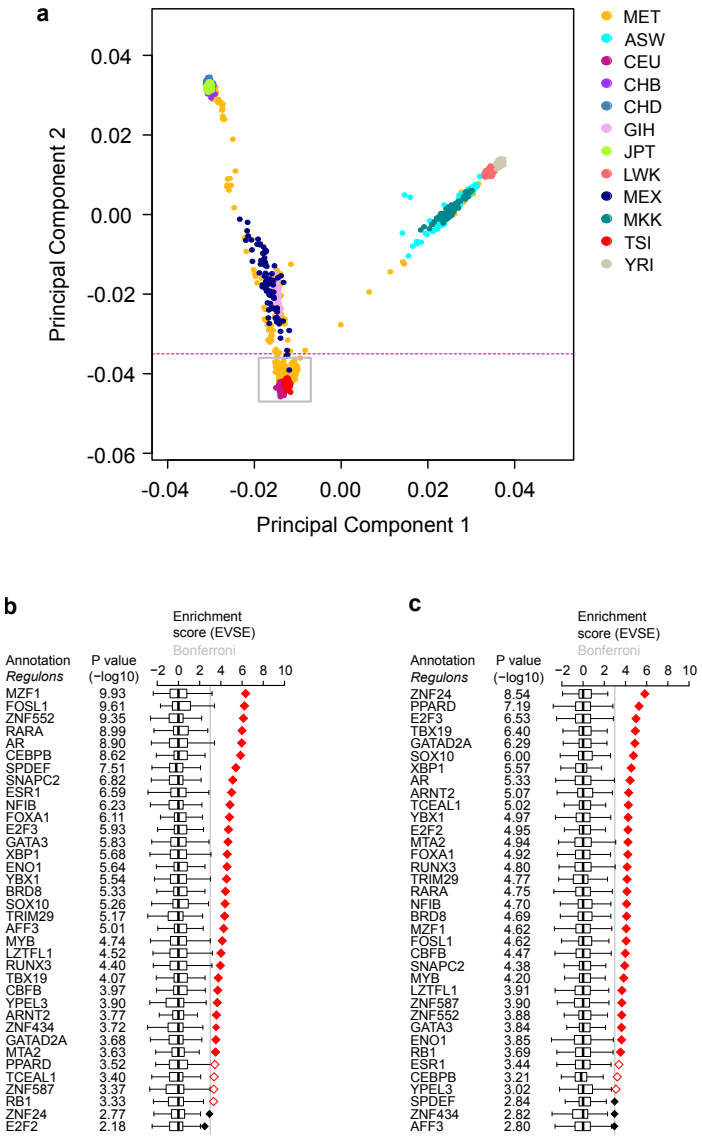


Supplementary Figure 3: **EVSE-based identification of risk-TFs.** This figure shows the results of the EVSE analysis for all those TFs that were not removed in the low resolution EVSE analysis. The EVSE analysis was carried out independently for cohort I (**a**) and cohort II (**b**) of the METABRIC sample set. The panels list the enrichment score and associated p-value for each TF-regulon, next to the enrichment scores (boxplot) obtained using the same regulon in an EVSE using random AVSs (size matched to the breast cancer AVS). Solid and open red diamonds highlight mapping tallies that satisfy a Bonferroni-corrected threshold for significance of $P<0.01$ and $P<0.05$, respectively. P-values are based on null distributions from 1,000 random AVSs. Non-significant enrichment scores are shown as black diamonds. The tagging SNP for each breast cancer GWAS hit[5] is listed above the matrix and the number of markers (SNPs in the AVS for which genotypes were available in METABRIC) obtained for each locus is given beneath the tagging SNP. The matrix shows the summary of the mapping tally, with significant results highlighted by a darker grey colour. Mapping tallies are summed on the right of the matrix, while the number of genes within a +/- 250kb window of the AVS is indicated at the bottom. The 36 TF-regulons with significant enrichment scores in both cohort I and II are also shown in Figure 1. (**c**) Analysis using normal tissue. For the 36 TFs that were significant in both cohorts, the EVSE analysis was run using gene expression data from 144 normal samples from the METABRIC data set. Annotations are as for a and b.

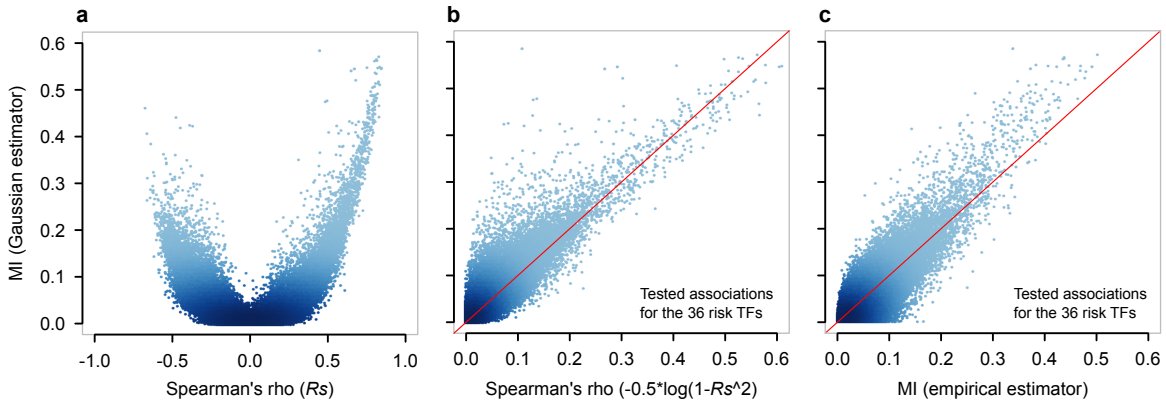

a
SSBP4 | rs4808801
CASZ1 | rs616488
ASB13 | rs2380205
CCND1 | rs614367
OCEL1 | rs8170, rs2363956
BHLHE40 | rs6762644
DYNLRB2 | rs13329835
MRPS30 | rs10941679
ITPR1 | rs6762644
TOM1L1 | rs6504950
COMMD3 | rs11814448
CCDC91 | rs10771399
NTN4 | rs17356907
SUMF1 | rs6762644
TOX3 | rs3803662
DNAJC1 | rs7072776
ZMIZ1 | rs704010
MDM4 | rs4245739
SLC4A7 | rs4973768
MAP3K1 | rs889312
LRRN2 | rs4245739
XRCC1 | rs3760982
FAM125A | rs2363956
KLHL26 | rs4808801
NR2F6 | rs8170, rs2363956
STXBP4 | rs6504950
RNPEP | rs6678914
C14orf159 | rs941764
MIER3 | rs889312
NDUFB3 | rs1045485
ST13 | rs6001930
MRPS35 | rs10771399
CASP8 | rs1045485
COX11 | rs6504950
ZAK | rs1550623
AKAP12 | rs12662670
SIPA1 | rs3903072
UPF1 | rs4808801
MUS81 | rs3903072
PDE4C | rs4808801
DRAP1 | rs3903072
C11orf68 | rs3903072
PLVAP | rs8170, rs2363956
MKL1 | rs6001930
CFLAR | rs1045485
DDA1 | rs8170, rs2363956
ANKLE1 | rs8170, rs2363956
CFL1 | rs3903072
NOL7 | rs204247
MMD | rs6504950
UBE2T | rs6678914
BANF1 | rs3903072
FOXF2 | rs11242675
CHST9 | rs1436904
LYPD5 | rs3760982
CRTC1 | rs4808801
PIK3C2B | rs4245739
PAX9 | rs2236007
ZNF224 | rs3760982
NXNL1 | rs2363956
KXD1 | rs4808801
SLC25A17 | rs6001930
SETD9 | rs889312
ELL | rs4808801
GTPBP3 | rs8170, rs2363956
FIBP | rs3903072
UBA52 | rs4808801
MRPL34 | rs8170, rs2363956
ORAOV1 | rs614367
SIRT5 | rs204247
ZNF428 | rs3760982
CPAMD8 | rs2363956, rs8170
EBF1 | rs1432679
FTO | rs11075995
FBXO18 | rs2380205
HAUS8 | rs2363956
PTPN7 | rs6678914
RAD51B | rs2588809, rs999737
TMEM237 | rs1045485
CTSW | rs3903072
GDI2 | rs2380205
CCDC90A | rs204247
KCNN4 | rs3760982
CDKN2A | rs1011970
RNF145 | rs1432679
IFI30 | rs4808801
IL2RA | rs2380205
eQTLs (genes I SNPs)
TBX19
FOSL1
PPARD
MTA2
RUNX3
CEBPB
ENO1
E2F3
GATAD2A
E2F2
YBX1
CBFB
TRIM29
SOX10
NFIB
ZNF434
RARA
SNAPC2
MZF1
YPEL3
ZNF587
ARNT2
AR
SPDEF
BRD8
LZTFL1
MYB
FOXA1
GATA3
ESR1
ZNF552
AFF3
XBP1
RB1
ZNF24
TCEAL1
Regulons
TF-target associations
Positive
Negative



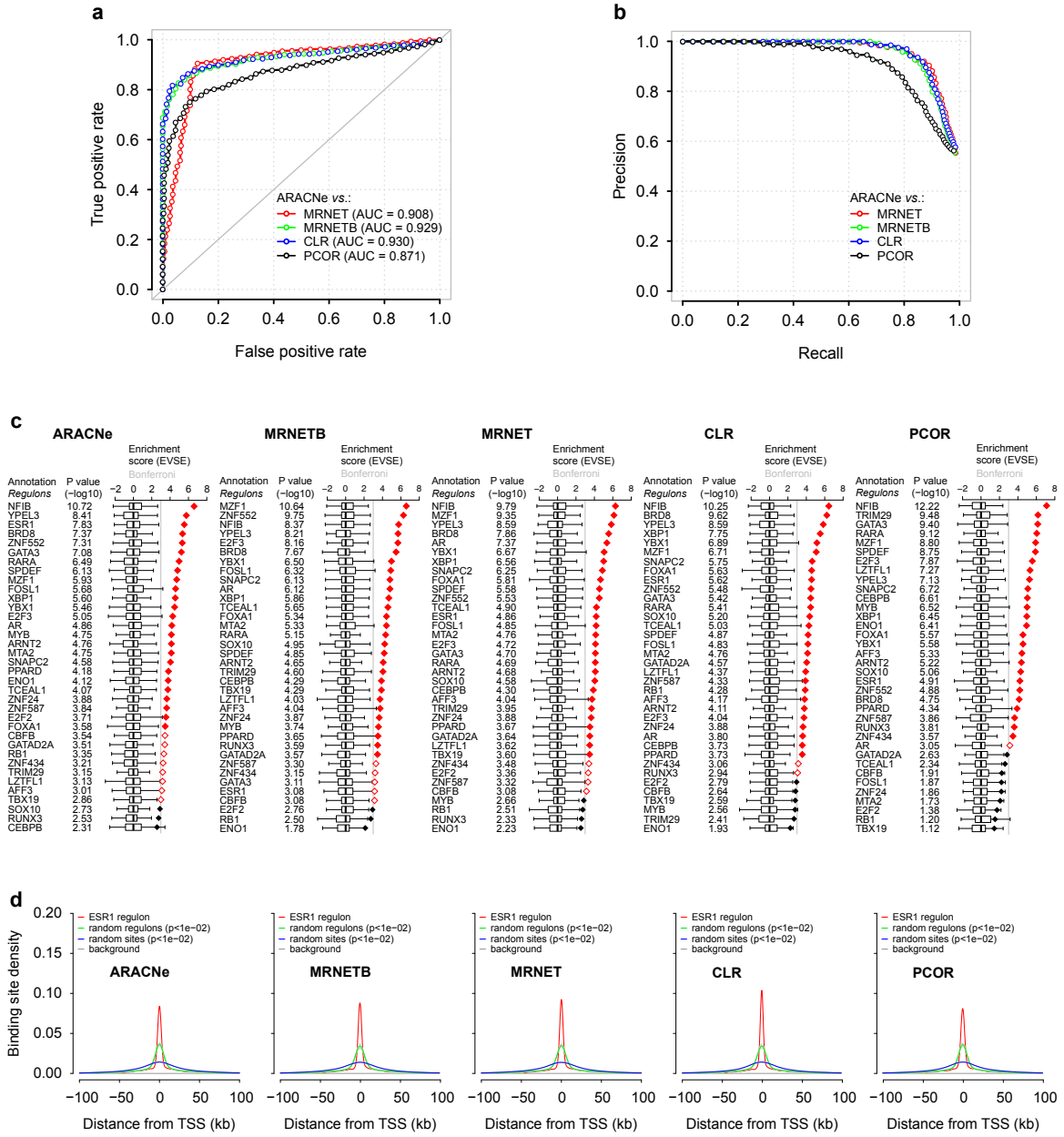
Supplementary Figure 4: **Tabulated eQTLs.** Graphical representation of genes and breast cancer risk loci mapped in the EVSE analysis for METABRIC cohort I (**a**) and II (**b**). Each square represents a TF-target association (positive or negative) indicating whether there is a positive or negative correlation between the TF and the target gene listed along the right hand of the matrix. Each gene is listed next to the tagging SNP denoting the AVS that led to the eQTL.

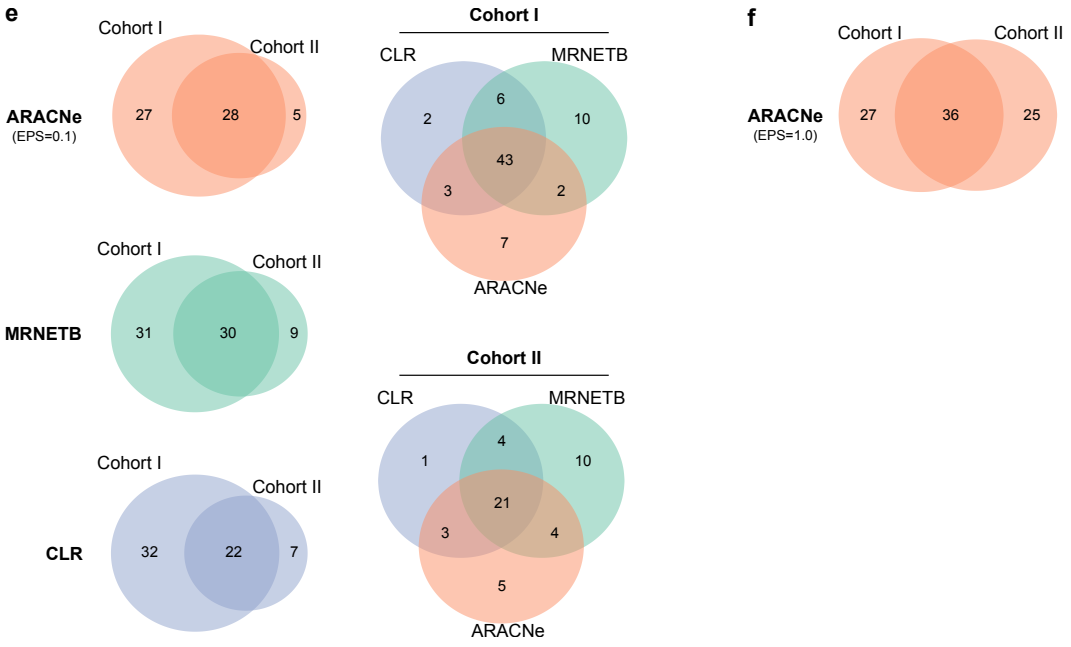


Supplementary Figure 5: **METABRIC samples with homogeneous genetic background.** (**a**) Projection of the METABRIC Samples onto the top two principal components of the HapMap genotypes. Most METABRIC samples (yellow) cluster around the coordinates (-0.015, -0.042), close to the HapMap CEU samples of Northern and Western European ancestry and the HapMap TSI samples of Tuscan ancestry. This square encompasses over 90% of the samples in the METABRIC cohort. All METABRIC samples above the red dashed line were excluded from the analysis. Abbreviations: MET: METABRIC, ASW: African ancestry in Southwest USA, CEU: Utah residents with Northern and Western European ancestry from the CEPH collection, CHB: Han Chinese in Beijing (China), CHD: Chinese in Metropolitan Denver (Colorado), GIH: Gujarati Indians in Houston (Texas), JPT: Japanese in Tokyo (Japan), LWK: Luhya in Webuye (Kenya), MEX: Mexican ancestry in Los Angeles (California), MKK: Maasai in Kinyawa (Kenya), TSI: Toscani in Italia, YRI: Yoruba in Ibadan (Nigeria). Results of the EVSE analysis carried out for the breast cancer AVS and regulons calculated for cohort I (**b**) and cohort II (**c**). The panels list the enrichment score and associated p-value, next to the enrichment scores (boxplot). Solid and open red diamonds indicate significant enrichment scores that satisfy a Bonferroni-corrected threshold for significance of $P<0.05$ and $P<0.01$, respectively. P-values are based on null distributions from 1,000 random AVSs. All other annotations are as for Supplementary Figure 3.

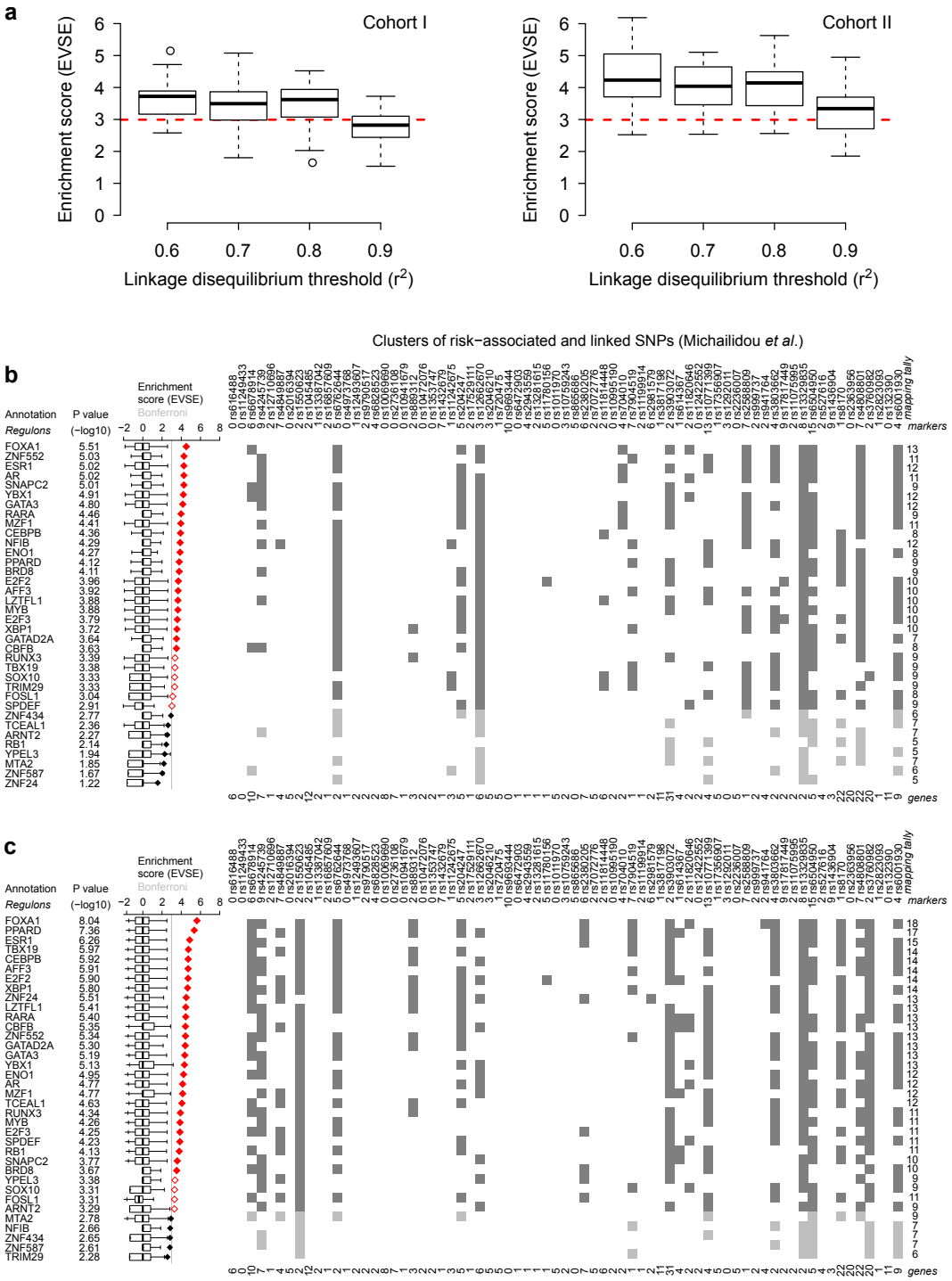


Supplementary Figure 6: **Different choices of association metrics.** (**a-c**) Scatterplots showing the relationship between two sets of data computed by the indicated statistics. It represents all tested associations for the 36 risk TFs (colors indicate the local densities at each point). Raw values for Spearman's correlation are shown in the first panel, and in order to share the scale among the estimators the corresponding MI transformation is provided. The distributions differ in some extent, but they are highly correlated (P<2.2e-16, F-statistic).

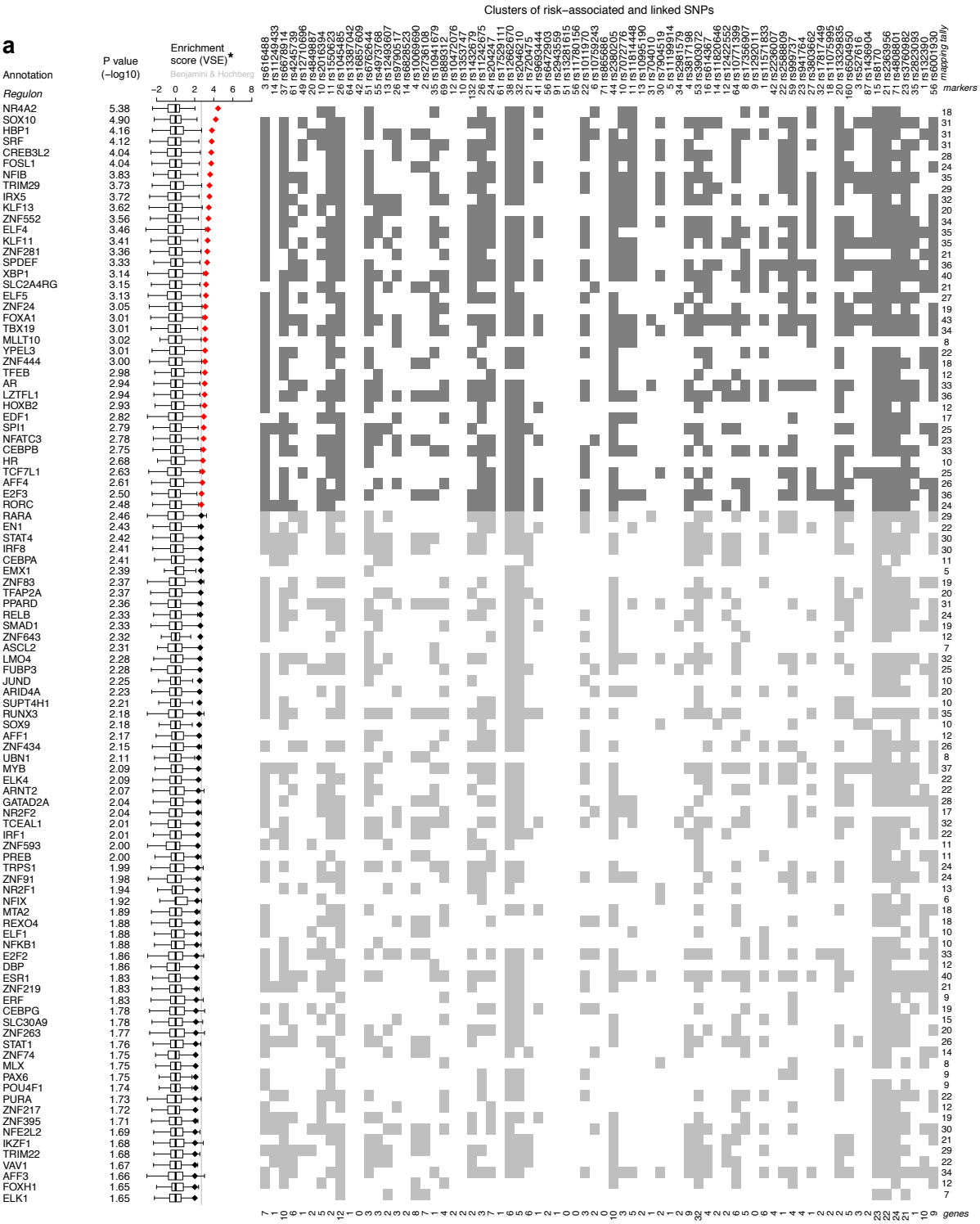

a
True positive rate
False positive rate
ARACNe vs.:
MRNET (AUC = 0.908)
MRNETB (AUC = 0.929)
CLR (AUC = 0.930)
PCOR (AUC = 0.871)
b
Precision
Recall
ARACNe vs.:
MRNET
MRNETB
CLR
PCOR
c
ARACNe
MRNETB
MRNET
CLR
PCOR
Enrichment score (EVSE)
Bonferroni
Annotation
Regulons
P value (−log10)
d
Binding site density
ESR1 regulon
random regulons (p<1e−02)
random sites (p<1e−02)
background
ARACNe
MRNETB
MRNET
CLR
PCOR
Distance from TSS (kb)



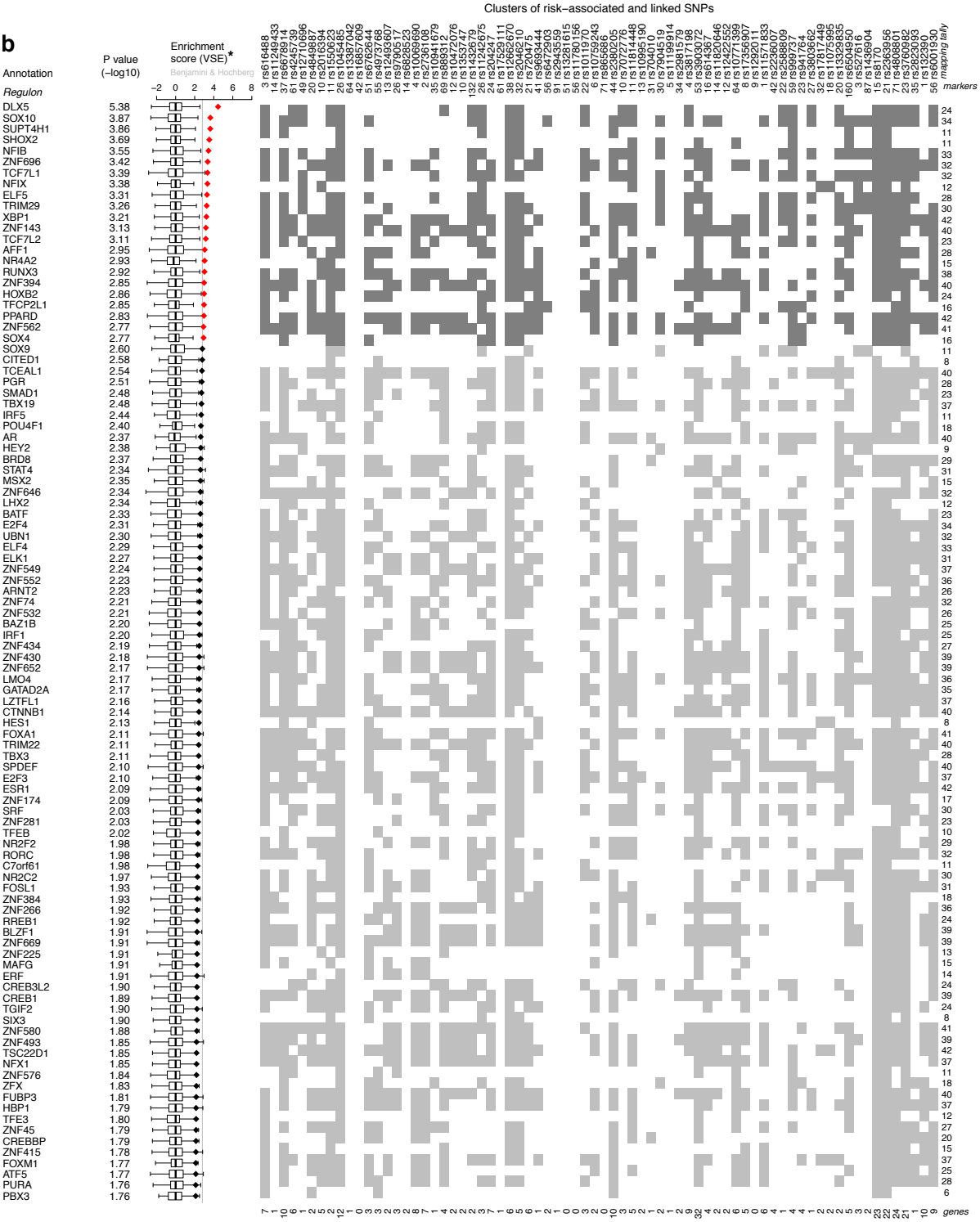
Supplementary Figure 7: **EVSE analysis using regulons inferred by different algorithms.** (**a-b**) Several competing algorithms were used to filter the same input adjacency matrix, which was computed for all TFs and their potential target genes using cohort I of the METABRIC data set: Maximum Relevance Minimum Redundancy (MRNET), Maximum Relevance Minimum Redundancy Backward (MRNETB), Context Likelihood (CLR) and the Partial Correlation (PCOR). ARACNe is used as the underlying reference network to compute the Receiver Operating Characteristic (ROC) curves. Additionally, Precision-Recall curves show the performance of the other algorithms to retrieve the relevant associations inferred by ARACNe. (**c**) EVSE analysis using the breast cancer AVS was carried out independently for the 36 risk-TFs defined in **Figure 1**. The panels list the enrichment score and associated p-value, next to the enrichment scores (boxplot). Solid and open red diamonds indicate significant enrichment scores that satisfy a Bonferroni-corrected threshold for significance of $P<0.05$ and $P<0.01$, respectively. P-values are based on null distributions from 1,000 random AVSs. All other annotations are as for Supplementary Figure 3. (**d**) The robustness of the regulatory network construction should also rely on experimental validation, and we validated 5 regulons derived from this same network construction approach in a previous study using ChIP-seq data obtained in MCF-7 cells (Fletcher *et al.*, 2013). The density distribution plots reproduce the results for the ESR1 regulon. They show the distribution of estrogen binding sites near to the transcription start sites (TSS) of the genes in the ESR1 regulons inferred by the indicated network construction algorithms (each distribution is compared with random regulons and random sites). A background distribution is also shown as a reference line (grey line) and represents the distance between the TSSs and random peaks. (**e**) The Venn diagrams show the results from EVSE analyses expanded to all regulons, with the overlap of the results obtained for MRNETB, CLR, and ARACNe, each in cohort I and II (these results are discussed in detail on page 8). (**f**) Results from the EVSE analysis described in Supplementary Figure 3 showing the intersect that produced the 36 TFs that were significant in both cohorts.



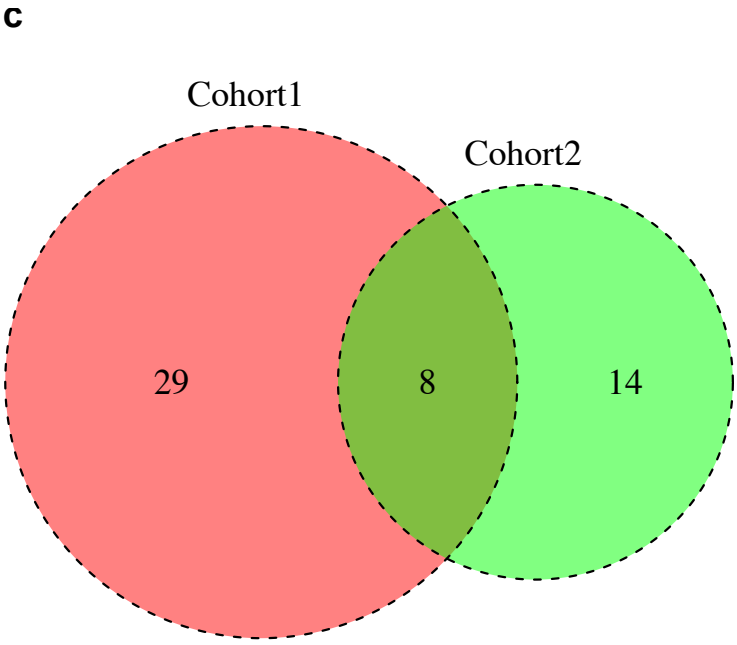
Supplementary Figure 8: **EVSE analysis using r^2 measure of linkage disequilibrium.** Computational validation of the EVSE analysis for the 36 risk-TFs using r^2 measure of linkage disequilibrium to derive the breast cancer AVS. (**a**) Averages of enrichment scores obtained with different r^2 thresholds. (**b-c**) EVSE analysis with $r^2 \geq 0.8$, showing enrichment score and p-value for cohort I (**b**) and cohort II (**c**) from the METABRIC dataset. Solid and open red diamonds indicate significant enrichment scores that satisfy a Bonferroni-corrected threshold for significance of $P<0.05$ and $P<0.01$, respectively. P-values are based on null distributions from 1,000 random AVSs. All other annotations are as for Supplementary Figure 3.



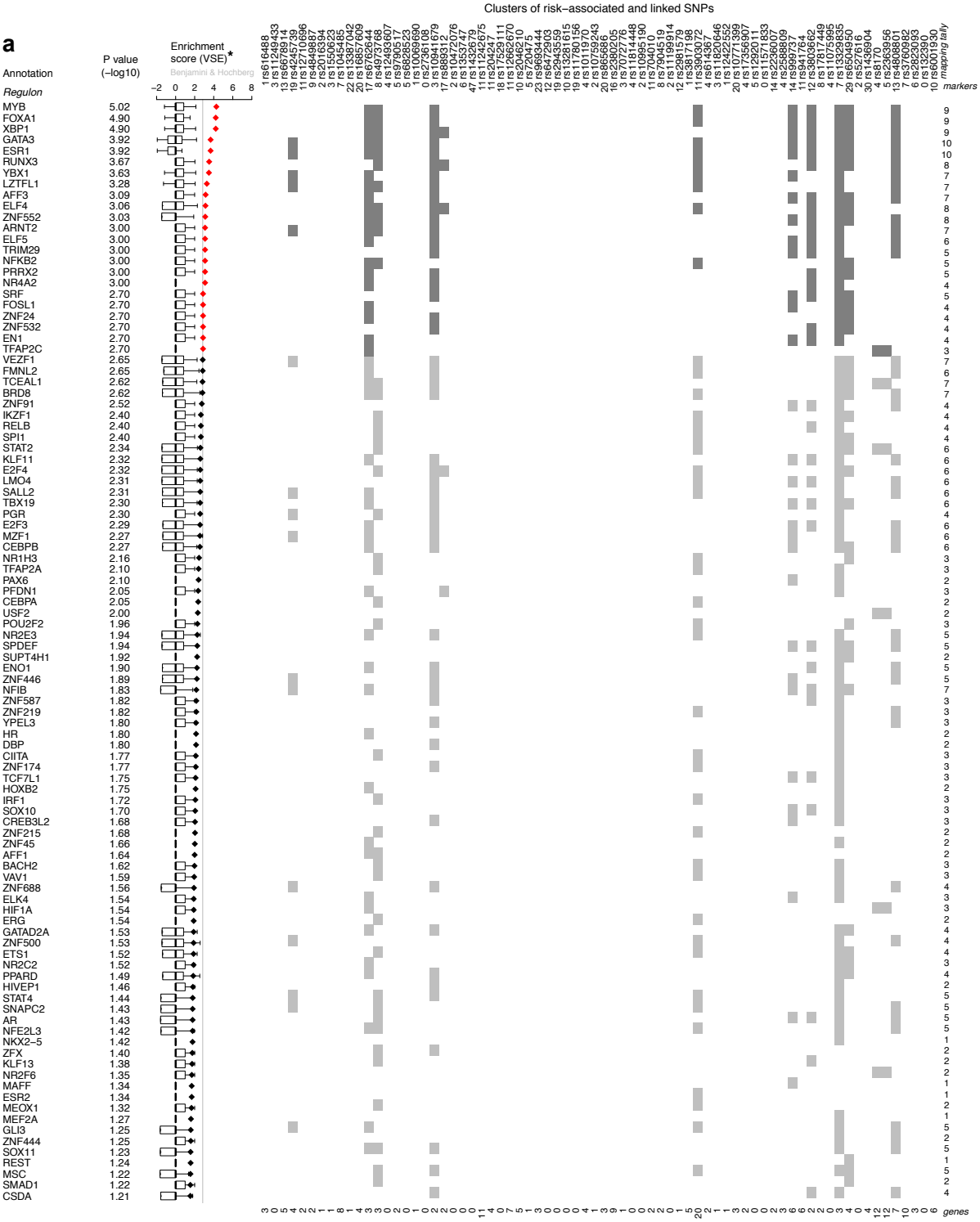
** VSE testing gene lists*



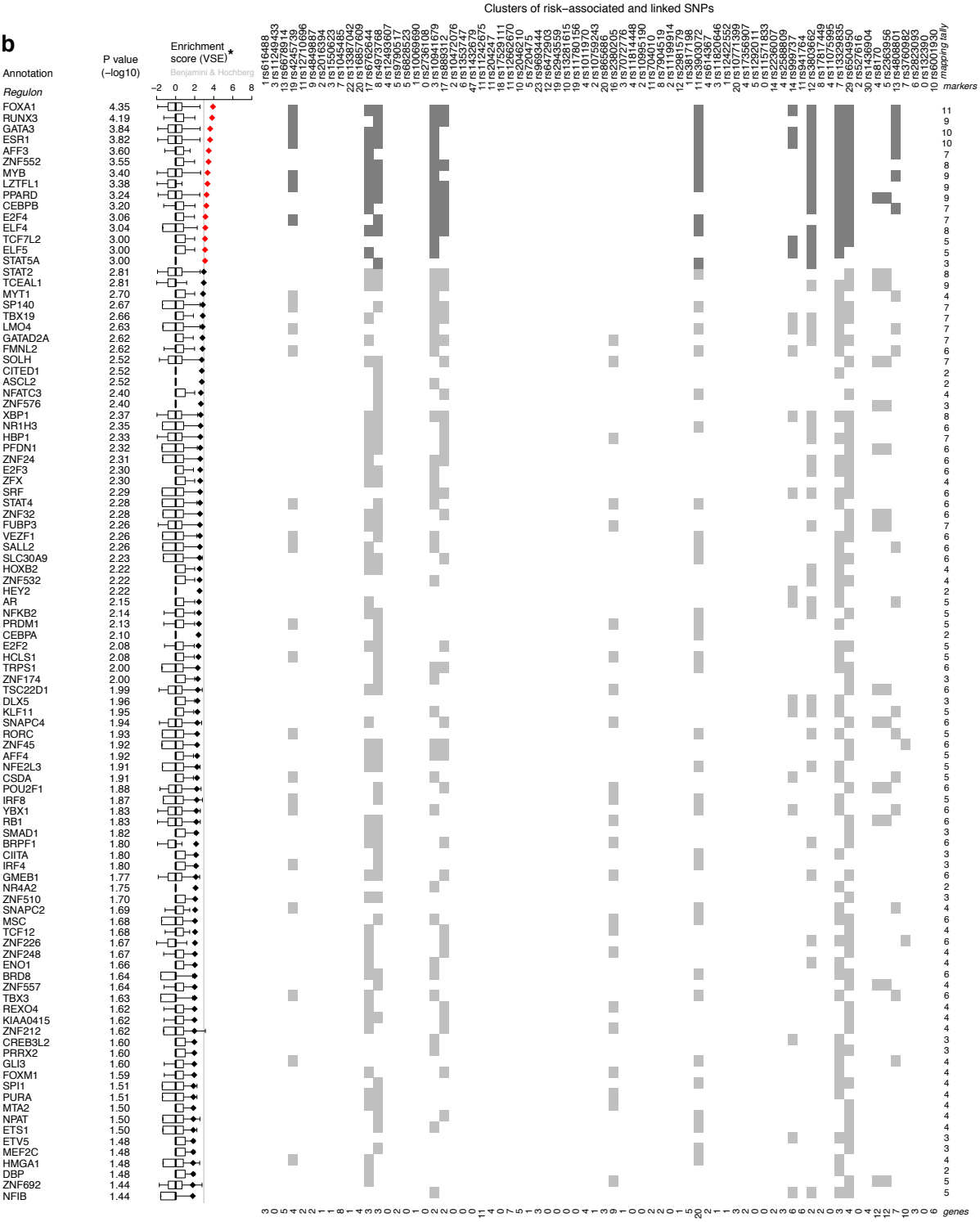
** VSE testing gene lists*



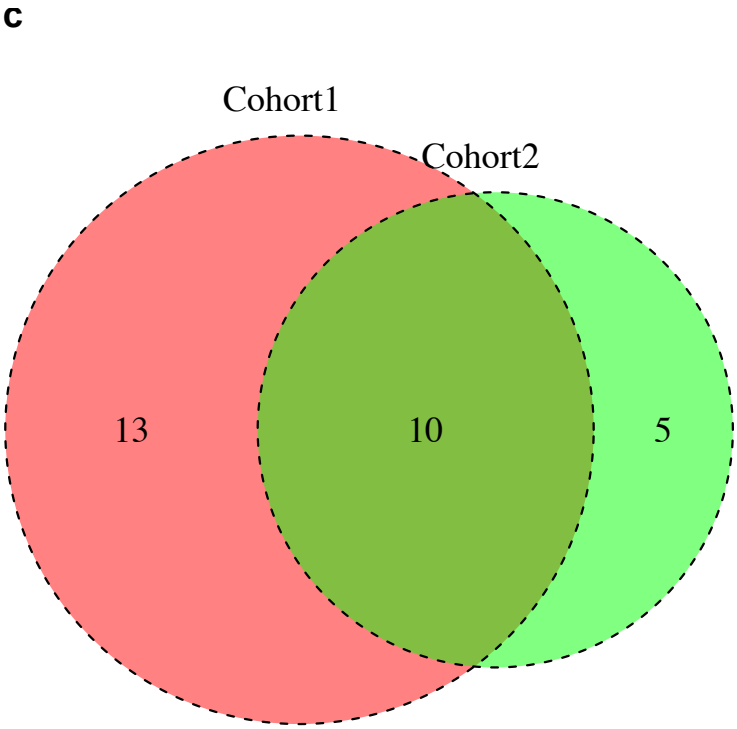
Supplementary Figure 9: **Distance-based VSE analysis.** Results of the distance-based VSE analysis identifying TF regulons associated with breast cancer GWAS hits for (**a**) cohort I and (**b**) cohort II. All genes in +/-250kb windows around the AVS were considered. The number of SNPs (markers) is shown to indicate the extent to which the tagging SNP was expanded to generate the AVS, but the analysis is based on the window size alone. Red diamonds indicate significant enrichment scores with $P<0.05$. P-values are based on null distributions from 1,000 random AVSs. Since no eQTL step is executed in this analysis, the number of markers listed beneath the tagging SNP represents all markers in the AVS. All other annotations are as for Supplementary Figure 3. (**c**) The overlap of the results obtained in a and b is plotted as a Venn diagram.



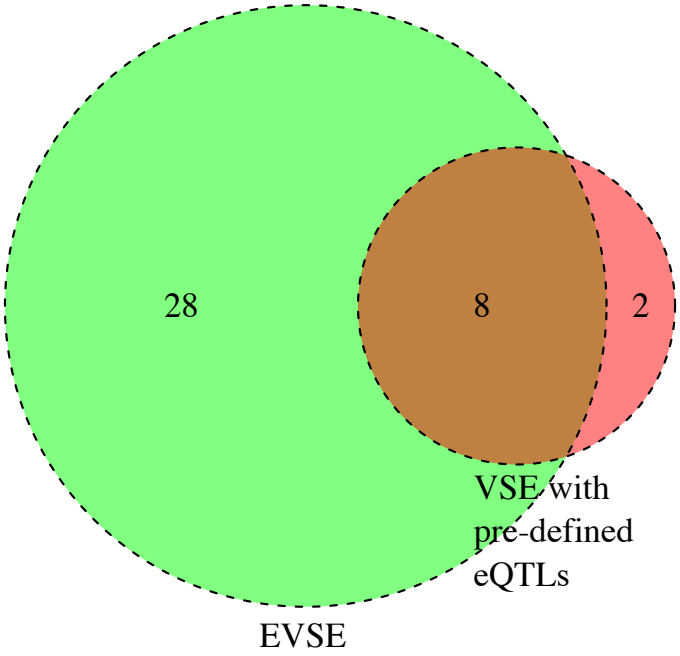
** VSE with pre-defined eQTLs*



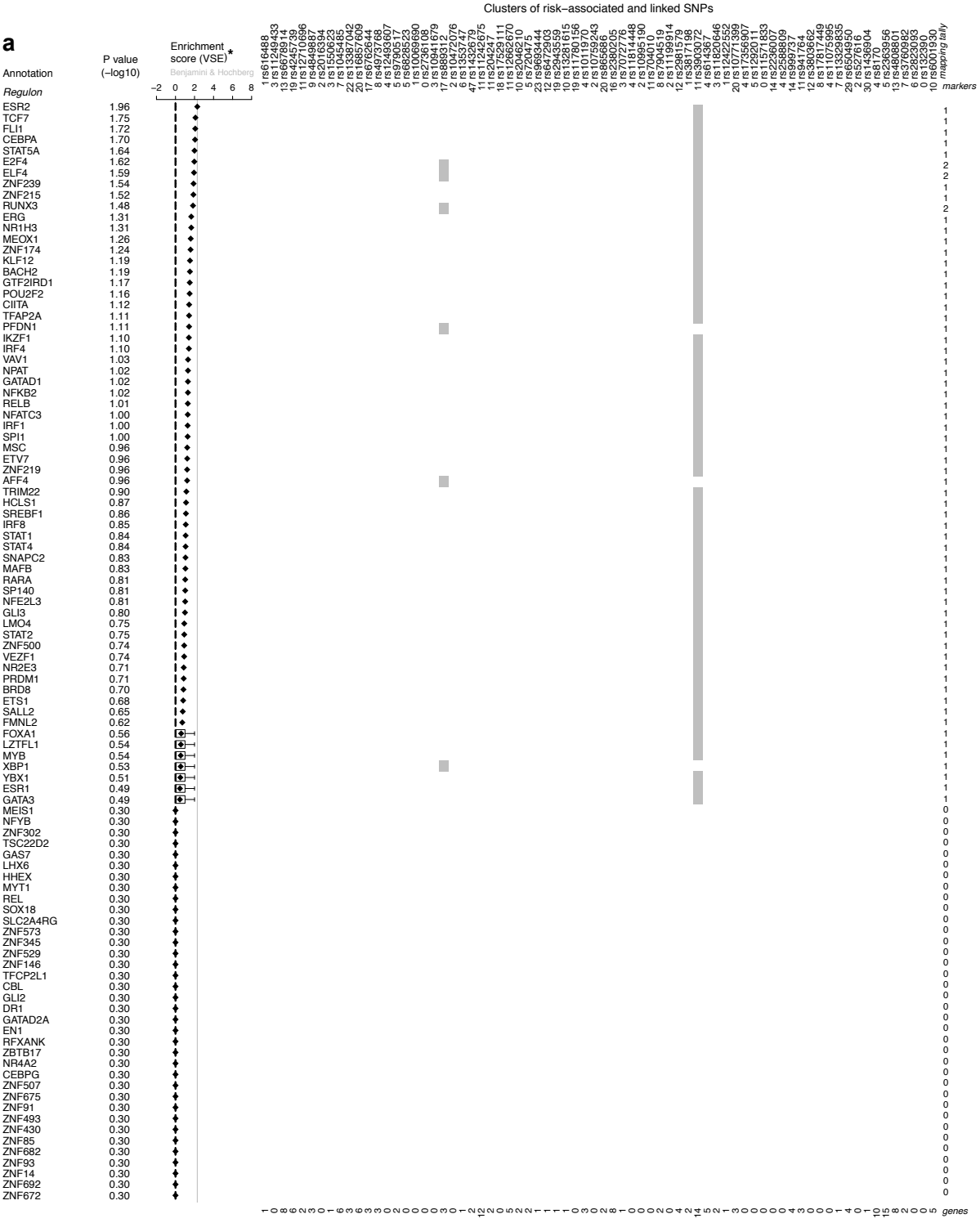
** VSE with pre-defined eQTLs*



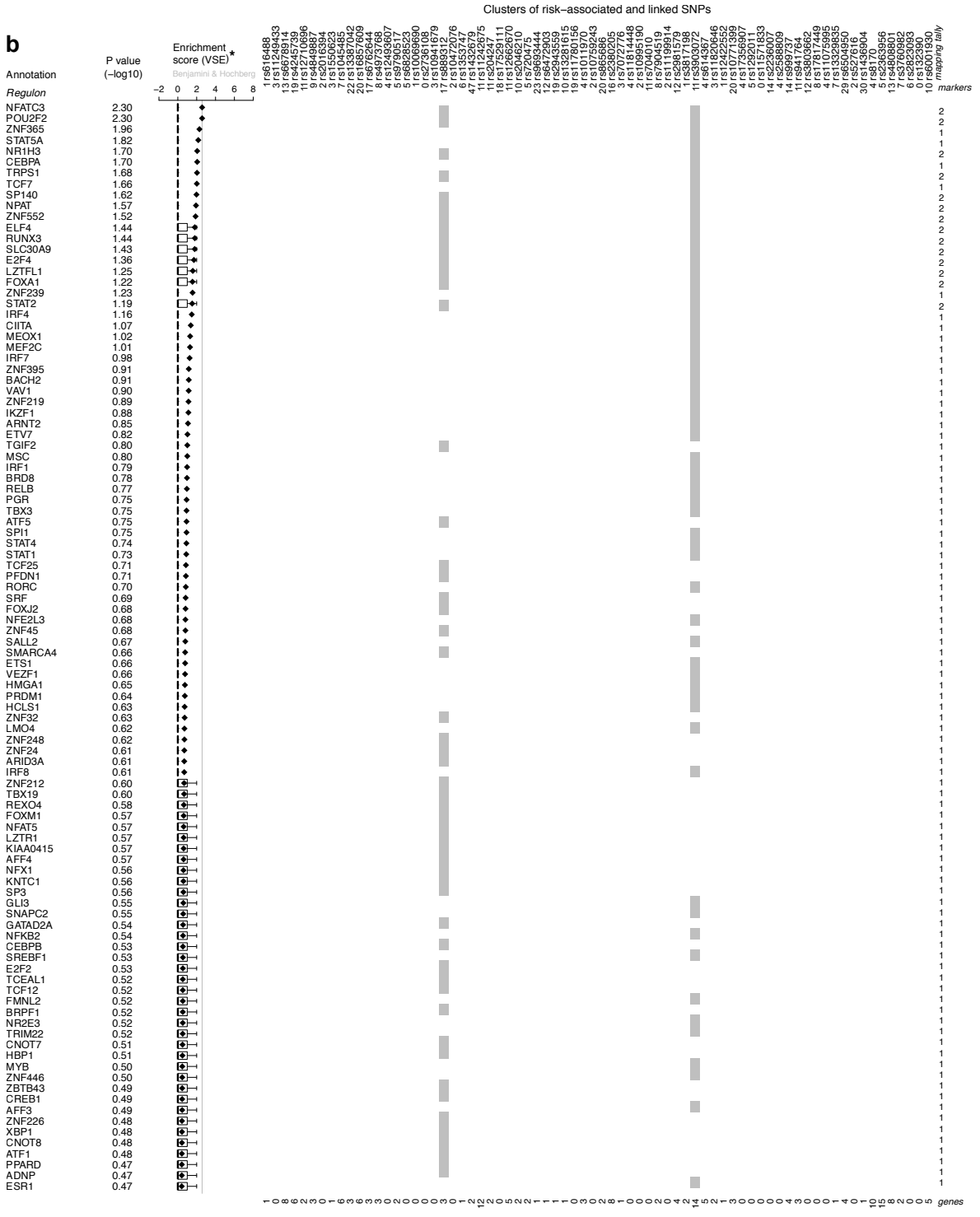
Supplementary Figure 10: **VSE analysis with pre-defined eQTLs for ER$^+$ tumours.** Risk associated TFs identified using the eQTL-conditioned VSE analysis with pre-defined eQTLs for ER$^+$ breast tumours from (**a**) cohort I and (**b**) cohort II of the METABRIC data set. Red diamonds indicate significant enrichment scores with $P<0.05$. P-values are based on null distributions from 1,000 random AVSs. Since pre-defined eQTLs are used in this analysis, the numbers beneath the tagging SNP refer to the markers in the input list, ie those genotyped in the METABRIC data set. All other annotations are as for Supplementary Figure 3, except that the gene list given at the bottom of the matrix represents the number of genes that were found to be linked to the AVS in the unconstrained eQTL calling. (**c**) The overlap of the results obtained in a and b is plotted as a Venn diagram.



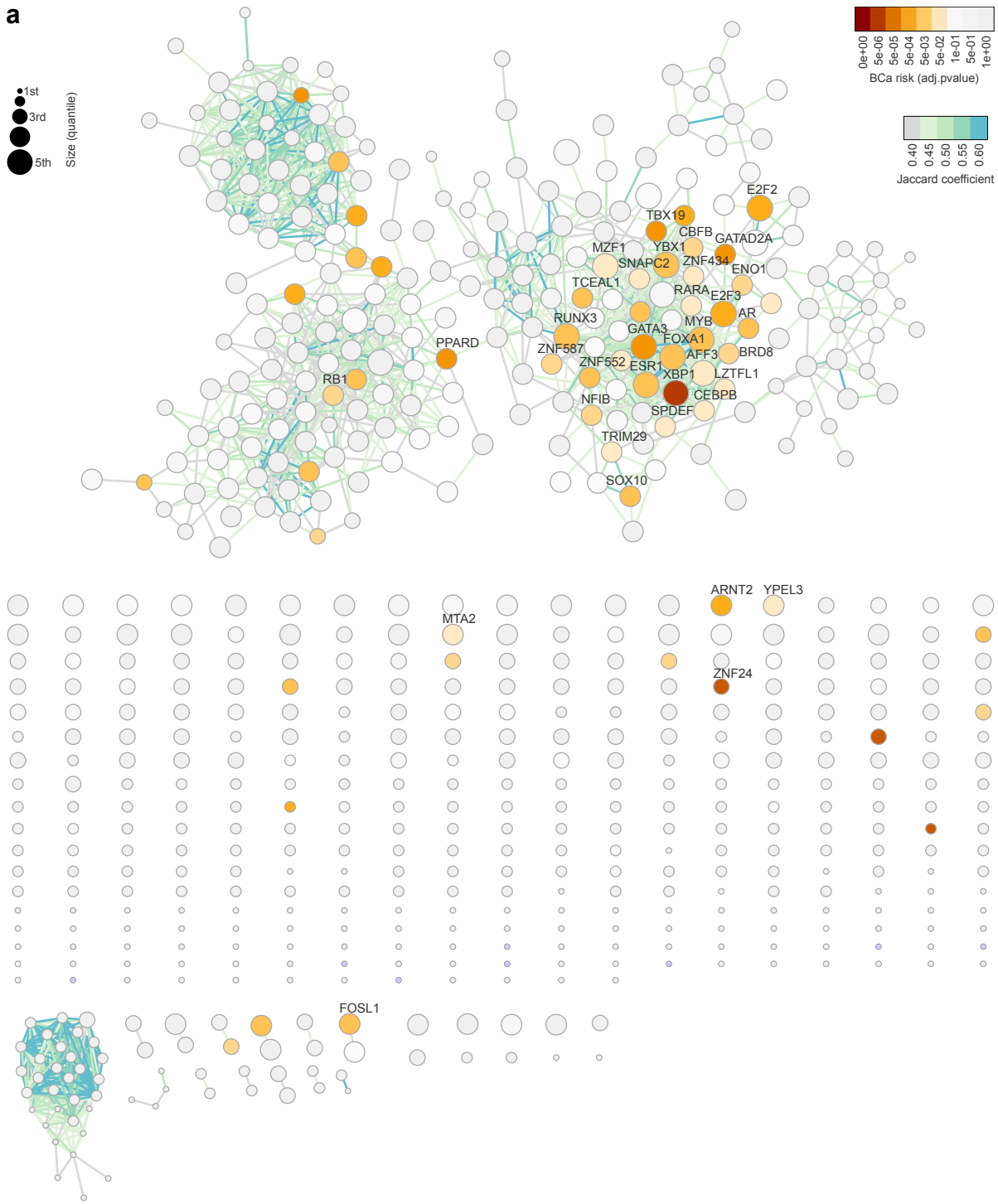
Supplementary Figure 11: **Venn diagram showing the overlap between EVSE and VSE with pre-defined eQTLs for ER$^+$ tumours.** The EVSE analysis identified 36 risk-TFs across cohort I and II (in green, listed in Figure 1a and b). The analogous analysis using eQTLs identified in an independent analysis yielded a consensus of 10 TFs across cohort I and II (in red, Supplementary Figure 10c). The Venn diagram shows the overlap of these two groups.

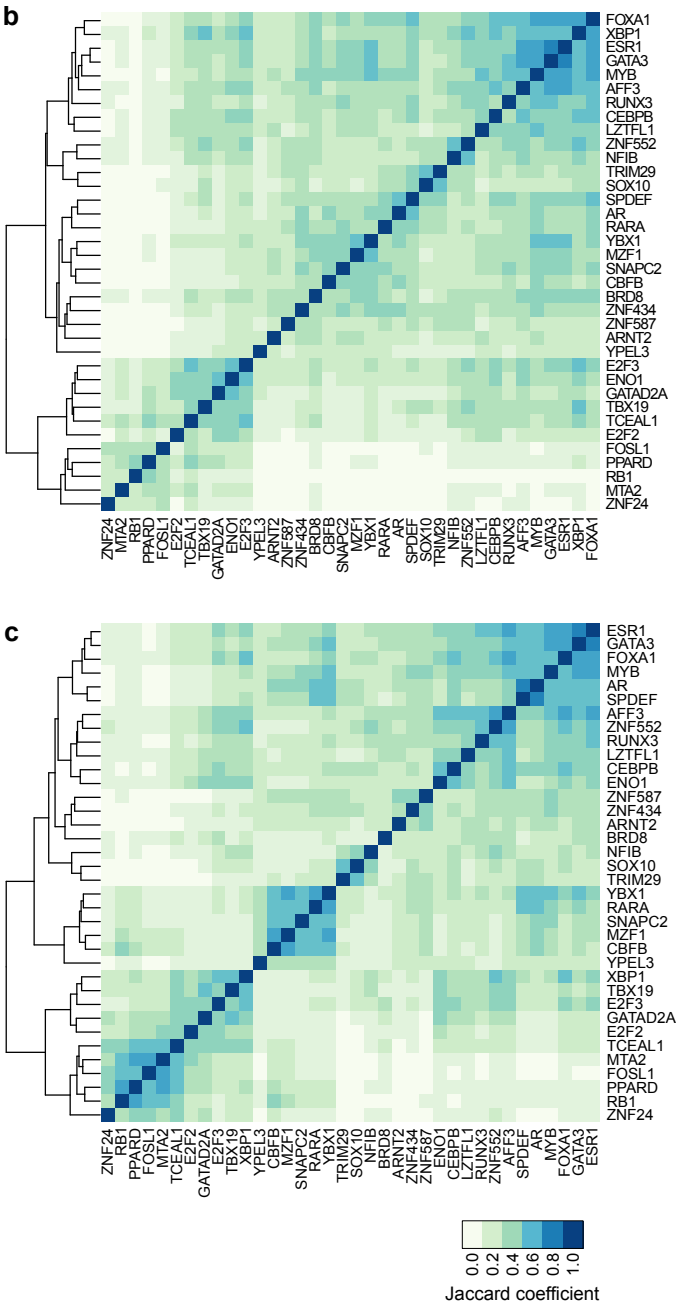


** VSE with pre-defined eQTLs*

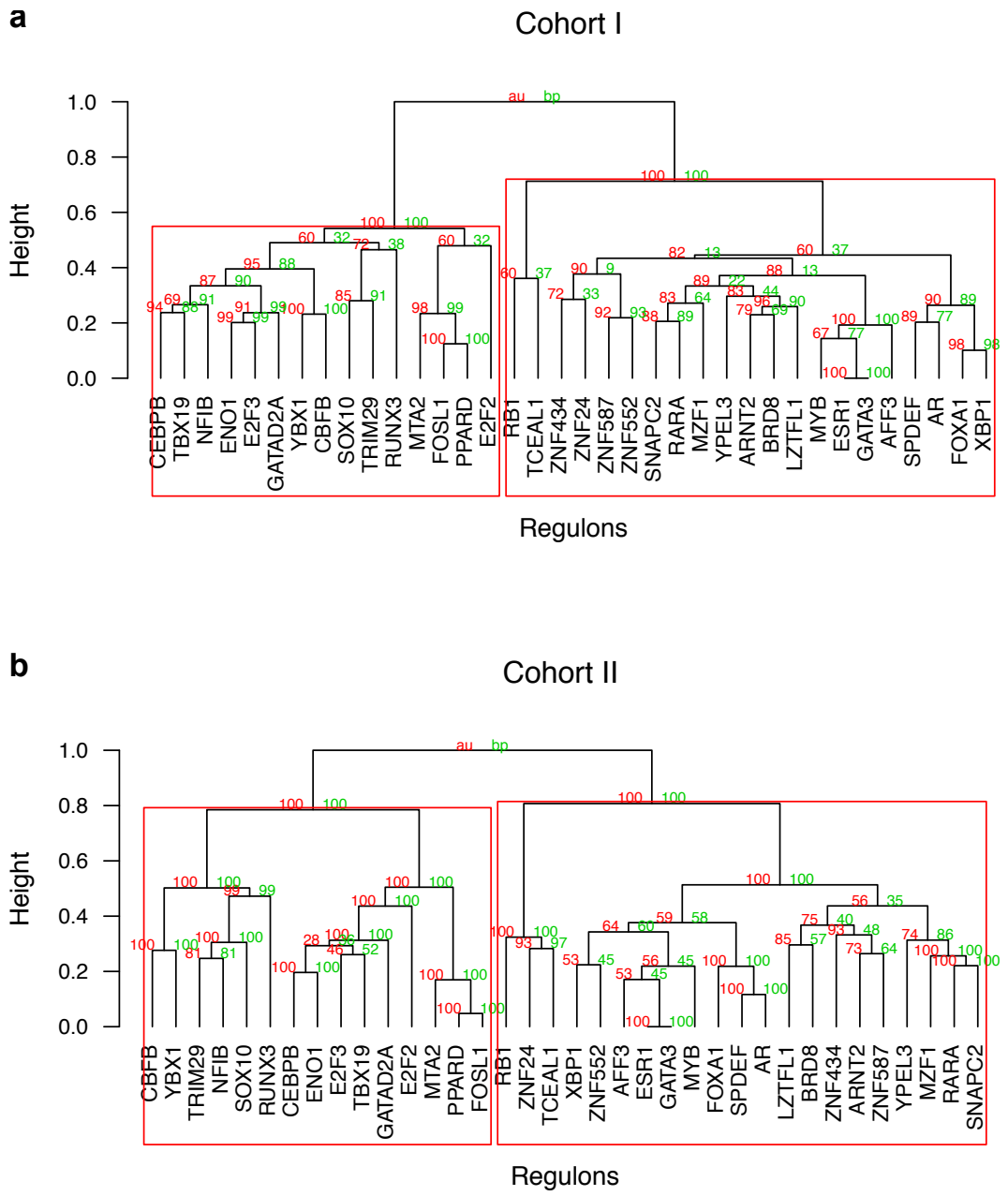


Supplementary Figure 12: **VSE analysis with pre-defined eQTLs for ER$^-$ tumours.** Risk-TFs identified using a VSE analysis with pre-defined eQTLs for ER$^-$ breast tumours from (**a**) cohort I and (**b**) cohort II of METABRIC (all other annotations are as for Supplementary Figure 10).

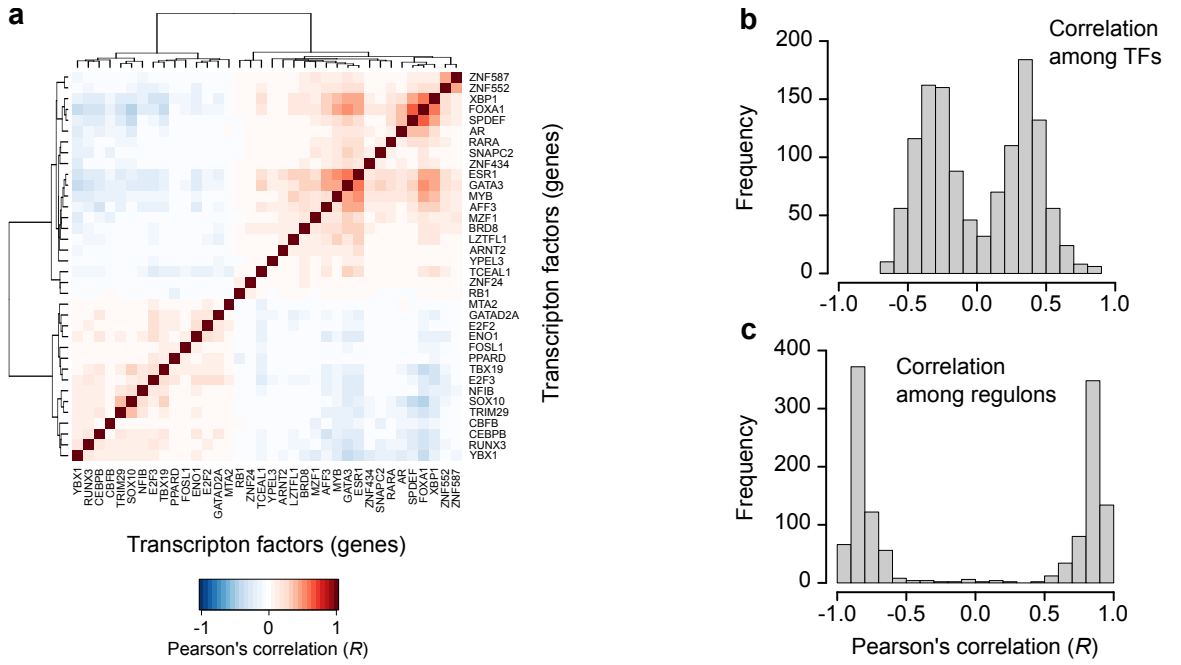

a
1st
3rd
5th
Size (quantile)
0e+00
5e-06
5e-05
5e-04
5e-03
5e-02
1e-01
5e-01
1e+00
BCa risk (adj.pvalue)
0.40
0.45
0.50
0.55
0.60
Jaccard coefficient
E2F2
TBX19
CBFB
GATAD2A
MZF1
YBX1
SNAPC2
ZNF434
ENO1
TCEAL1
RARA
E2F3
AR
RUNX3
GATA3
MYB
FOXA1
PPARD
ZNF587
AFF3
BRD8
ZNF552
ESR1
XBP1
LZTFL1
RB1
NFIB
CEBPB
SPDEF
TRIM29
SOX10
ARNT2
YPEL3
MTA2
ZNF24
FOSL1



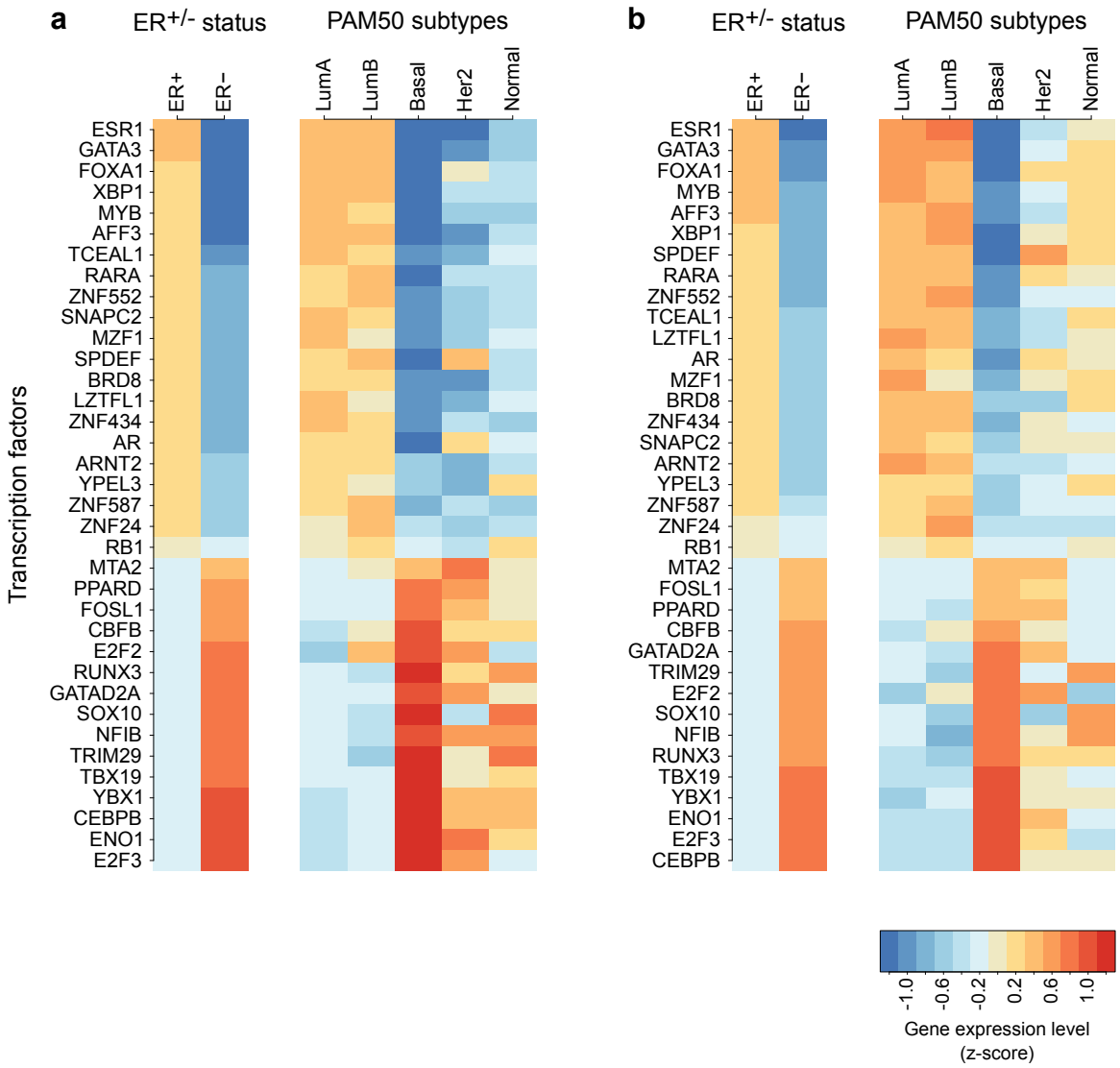
Supplementary Figure 13: **Regulatory network and hierarchical clustering on the Jaccard similarity coefficient.** (**a**) To avoid the use of the same data for calculating p-values and regulon overlap, we show the regulatory network based on cohort I, with risk association shown in yellow to red (based on cohort II of METABRIC). The 36 consensus risk-TFs identified in both cohorts are labelled. The colouring of the edges (shown in light green to blue) indicates the overlap as measured by Jaccard coefficient (JC) and size of circles represents the size of each regulon. In contrast to Figure 3, we also display those regulons not linked to the main network, based on a cut-off value of JC$\geq$0.4. (**b-c**) Hierarchical clustering on the Jaccard similarity coefficient focused on the overlap between the 36 risk TF-regulons inferred from the METABRIC cohort I (b) and II (c). Note that the Jaccard coefficient represents the fraction of common targets among the regulons and, therefore, does not take into account the directionality of the TF-target associations.



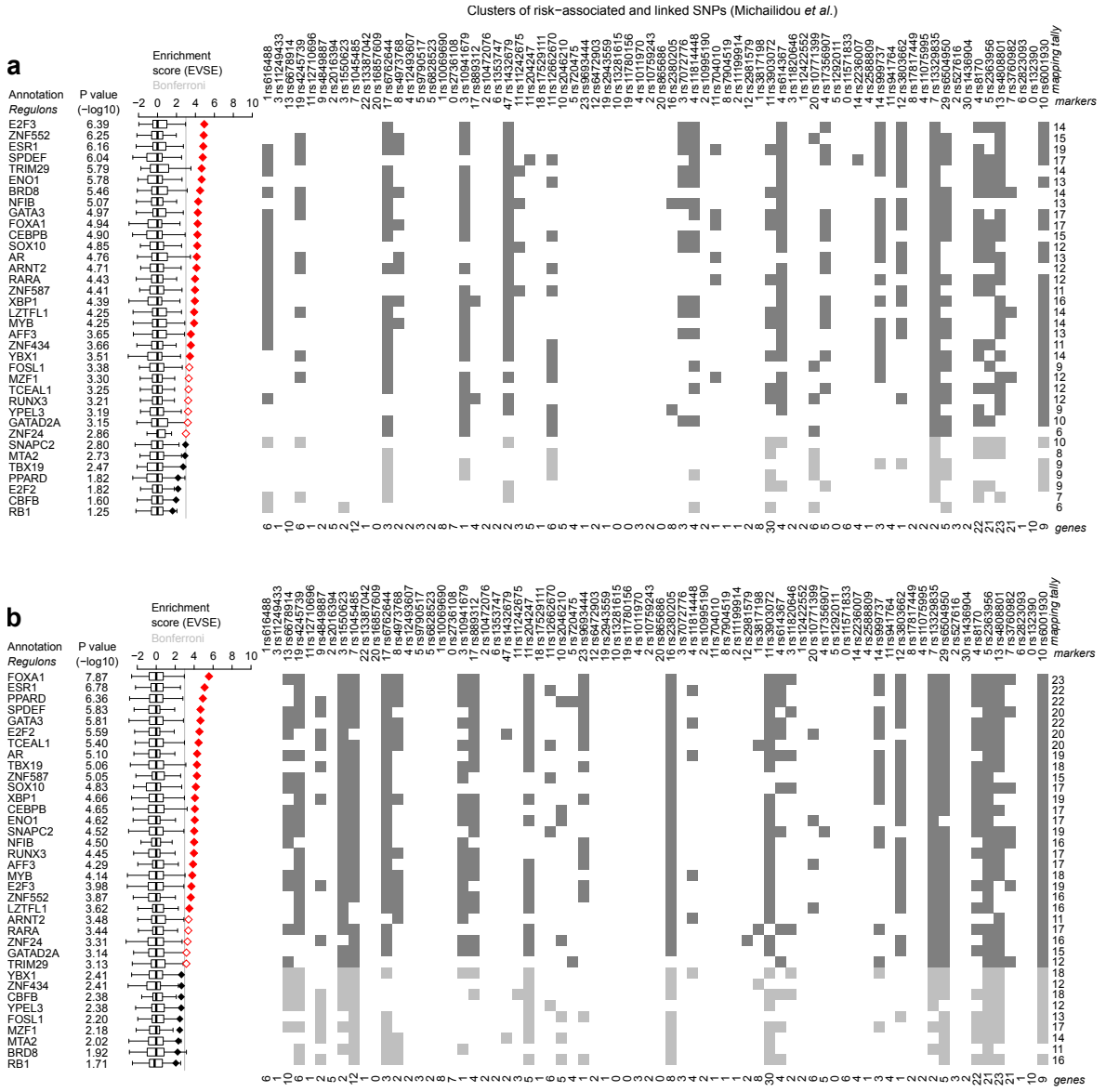
Supplementary Figure 14: **Stability of the clustering of the heat map depicted in Figure 4e.** Dendrograms were obtained as described for Figure 4 and unsupervised clustering results obtained for cohort I (**a**) and II (**b**) in multi-scale bootstrap-resampling are depicted. AU: approximate unbiased p-values (in red); BP: bootstrap probability p-values (in green). The analysis was performed using the R package *pvclust*[66], which executes a bootstrap analysis (n=1000) and counts how many times a given cluster in the hierarchical clustering can be observed from the bootstrap subsamples (the AU and BP values indicate how strongly the cluster is supported by the data and are expressed as percentages).



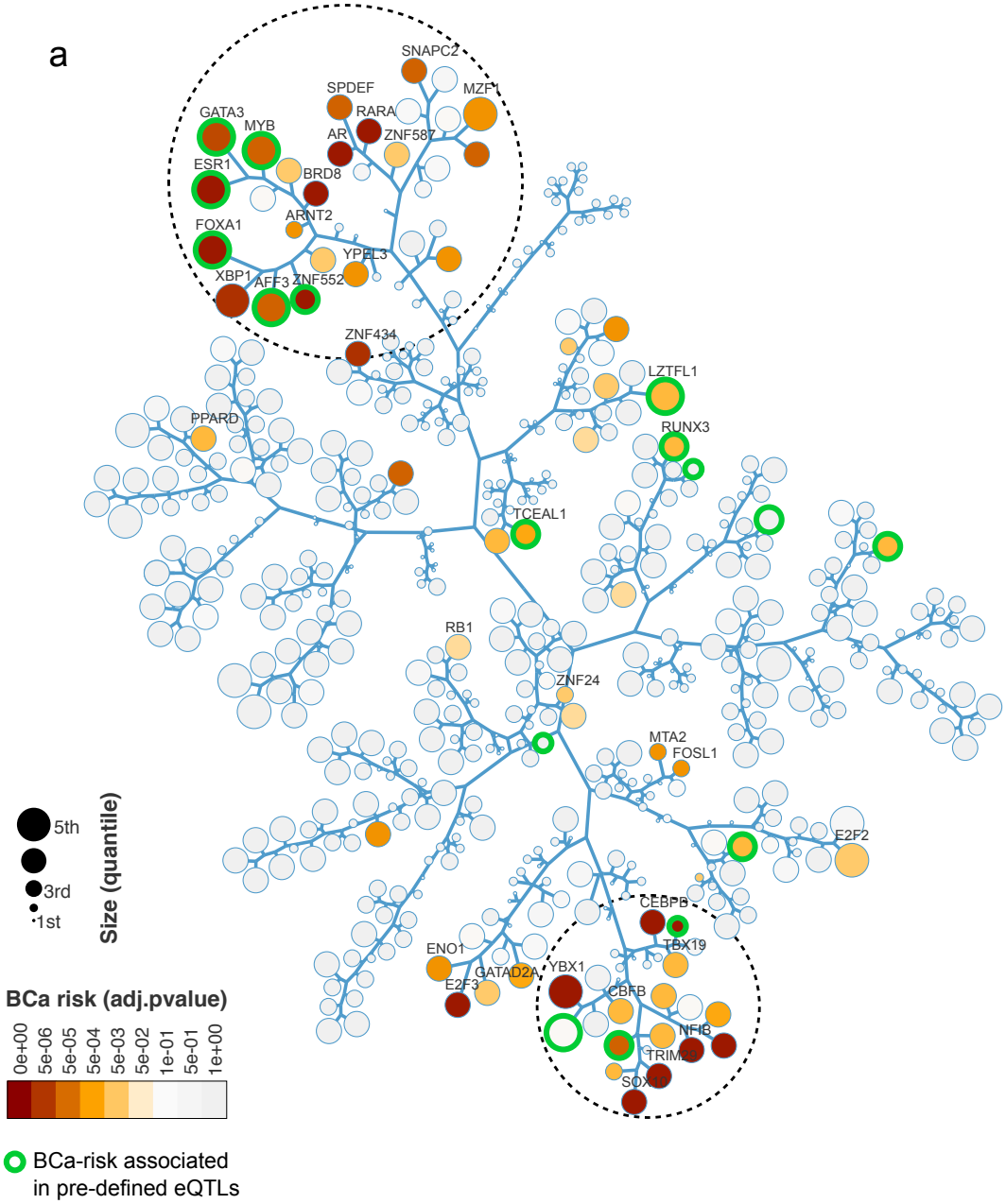
Supplementary Figure 15: **Comparison of the correlation of expression of TF-targets shared between TF pairs expressed in breast tumours shown in Figure 4e and the correlation obtained for the expression of the TFs themselves.** (**a**) Correlation of gene expression of the 36 risk-TFs in METABRIC cohort I. The scale for correlation is the same as that used in Figure 4e. (**b**) Histogram of Pearson's Correlation values obtained in a. (**c**) Histogram of correlation values obtained in the analysis shown in Figure 4e. The two distributions show that the correlation values obtained with TF-regulons were much higher than those obtained with TF genes only, with the majority of values being close to +1 or -1. In contrast the majority of correlation values obtained examining TF gene expression were around 0.3 to 0.4 or -0.3 to -0.4.

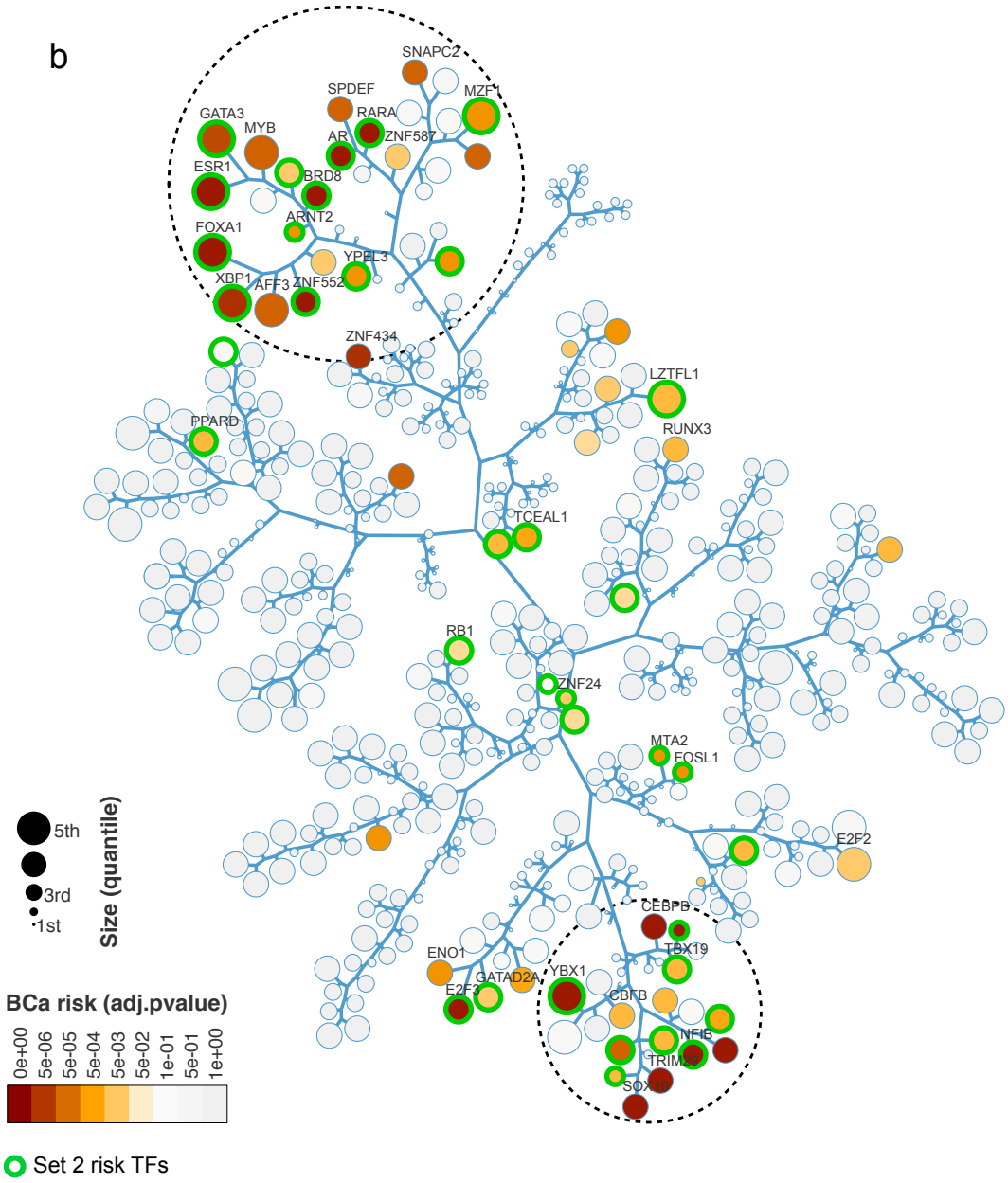


Supplementary Figure 16: **Relative gene expression of the 36 risk-TFs in ER$^+$ and ER$^-$ tumours and the 5 intrinsic (PAM50) subtypes of breast cancer in METABRIC.** (**a**) cohort I and (**b**) cohort II. Z-scores were obtained as described in the methods section. The TFs are ranked by their differential expression between ER$^+$ and ER$^-$ tumours.

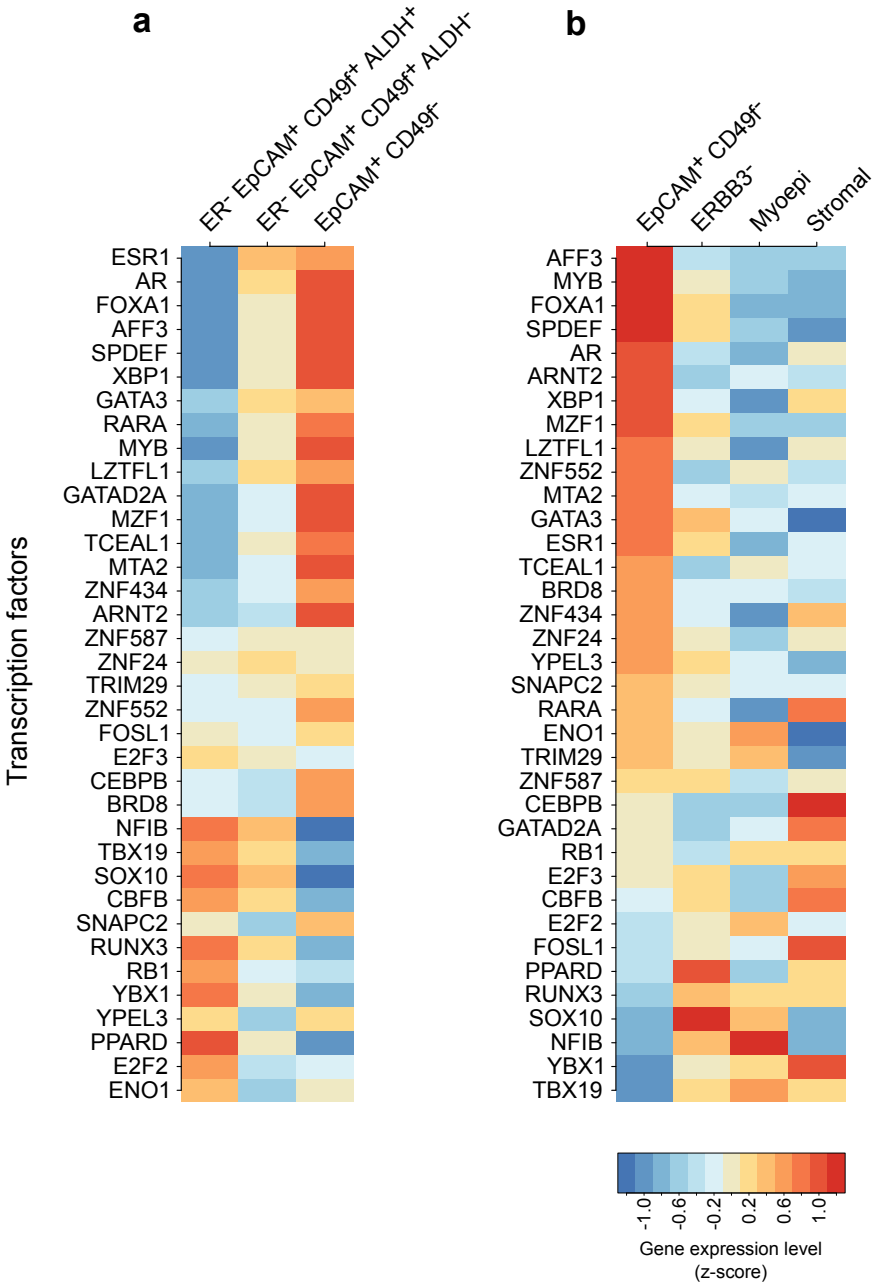


Supplementary Figure 17: **EVSE analysis of the 36 risk-TFs using ER$^+$ tumour samples from (a) cohort I and (b) cohort II of the METABRIC dataset.** The EVSE analysis was carried independently for each cohort using the breast cancer AVS. The panels list the enrichment score and associated p-value, next to the enrichment scores (boxplot). Solid and open red diamonds indicate significant enrichment scores that satisfy a Bonferroni-corrected threshold for significance of $P<0.05$ and $P<0.01$, respectively. P-values are based on null distributions from 1,000 random AVSs. All other annotations are as for Supplementary Figure 3.

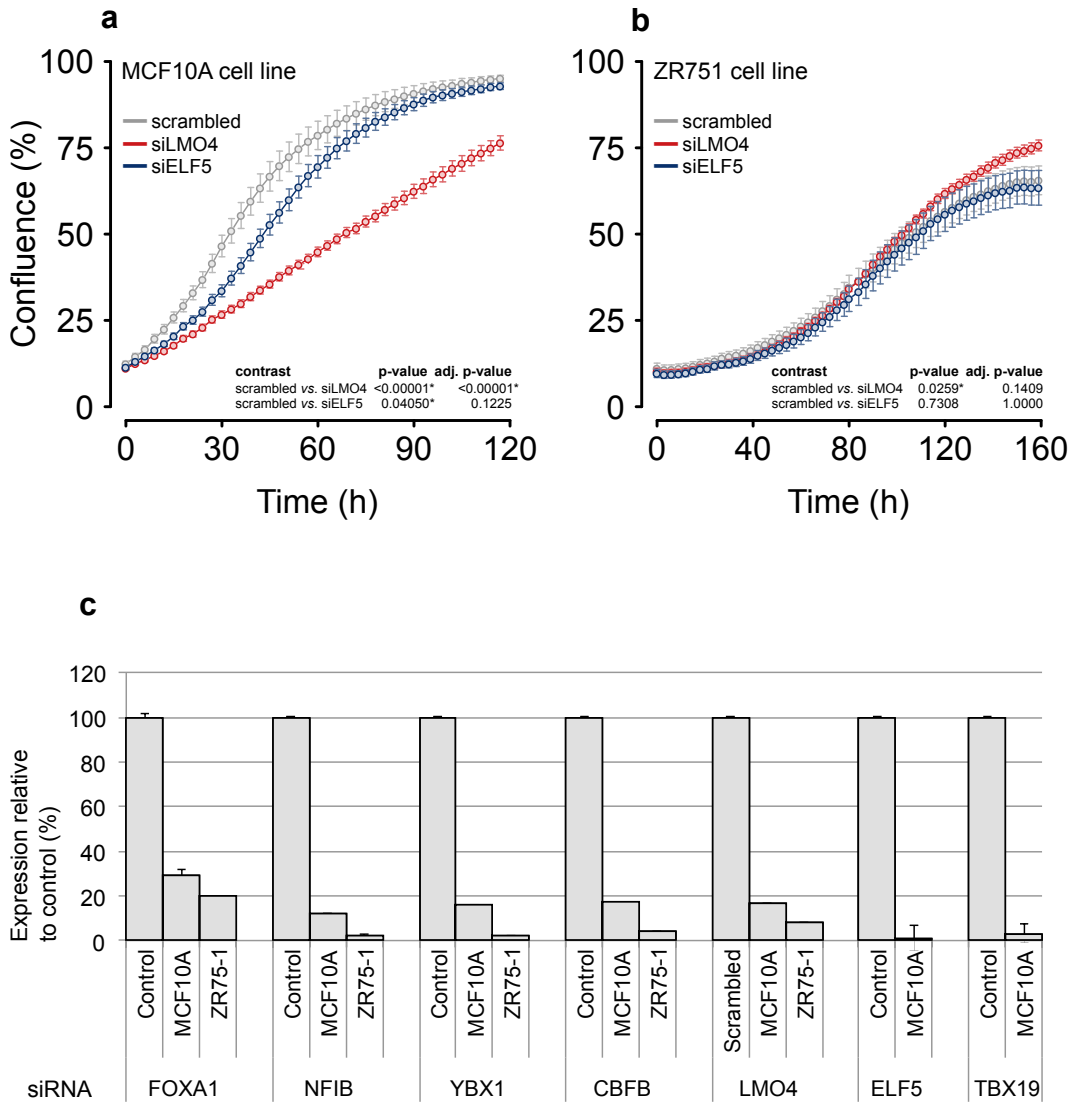

a
SNAPC2
SPDEF
MZF1
GATA3
MYB
RARA
AR
ZNF587
ESR1
BRD8
ARNT2
FOXA1
YPEL3
XBP1
AFF3
ZNF552
ZNF434
LZTFL1
PPARD
RUNX3
TCEAL1
RB1
ZNF24
MTA2
FOSL1
E2F2
CEBPB
TBX19
ENO1
GATAD2A
YBX1
E2F3
CBFB
NFIB
TRIM2
SOX10
5th
3rd
1st
Size (quantile)
BCa risk (adj.pvalue)
0e+00
5e-06
5e-05
5e-04
5e-03
5e-02
1e-01
5e-01
1e+00
BCa-risk associated
in pre-defined eQTLs



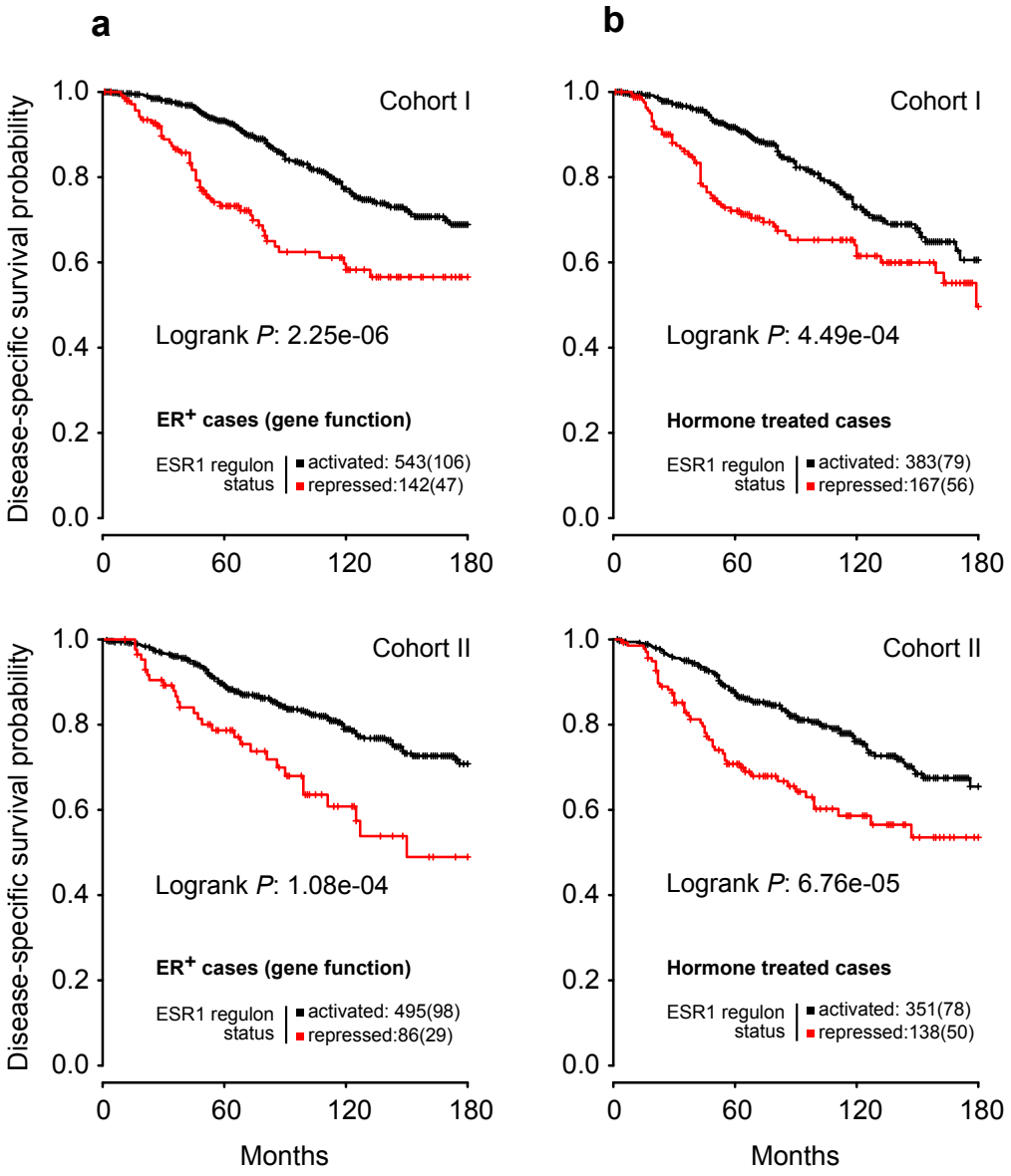
Supplementary Figure 18: **Tree and leaf diagram of all tested TFs as shown in Figure 5, highlighting different sets of risk-TFs.** (**a**) TFs that are associated with risk in the eQTL-conditioned analysis using predefined eQTLs in ER$^+$ tumours are highlighted by green circles. As for the EVSE analysis there is clear clustering of risk-TFs within cluster 1. The size of regulons is represented by circle size as indicated and p-values for enrichment of regulons for breast cancer GWAS loci in cohort I are shown in colour (same as depicted in **Fig. 5a**). (**b**) Tree and leaf diagram showing the position of a second set of risk-TFs (highlighted by green circles) that were defined in an EVSE analysis where regulons were inferred using three different algorithms (MRNETB, CLR, ARACNe with DPI threshold of 0.1). This Set 2 TFs comprises all TFs enriched in 4 out of the 6 analyses carried out (listed in Supplementary Table 4, also see page 8, **Supplementary Information**).



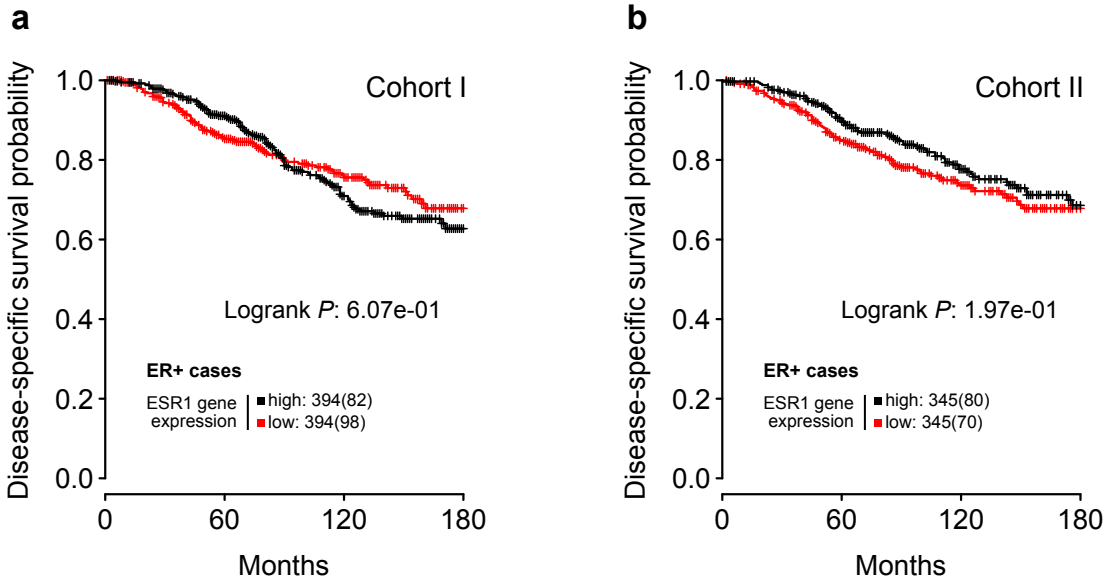
Supplementary Figure 19: **Relative gene expression levels of the 36 risk-TFs in the normal mammary cell populations isolated by Shehata *et al.*[18].** In contrast to **Figure 5f** results for all risk-TFs are shown. In (**a**) Z scores were calculated relative to the average of the 3 populations analysed (ER^- $EpCAM^+$ $CD49f^+$ $ALDH^+$, ER^- $EpCAM^+$ $CD49f^+$ $ALDH^-$ and $EpCAM^+$ $CD49f^-$) while in (**b**) Z scores were calculated relative to the average of the 4 populations analysed ($EpCAM^+$ $CD49f^-$, $ERBB3^-$, myoepithelial and stromal). TFs are ranked in (a) by the differential expression between $ALDH^+$ and $ALDH^-$ cell populations, while in (b) by the expression levels of $EpCAM^+$ $CD49f^-$ cells.



Supplementary Figure 20: **Effect of siRNA knock-down of TFs.** Growth curves for (**a**) ER$^-$ cell line MCF10A and (**b**) ER$^+$ cell line ZR751 after transient transfection of the siRNAs as indicated. LMO4 and ELF5 map to cluster 2 but are only significantly enriched in one of the cohorts and are therefore shown separately from Figure 6. The results show the averages of multiple transfections (MCF10A: n=3; ZR751: n=2; error bars show the SEM). The statistical analysis (insets) compares the growth curves using 100,000 simulations ($^*P<0.05$), with p-values adjusted by the BY correction method. (**c**) RT-PCR experiments demonstrate the reduction of TF gene expression 24 hours after siRNA transfection. Expression levels for each TF are given relative to a control transfection with scrambled siRNA, which is set as 100% for each cell line. The cell lines tested and the siRNAs assayed are listed along the x-axis. A representative experiment is shown and error bars represent the standard deviation for five technical repeats.



Supplementary Figure 21: **Survival analysis for ER$^+$ tumours defined by different evidences and stratified by ESR1 regulon status.** In the METABRIC study tumours are defined as ER$^+$ either using immunohistochemistry (**Fig. 7e-f**) or on the basis of gene expression data shown here. (**a**) Survival analysis of ER+ tumours defined by gene expression, stratified by ESR1 regulon status. (**b**) As additional evidence, similar results are obtained for tumours from those patients that underwent hormone therapy (**Figure 7** provides further detail on how regulon status was determined).



Supplementary Figure 22: **Survival analysis for ER$^+$ tumours stratified by ESR1 gene expression.** Kaplan-Meier survival curves for ER$^+$ tumours in cohort I (**a**) and II (**b**) of the METABRIC patients comparing those in which ESR1 gene expression is high to those with low ESR1 gene expression.

Supplementary Table 1: Mutations and copy number aberrations (CNA) in the 36 risk-TFs in TCGA.

Gene Symbol	Samples	Mutation	CNA	Altered	Frequency
GATA3	482	50	12	62	12.86
RARA	482	1	31	32	6.64
CEBPB	482	0	24	24	4.98
MYB	482	7	11	18	3.73
TBX19	482	1	17	18	3.73
FOXA1	482	8	8	16	3.32
CBFB	482	8	8	15	3.11
RB1	482	8	7	15	3.11
ZNF587	482	3	12	15	3.11
ZSCAN32	482	3	12	15	3.11
YPEL3	482	0	14	14	2.9
ZNF552	482	1	12	13	2.7
ESR1	482	2	10	12	2.49
MZF1	482	0	12	12	2.49
ARNT2	482	0	7	7	1.45
E2F3	482	1	6	7	1.45
AFF3	482	4	2	6	1.24
GATAD2A	482	0	6	6	1.24
NFIB	482	1	5	6	1.24
SNAPC2	482	0	6	6	1.24
YBX1	482	0	6	6	1.24
XBP1	482	2	3	5	1.04
BRD8	482	3	1	4	0.83
FOSL1	482	0	4	4	0.83
MTA2	482	1	3	4	0.83
SOX10	482	1	3	4	0.83
SPDEF	482	1	3	4	0.83
AR	482	3	0	3	0.62
ENO1	482	1	2	3	0.62
TRIM29	482	1	2	3	0.62
ZNF24	482	0	3	3	0.62
E2F2	482	0	1	1	0.21
LZTFL1	482	0	1	1	0.21
PPARD	482	0	1	1	0.21
RUNX3	482	0	1	1	0.21
TCEAL1	482	0	0	0	0

The median frequency of the observed aberration was compared to that observed in 10,000 random sets of genes, resulting in an empirical p-value of $p<10^{-4}$.

Supplementary Table 2: Consensus list of Master regulators (MR) of the E2 and FGFR2 response in three breast cancer cell lines, MCF7, T47D and ZR751.

E2 response				**FGF10 response**			
MCF-7 T47D	ZR751 T47D	ZR751 MCF7	consensus	MCF7 T47D	T47D ZR751	MCF7 ZR751	consensus
ASCL2	ASCL2	ASCL2	ASCL2	ASCL2	CREB5	CSDA	CSDA
CITED1	CITED1	CCRN4L	CITED1	CEBPB	CSDA	E2F3	E2F5
E2F2	E2F2	CITED1	E2F2	CSDA	E2F5	E2F5	ELF3
E2F3	E2F5	E2F2	E2F5	E2F2	ELF3	ELF3	ESR1
E2F5	ESR1	E2F5	ESR1	E2F5	ESR1	ESR1	FOXM1
ELF3	FOXM1	ENO1	FOXM1	ELF3	FOXA1	FOXM1	GATA3
ESR1	GATA3	ESR1	GATA3	ESR1	FOXM1	GATA3	HBP1
FOXA1	PGR	FOXM1	PGR	FOXM1	GATA3	HBP1	KLF6
FOXM1	PTTG1	GATA3	PTTG1	GATA3	HBP1	KLF6	PTTG1
GATA3	RUNX1	MYCN	RUNX1	GATAD2A	KLF6	NFIL3	SPDEF
HMGB2	SPDEF	PAX9	SPDEF	HBP1	PTTG1	PTTG1	XBP1
HOXD13	TFAP2B	PGR	TFAP2B	KLF10	SPDEF	SOX4	ZFP36L2
ILF2	ZNF395	PTTG1	ZNF395	KLF6	TRIM29	SOX9	
PGR	ZNF671	PURA	ZNF671	NFKB1	XBP1	SPDEF	
PTTG1		RUNX1		PBX1	ZFP36L2	TARDBP	
RUNX1		SPDEF		PGR	ZNF696	TFAP2B	
RXRA		TFAP2B		PTTG1		XBP1	
SIX5		ZNF395		PURA		YPEL3	
SOX13		ZNF484		RORC		ZFP36L2	
SPDEF		ZNF671		SMAD3		ZNF395	
TEAD4				SPDEF		ZNF45	
TFAP2B				TCF25			
YEATS4				TEAD4			
ZNF175				TRPS1			
ZNF264				XBP1			
ZNF395				YEATS4			
ZNF671				ZFP36L2			
				ZNF484			

MR of the Meta-PCNA signature in cohort I or II are shown in red writing; risk-TF in cohort I or II are indicated in green; TFs are ordered alphabetically.

Supplementary Table 3: Master regulator analysis using the basal gene signature derived by Bertucci et al.[17].

Cohort I	Adjusted p-value	Cohort II	Adjusted p-value	Consensus (Combined rank score)
SOX10	1.10E-24	SOX10	4.50E-19	SOX10
TRIM29	1.90E-08	TRIM29	5.80E-12	TRIM29
PLAGL1	2.40E-06	ATF3	7.50E-05	ETV5
ETV5	0.00055	CSDA	7.50E-05	PLAGL1
KLF6	0.00055	ETV5	0.00012	KLF6
LZTS1	0.00095	NFIB	0.00036	ATF3
SNAI2	0.00095	SOX9	0.00044	
TCF7L2	0.0013	PLAGL1	0.00047	
HEY2	0.0013	BCL6	0.00096	
CREB5	0.0013	KLF6	0.00096	
HOXD12	0.0013	HLF	0.0043	
OVOL1	0.0016	ZNF197	0.0044	
MEIS2	0.0016	ELF5	0.0062	
PRRX2	0.0050	ZFP36L2	0.0062	
EN1	0.0052			
ATF3	0.0095			

TF regulons enriched for the basal signature are listed. As significance cut-off a BH-adjusted p-value < 0.01 (MRA analysis) was used for cohort I and II. TFs found in both analyses are listed as consensus, ordered by summed ranks in cohort I and II. TFs in node 2 are highlighted in green.

Supplementary Table 4: EVSE analyses for all regulons derived from MRNETB, CLR and ARACNE algorithms. Results are shown for the previously identified 36 risk-TFs and those TFs where 4 out of 6 analyses were positive (intersect ≥4, which we refer to as Set 2 risk-TFs).

		Cohort I			Cohort II				
Regulon	Cluster	ARACNe	MRNETB	CLR	ARACNe	MRNETB	CLR	Intersect	Set 2
NFIB	2	1	1	1	1	0	1	5	Yes
YBX1	2	1	1	1	0	1	1	5	Yes
TBX19	2	0	1	0	1	1	1	4	Yes
SRF	2	1	1	1	1	1	1	6	Yes
ELF5	2	1	1	1	1	1	0	5	Yes
CREB3L2	2	1	1	1	0	1	0	4	Yes
KLF11	2	1	1	1	1	0	0	4	Yes
TCF7L1	2	0	1	0	1	1	1	4	Yes
ZNF552	1	1	1	1	1	1	1	6	Yes
XBP1	1	1	1	1	1	1	1	6	Yes
FOXA1	1	1	1	1	1	1	1	6	Yes
ARNT2	1	1	1	1	1	1	1	6	Yes
BRD8	1	1	1	1	0	1	1	5	Yes
ESR1	1	1	0	1	1	1	1	5	Yes
GATA3	1	1	0	1	1	1	1	5	Yes
YPEL3	1	1	1	1	0	0	1	4	Yes
MZF1	1	1	1	1	0	1	0	4	Yes
RARA	1	1	1	1	0	1	0	4	Yes
AR	1	1	1	1	0	1	0	4	Yes
TRPS1	1	1	1	1	1	1	1	6	Yes
VEZF1	1	1	1	1	1	1	0	5	Yes
FOSL1	0	1	1	1	1	1	1	6	Yes
MTA2	0	1	1	1	1	1	1	6	Yes
TCEAL1	0	1	1	1	1	1	1	6	Yes
ZNF24	0	1	1	1	1	1	1	6	Yes
GATAD2A	0	1	1	1	1	1	1	6	Yes
PPARD	0	1	1	1	1	1	0	5	Yes
E2F3	0	1	1	1	0	1	0	4	Yes
LZTFL1	0	0	1	1	0	1	1	4	Yes
RB1	0	0	0	1	1	1	1	4	Yes
HBP1	0	1	1	1	1	1	1	6	Yes
NFKB2	0	1	1	1	1	1	1	6	Yes
ELF4	0	1	1	1	1	0	1	5	Yes
AFF4	0	1	0	0	1	1	1	4	Yes
SMAD1	0	0	1	0	1	1	1	4	Yes
ZNF493	0	1	1	1	1	0	0	4	Yes
SOX10	2	0	1	1	0	0	0	2	No
CEBPB	2	0	1	1	0	0	0	2	No
TRIM29	2	0	1	0	0	0	0	1	No
CBFB	2	1	0	0	0	0	0	1	No
SPDEF	1	1	1	1	0	0	0	3	No
MYB	1	1	1	0	1	0	0	3	No
SNAPC2	1	1	1	1	0	0	0	3	No
AFF3	1	0	1	1	0	0	0	2	No
ZNF587	1	1	0	1	0	0	0	2	No
RUNX3	0	0	1	0	0	1	1	3	No
E2F2	0	1	0	0	1	0	1	3	No
ENO1	0	1	0	0	1	0	0	2	No
ZNF434	0	0	0	0	0	0	0	0	No

Significant results are indicated by "1" ($P < 0.05$, Bonferroni-corrected). Intersect represents the overlap between the 6 tests in the two cohorts.

Risk associated regulons mapped to clusters 1 or 2:			
1	in Cluster 1, Figure 5b	1	Set 2 risk TF, mapped to cluster 1
2	in Cluster 2, Figure 5c	2	Set 2 risk TF, mapped to cluster 2
0	not in Cluster 1 or 2, Figure 5a	0	Set 2, not in cluster 1 or 2

Supplementary Table 5: Oligonucleotide used in RT-PCR analysis of gene expression after siRNA transfections

oligonucleotide	sequence 5' to 3'
TBX19	AAGAATGGCAGACGGATGTTT TGGGGTGTGGAGGATAAGGAA
ELF5	CAAGACTGTCACAGTCATAG GTCAACCCGCTCCAAAATTC
LMO4	AAAGTGGCATGATCCTTTGC ACGAGTTCACTCGCAGGAAT
CBFB	ACTGCCAGCAGCTGTGAAAC TGATCTCCAAAGACTGGATGG
YBX1	AACTGGGAACACCACCAGGAC GGAGTTTGATGTTGTTGAAGGA
NFIB	TCTTGGGGAAGAATCCTGTG AAACCCAGCACTTTGTGTCC
TRIM29	TTGGGGCTTTGGCTCCGCATGA GGAGAAGCAAAAGGAGGAAGTG
SOX10	CGCTTGTCACTTTCGTTCAG GACCAGTACCCGCACCTG
FOXA1	GGGGGTTTGTCTGGCATAGC GCACTGGGGGAAAGGTTGTG
DGUOK	GCCTGAACTTCATGGTATTGG GCTGGTGTTGGATGTCAATG

Supplementary References

57 Patterson N, Price LA, Reich D. Population Structure and Eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

58 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904-909, 2006.

59 Purcell S, Chang C. PLINK version 1.9. *Software*, https://www.cog-genomics.org/plink2, 2015.

60 Meyer PE, Lafitte F, Bontempi G. MINET: An open source R/Bioconductor Package for Mutual Information based Network Inference. *BMC Bioinformatics*, 9:461, 2008.

61 Altay G, Emmert-Streib F. Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, 26(14):1738-1744, 2010.

62 Schafer J, Opgen-Rhein R, Zuber V, Ahdesmaki M, Silva P, Strimmer K. Corpcor: Efficient Estimation of Covariance and (Partial) Correlation. *R package*, 2015.

63 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56-65, 2012.

64 Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799-805, 2004.

65 Scutari M. Bnlearn: Bayesian network structure learning, parameter learning and inference. *R package*, 2015.

66 Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540-1542, 2006.