Vignette for *Fletcher2013b*: master regulators of FGFR2 signalling and breast cancer risk.

Mauro AA Castro[*], Michael NC Fletcher[*], Xin Wang, Ines de Santiago, Martin O'Reilly, Suet-Feung Chin, Oscar M Rueda, Carlos Caldas, Bruce AJ Ponder, Florian Markowetz and Kerstin B Meyer [†]

florian.markowetz@cancer.org.uk

kerstin.meyer@cancer.org.uk

April 18, 2015

Contents

[*]joint first authors

[†]Cancer Research UK - Cambridge Research Institute, Robinson Way Cambridge, CB2 0RE, UK.

1 Description

The package *Fletcher2013b* contains a set of transcriptions networks and related datasets that can be used to reproduce the results in Fletcher et al. [1]. The first part of this study is available in the package *Fletcher2013a*, which contains the time-course gene expression data and has been separated for better organization on the data distribution. Here we provide the R scripts to reproduce the bioinformatics analysis. Please refer to Fletcher et al. [1] for more details about the biological background and experimental design of the study.

2 Network inference and analysis

2.1 Data sources for regulatory network inference

The METABRIC breast cancer gene expression dataset [2] was used in two cohorts, a discovery set (n = 997) and a validation set (n = 995). The METABRIC normal breast expression dataset (n = 144) was used as a non-cancer, tissue control and a T-cell acute lymphoblastic leukaemia gene expression dataset (n = 57) was included as a non-related tissue, cancer control [3]. These data sets are publicly available at:

- METABRIC discovery set EGAD00010000210
- METABRIC validation set EGAD00010000211
- METABRIC normals EGAD00010000212
- T-cell ALL GSE33469

2.2 Reconstruction of the breast cancer transcription networks

Due to the large-scale datasets and the parallel processing required to compute the transcription networks, this package provides 4 pre-processed networks named: `rtni1st` (METABRIC discovery set), `rtni2nd` (METABRIC validation set), `rtniNormals` (METABRIC normals) and `rtniTALL` (T-cell ALL). These R objects will be required to reproduce the analyses along the vignette. Next we describe the main methods used to compute the transcription networks, and in the R package *RTN* we provide a short tutorial demostrating the inference pipeline.

2.2.1 Transcription network inference pipeline

In order to make all methods used in this study available for different users, we implemented the R package called *RTN: reconstruction of transcriptional networks and analysis of master regulators*, which is designed for the reconstruction of transcriptional networks using mutual information [4]. It is implemented by S4 classes in *R* [5] and extends several methods previously validated for assessing transcriptional regulatory units, or regulons (*e.g.* MRA [6], GSEA [7], synergy and shadow [8]). The main advantage of using *RTN* lies in the provision of a statistical pipeline that runs the network inference in a stepwise process together with a parallel computing algorithm that demands high performance. The *RTN* package should be installed prior to running this vignette. Additionally, in *RTN* we provide a tutorial showing how to compute a transcriptional network using a toy example, which is generated with default options and pValueCutoff=0.05. Here, the

pre-processed breast cancer transcription networks were generated by a more stringent threshold, with pValueCutoff=1e-6. To reproduce these large networks we suggest as minimum computational resources a cluster >= 8 nodes and RAM >= 8 GB per node (specific routines should be tuned for the available resources). The inference pipeline is executed in four steps: (***i***) check the consistency of the input data and remove non-informative probes, (***ii***) compute the mutual information and remove the non-significant associations by permutation analysis, (***iii***) remove unstable interactions by bootstrap and (***iv***) apply the data processing inequality filter. These steps are described next.

2.2.2 Pre-processing of gene expression data

Non-informative microarray probes with low dynamic range of expression were removed from the gene expression matrices. This procedure aims to filter out probes that exhibit low coefficient of variation (CV), below the CV median value. For breast cancer samples, this CV threshold yields a good overlap (>90%) with the corresponding differential expression analysis of cancer vs. normal cohort samples. The differential expression analysis therefore was used for quality control purposes. The advantage of using the CV here is that the same procedure could be applied across all samples, guaranteeing statistical independence between cancer and normal cohorts. In an alternative approach, for a given gene with multiple probes the *RTN* package selects the probe exibiting the maximum CV, which yields higher gene representativity. We have carried out both approaches and the overall results converged to the same scenario as described in [1].

2.2.3 Mutual information (MI) computation

The MI algorithm used in the *RTN* package extends the methods available in *minet* [9]. The structure of the regulatory network was derived by mapping all significant interactions between TF and target probes. The TF list was derived from that used in a previous ARACNe/MRA publication [6] by converting Affymetrix probe IDs into the equivalent probes on the Illumina Human-HT12 Expression BeadChip. Non-significant interactions were removed by permutation analysis. Unstable interactions were additionally removed by bootstrap analysis in order to create a consensus bootstrap network (referred to as the transcriptional network (TN)).

2.2.4 Application of data processing inequality (DPI)

DPI was applied to the RN with tolerance = 0.0 to remove interactions likely to be mediated by another TF [10]. As DPI removes the weakest edge of each network triplet, the vast majority of indirect interactions are likely to be removed. We also tested DPI tolerance ranging from 0.1 to 0.5 in order to assess the stability of the regulatory units identified in the transcriptional networks. Both the TN and the post-DPI network (filtered transcriptional network) were used in the MRA analysis.

2.3 Master Regulator Analysis (MRA)

The application of MRA has been described in detail in a previous publication [6]. MRA computes the overlap between two lists: the TFs and their candidate regulated genes (referred to as regulons) and the gene expression signatures from other sources. In this case, the MRA analytical pipeline estimates the statistical significance of the overlap between all the regulons in each TN using a hypergeometric test. The stability of MRA results was tested by comparing the MRA results

between the filtered and unfiltered TN networks, removing master regulators inconsistent with the previous analysis (*i.e.* selected regulons must be significant in both TN networks). Next we retrieve one of the FGFR2 signatures (*i.e.* differentially expressed genes from *Exp1*) and run the MRA analysis on METABRIC discovery set:

```
>   library(Fletcher2013b)
>   sigt <- Fletcher2013pipeline.deg(what="Exp1",idtype="entrez")
>   MRA1 <- Fletcher2013pipeline.mra1st(hits=sigt$E2FGF10, verbose=FALSE)

Running analysis pipeline ...

>   MRA1

      Regulon Universe.Size Regulon.Size Total.Hits Expected.Hits
2625    GATA3         19747          343        787         13.67
2099     ESR1         19747          367        787         14.63
25803   SPDEF         19747          187        787          7.45
      Observed.Hits  Pvalue Adjusted.Pvalue
2625             36 4.2e-08         2.2e-05
2099             37 8.1e-08         4.3e-05
25803            24 1.2e-07         6.2e-05
```

We provide the following functions to run the MRA analysis on the other 3 TN networks:

```
>   MRA2 <- Fletcher2013pipeline.mra2nd(hits=sigt$E2FGF10)
>   MRA3 <- Fletcher2013pipeline.mraNormals(hits=sigt$E2FGF10)
>   MRA4 <- Fletcher2013pipeline.mraTALL(hits=sigt$E2FGF10)
```

Each of these MRA pipelines constitutes a wrapper function that uses the pre-processed transcriptional networks together with the MRA algorithm implemented in the *RTN* package. Therefore, different signatures can also be interrogated on METABRIC datasets using these functions (for detailed description and default settings, please see the package's documentation).

3 Consensus breast cancer master regulators (MRs)

To define a smaller set of functionally important regulons, we applied the MRA functions described in the previous step to all transcriptional networks using all FGFR2 signatures (*i.e.* 2 TN networks vs. 3 FGFR2 signatures). We found that 20 regulons are reproducibly enriched across the two breast cancer cohorts in at least one experiment. This analysis is fully executed next:

```
>   masters <- Fletcher2013pipeline.masters()
>   Fletcher2013mra.consensus()
```

The overall agreement between the two cohorts was very high when a DPI tolerance of 0.05 is allowed, with regulons of five MRs enriched in both cohorts in all three experimental systems (DPI tolerance from 0.01 to 0.05 gives the same consensus). These were SPDEF, ESR1 and its co-factors FOXA1 and GATA3, and PTTG1 (Figure 1).

4 MRA aggrement among FGFR2 signatures and cohorts

The agreement among FGFR2 signatures was obtained by ranking regulons according to p-values derived from the MRA analyses, and then computing the Spearman's rank correlation coefficient for each pairwise ranking (Figure 2).

```
> Fletcher2013mra.agreement.cohort1()
```

Figure 2 shows that there is good agreement of the regulon rank when comparing the three different gene expression signatures (*Exp1-3*), both for the total set of regulons as well as the top 50 regulons, suggesting that our three model systems identify similar sets of dysregulated genes following FGFR2 signalling. The agreement among breast cancer cohorts are equally good (Figures 3, 4 and 5), and can be reproduced by the following functions:

```
> Fletcher2013mra.agreement.exp1()
> Fletcher2013mra.agreement.exp2()
> Fletcher2013mra.agreement.exp3()
```

5 Transcriptional network of consensus master regulators

Next, the pipeline function plots a graph representing all regulons identified in the consensus MRA analysis. The network is generated by the R package *RedeR* [11] and should require some user input in order to tune the layout in the software's interface (Figure 6).

```
> Fletcher2013pipeline.consensusnet()
```

As a suggestion, set 'anchor' to the master regulators at the end of the 'relax' algorithm for a better layout control! right-click the square nodes and then assign 'transform' and 'anchor'!!!

6 Clustering analysis

The non-supervised clustering analysis was performed on the adjacency matrix derived from the RN network. The Jaccard similarity coefficient (JC) was used as metric to compute the *manhattan* distance. For any two regulons, *R1* and *R2*, *JC* is simply obtained by dividing the number of common targets by the number all targets of the regulon pair, $JC = (R1 \cap R2)/(R1 \cup R2)$. The distance matrix was then used as input for the R function *hclust*, setting Ward's minimum variance method for agglomeration.

```
> Fletcher2013mra.heatmap1()
> Fletcher2013mra.heatmap2()
```

This code chunk reproduces two heatmaps, one showing all regulons clustered in the relevance network (Figure 7a), and the other focusing on the selected master regulators (Figure 7b).

7 Enrichment maps

In addition to the clustering analysis, the regulons were also represented in an association map showing the degree of similarity among them, the number of common targets. Likewise, the similarity is

assessed by the Jaccard coefficient, which is plotted in the association map by the R package *RedeR* [11]. In the next pipeline, a graph representation is generated for regulons exhibiting $JC \geq 0.4$ (Figure 8).

```
> Fletcher2013pipeline.enrichmap()
```

Suggestion: zoom in/out with a scroll wheel, and adjust the graph settings interactively!

8 GSEA analysis of master regulators

As a complementary approach, we assessed the enrichment of the master regulators using all information available in the FGFR2 signatures. In contrast to the MRA analysis that considers only the top differentially expressed genes, the GSEA uses the complete rank information. In the GSEA analysis [7], the association of a known set of genes is tested against the phenotypic difference. Here regulons are treated as *gene sets* and the FGFR2 perturbation experiments as *phenotypes*, an extension of the GSEA analysis as previously described [8]. Figure 9 shows the results computed in the next code chunk:

```
> Fletcher2013gsea.regulons(what="Exp1")
> Fletcher2013gsea.regulons(what="Exp2")
> Fletcher2013gsea.regulons(what="Exp3")
```

These functions evaluate the statistical significance of the gene set enrichment scores (ES) by performing 1000 permutations in the R package *RTN* (*a better statistical resolution as in [1] can be obtained using additional permuation steps*).

9 Synergy and shadow analyses

Regulon shadowing has been described as a potential confounding factor when assessing master regulators [8]. If two enriched regulons overlap significantly, one of them may appear enriched because of the common enriched targets. In order detect this potential confounding factor, we have applied for regulons a pairwise GSEA analysis restricted to non-common-targets, and the obtained ES score was then compared to the full regulon. This analysis was executed between all regulon pairs that exhibit a significant overlap. We have implemented the shadow analysis in the R package *RTN* following the method described in Lefebvre et al. [8]. Given two enriched regulons, *R1* and *R2*, the shadow analysis is run in 5 steps: (***i***) execute a hypergeometric test to assess the overlap between regulons; (***ii***) if the overlap is significant, compute the ES score for the full regulons; (***iii***) compute the ES score of the non-common-targets, $S1 = R1 \setminus (R1 \cap R2)$ and $S2 = R2 \setminus (R2 \cap R1)$; (***iv***) compute the ES scores for 1000 random subsets of the same size as *S1* and *S2*, taking the random samples from *R1* and *R2*, respectively; and (***v***) compute the empirical p-value of observing an ES smaller in *S1* than *R1*, and an ES smaller in *S2* than *R2*, having also observed the ES score signals. Therefore, each regulon pair is tested in the two directions, and a shadow is identified only in case the results are not symmetrical. As a natural extension of this approach, we implemented the synergy analysis in the same pipeline, which examines if the enrichment of the applied gene expression signature is greater in the intersect of two regulons, $RI = R1 \cap R2$, than the enrichment found in the union of two regulons, $RU = R1 \cup R2$. The empirical p-value is computed from 1000 random subsets of the same size as RI by taking random samples from RU.

```
> Fletcher2013pipeline.synergyShadow()
```

The pipeline *Fletcher2013pipeline.synergyShadow* is a wrapper for the functions available in *RTN* package, computing at once the synergy and shadow analyses for all master regulators (Figure 10)

10 Network validation

10.1 Motif analysis of regulons and binding sites

The position weight matrices (PWM) of known ESR1, FOXA1 and GATA3 motifs were collected from TRANSFAC database [12] and used as input for the FIMO DNA motif identification tool [13] to scan motif sites across the human genome (run with default parameters and p-value threshold 1e-4). Only PWMs inferred from MFC-7 cell line were considered in the analysis (the data collected from these webtools are available in *Fletcher2013b*). For each regulon we computed the distances from the transcription start sites to the nearest motif. The observed median distance was then compared to a random distribution derived from a random PWM using the Mann-Whitney test (Figure 11a).

```
> Fletcher2013pipeline.motifs()
```

10.2 ChIP-Seq analysis of regulons and binding sites

Next we examined the actual TF binding sites of SPDEF, ESR1, GATA3 and FOXA1 in MCF-7 cells (see experimental details in [1]). For each ChIP-Seq experiment and for each transcription start site annotated in the RefSeq collection, the distances from the transcription start sites to the nearest peak were determined (TSS-NP distance). Two complementary permutation analyses were performed to access whether the proximity of a given regulon to the master regulator binding sites is smaller than would be expected by chance. In the first one, a null distribution was obtained by computing the median TSS-NP distance of 1000 random regulons, and in the second approach a null distribution was computed by placing peaks at random locations on the genome (1000 times) and then determining the median TSS-NP distance. The empirical p-value was calculated as the probability of getting the observed median distance as near as, or smaller than, the random distributions. Rejecting *H0* in the former approach indicates that the observed regulon is different from random regulons, while in the later indicates that the distribution of the binding sites related to the observed regulon is different from random binding sites. This statistical analysis is fully executed in the next code chunk using SPDEF ChIP-seq data (Figure 11b):

```
> Fletcher2013pipeline.chipseq(what="SPDEF")
```

In order to run the same analysis for all vs. all regulons and datasets, please set the argument 'what' to a different ChIP-seq data (i.e. ESR1, GATA3 or FOXA1).

11 Analysis of siRNA data

As additional validation of the newly identified regulons, we carried out siRNA knock-down experiments using siRNA against SPDEF and PTTG1, and used previously published ESR1 data

(included as positive control) to confirm that the responsive gene sets are indeed enriched in the relevant regulons (the siRNA gene expression data is avaible in *Fletcher2013a*). For each of these three putative master regulators we found that its own regulon was significantly enriched. The MRA analysis that reproduces these results is provided in the following code chunk:

```
>   siESR1  <- Fletcher2013pipeline.siRNA(what="siESR1")
>   siSPDEF <- Fletcher2013pipeline.siRNA(what="siSPDEF")
>   siPTTG1 <- Fletcher2013pipeline.siRNA(what="siPTTG1")
```

12 Analysis of meta-PCNA signature

The meta-PCNA signature corresponds to a proliferation-based gene set (n=131 genes) inferred from the top 1% list of genes correlated with the PCNA gene expression across many tissues [14]. While PCNA is a well known proliferation marker, the meta-PCNA signature has been originally designed to address a potential confounding variable problem: most cancer signatures are related to bad clinical outcome because they are associated with proliferation. Additionally, meta-PCNA has been shown as a highly effective predictor of breast cancer outcome [14]. Here, the meta-PCNA was used to assess the hypothesis that the master regulators have been selected only for being related to proliferation; alternatively, a negative result would point to a more specific association with the FGFR2 signalling. Next, we run the same MRA analysis as described in the first sections of this vignette, but using meta-PCNA genes:

```
>   data(miscellaneous)
>   mPCNAmra <- Fletcher2013pipeline.mra1st(hits=metaPCNA, idtype="entrez",
+                                           pAdjustMethod="BH", ntop=-1)
```

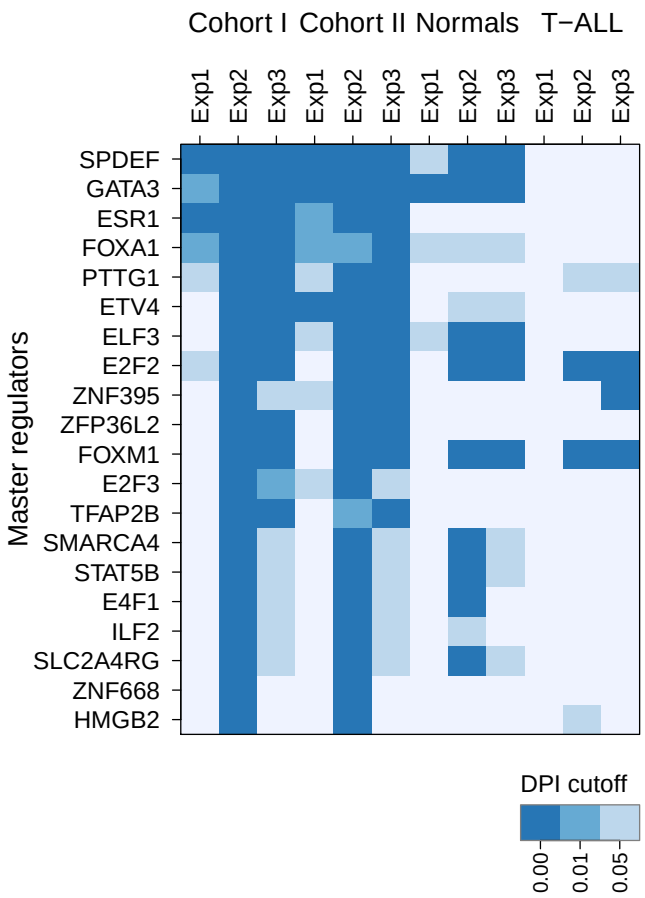

Figure 1: **Regulons enriched for FGFR2 signatures (Exp1-3) in breast cancer cohort I and cohort II.** There is substantial overlap between the MRs derived for different FGFR2 signatures, and the consensus corresponds to the 5 MRs for DPI threshold from 0.0 to 0.05.

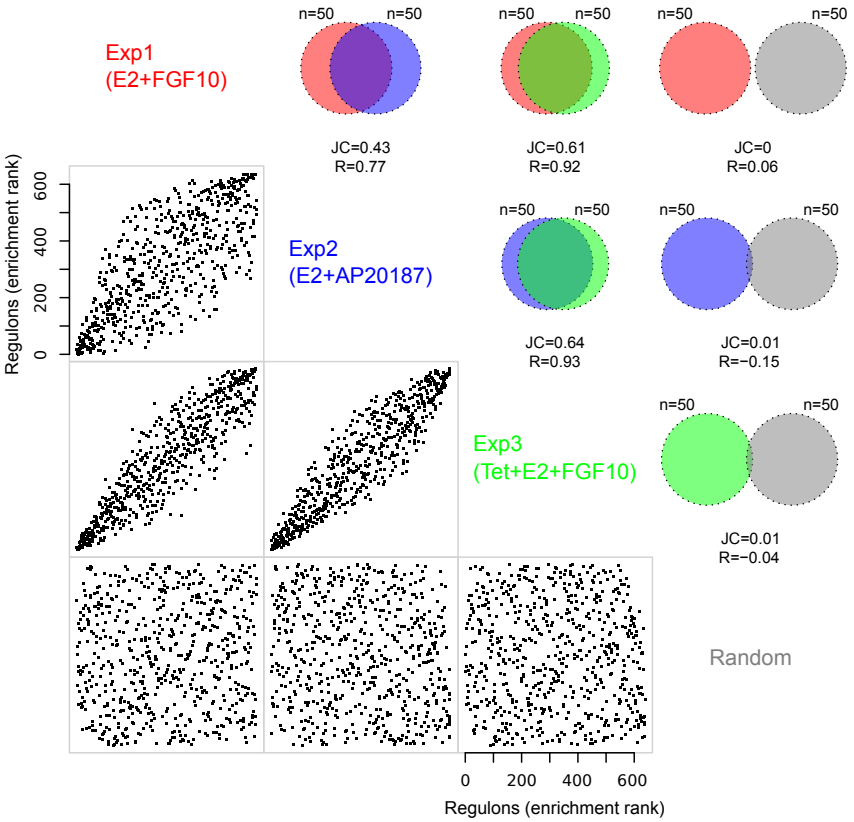

Figure 2: **MRA agreement among different FGFR perturbation experiments.** The scatter plots show the agreement in the ranking of all regulons by the enrichment p-value, between the different experimental perturbations of FGFR2 signalling: Exp1=E2+FGF10, Exp2=E2+AP20187 and Exp3=Tet+E2+FGF10. Each dot represents one regulon (i.e. one TF and all its targets) in the relevance network derived from cohort I. The correlation coefficient R is given for each pairwise ranking. The corresponding Venn diagrams show the level of agreement on the ranking for the top 50 enriched regulons, expressed by the Jaccard Coefficient (JC).

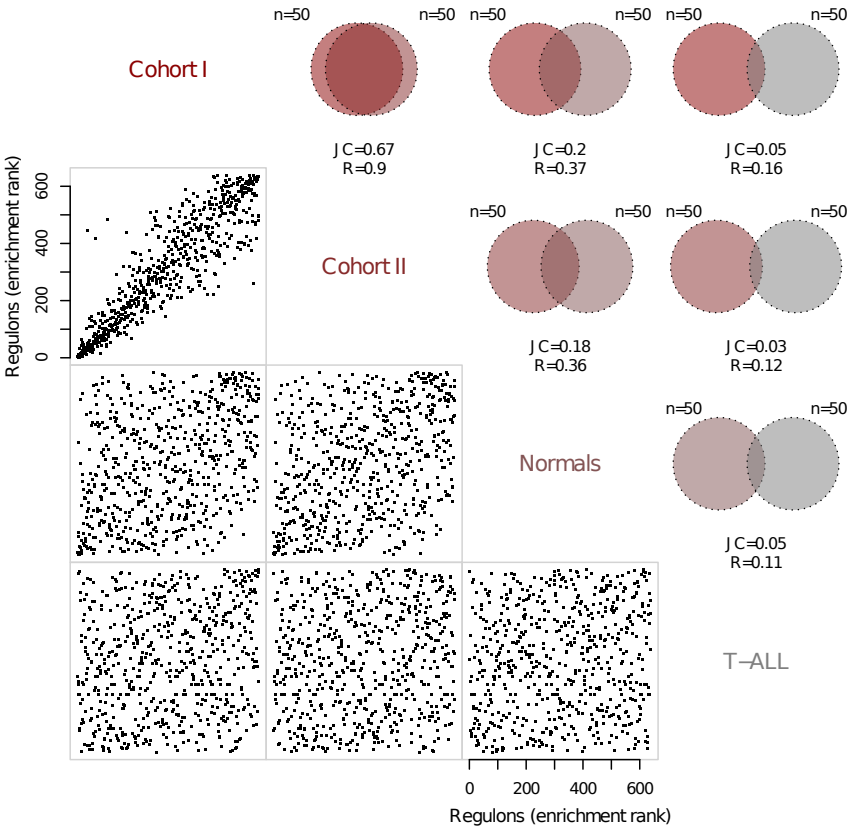

Figure 3: **MRA agreement among regulons derived from different relevance networks.** Regulons are ranked by the enrichment p-value estimated for E2.FGF10 signature (Exp1) and the graphs show the comparisons of regulon rank for cohort I, cohort II, normal breast tissue and T-ALL for all regulons. The correlation coefficient R is given for each comparison. The Venn diagrams depict the same comparison, but showing only the overlap obtained for the top 50 ranks, quantified by the Jaccard Coefficient (JC).

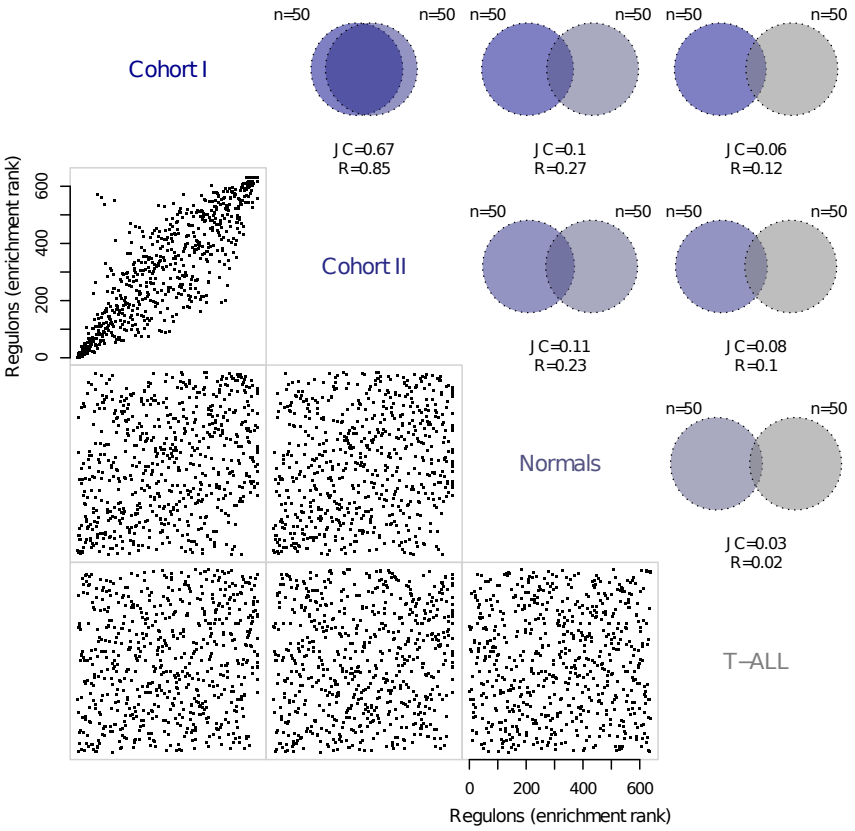

Figure 4: **MRA agreement among regulons derived from different relevance networks.** Regulons are ranked by the enrichment p-value estimated for E2.AP20187 signature (iF2 construct perturbation experiments) (Exp2) and shown as in Figure 3.

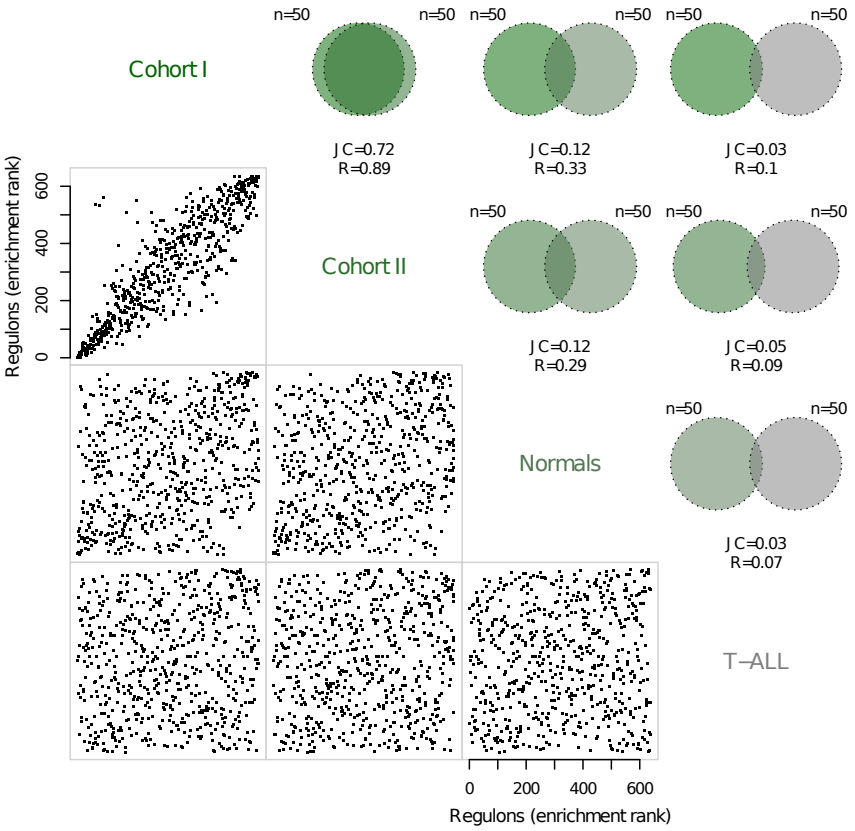

Figure 5: **MRA agreement among regulons derived from different relevance networks.** Regulons are ranked by the enrichment p-value estimated for PT.E2.FGF10 signature (FGFR2b perturbation experiments) (Exp3) and shown as in Supplementary Figure 4.

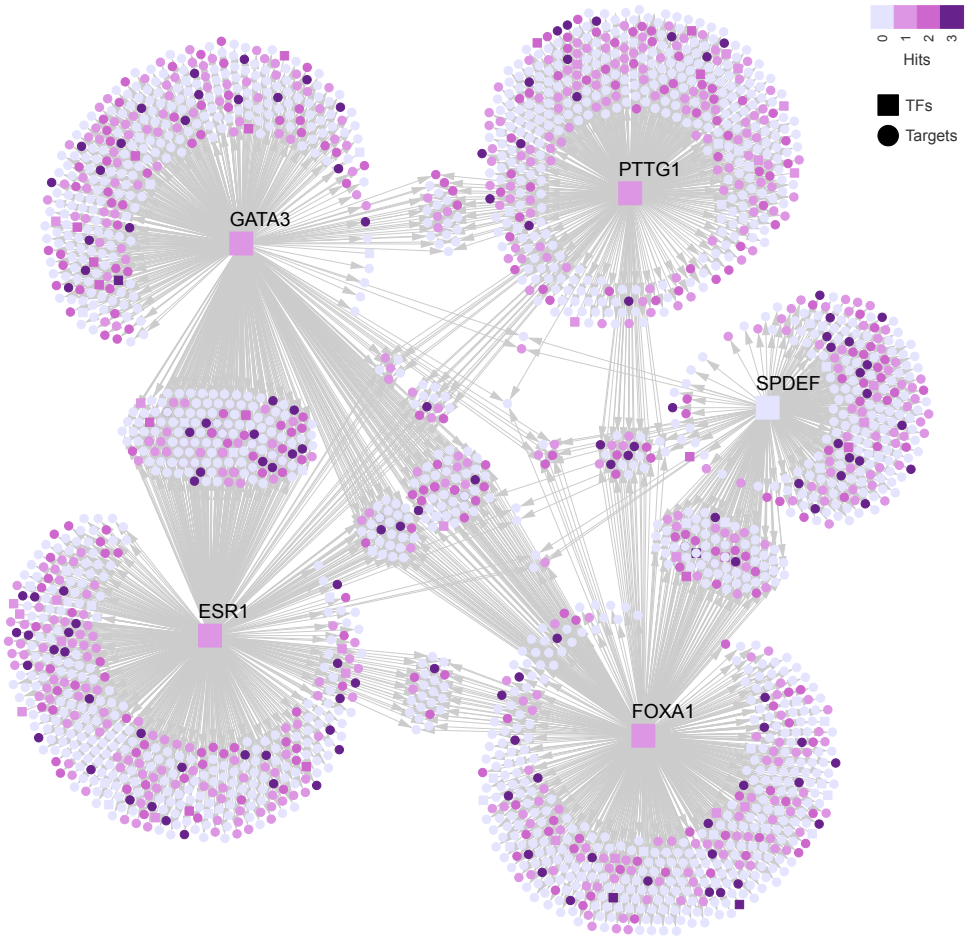

Figure 6: **Breast cancer transcriptional network (TN) enriched for the FGFR2 responsive genes.** The network shows the 5 MRs, each one comprising one TF (square nodes) and all inferred targets (round nodes) applying a DPI threshold of 0.01.

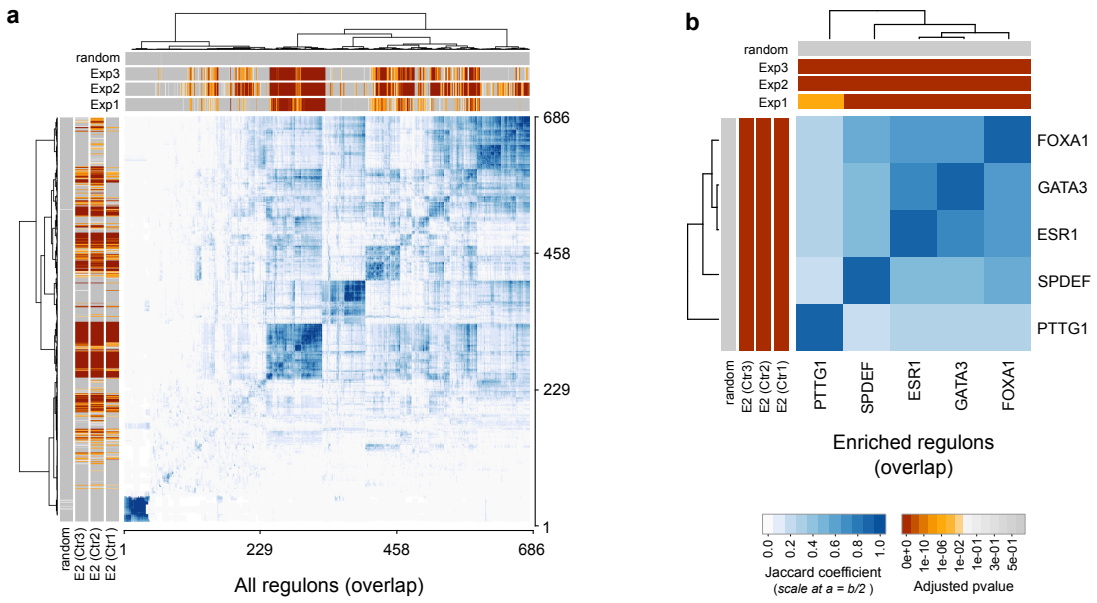

Figure 7: **Overlap between regulons in the relevance network.** (a) The heatmap shows the hierarchical clustering on the Jaccard similarity coefficient (in shades of blue) computed among all regulons in the relevance network derived from cohort I. Sidebars show the enrichment p-values (shades of orange) from the MRA analysis for FGFR2-associated gene expression signatures (Exp1-3) at the top of the graph and the MRA analysis for E2-associated gene expression signatures (E2 Ctrl1-3) derived for each experiment on the left. (b) Hierarchical clustering on the Jaccard similarity coefficient focused on the overlap between the 5MRs of the FGFR2 response.

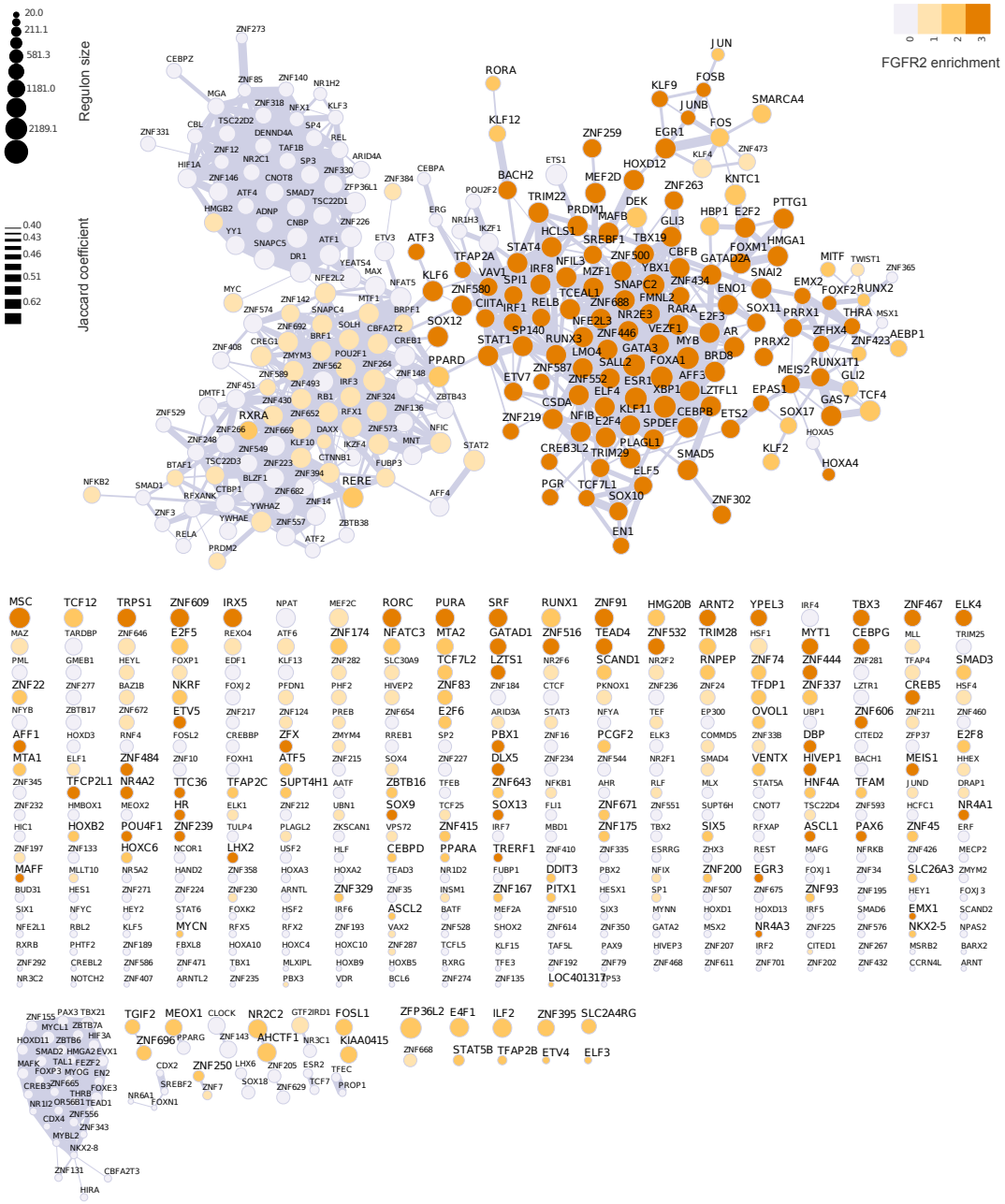

Figure 8: **Enrichment map derived from the relevance network in breast cancer.** Edge width depicts the overlap of regulons, and shades of orange indicate degree of enrichment of a regulon in at least one of the three FGFR2 gene signatures.

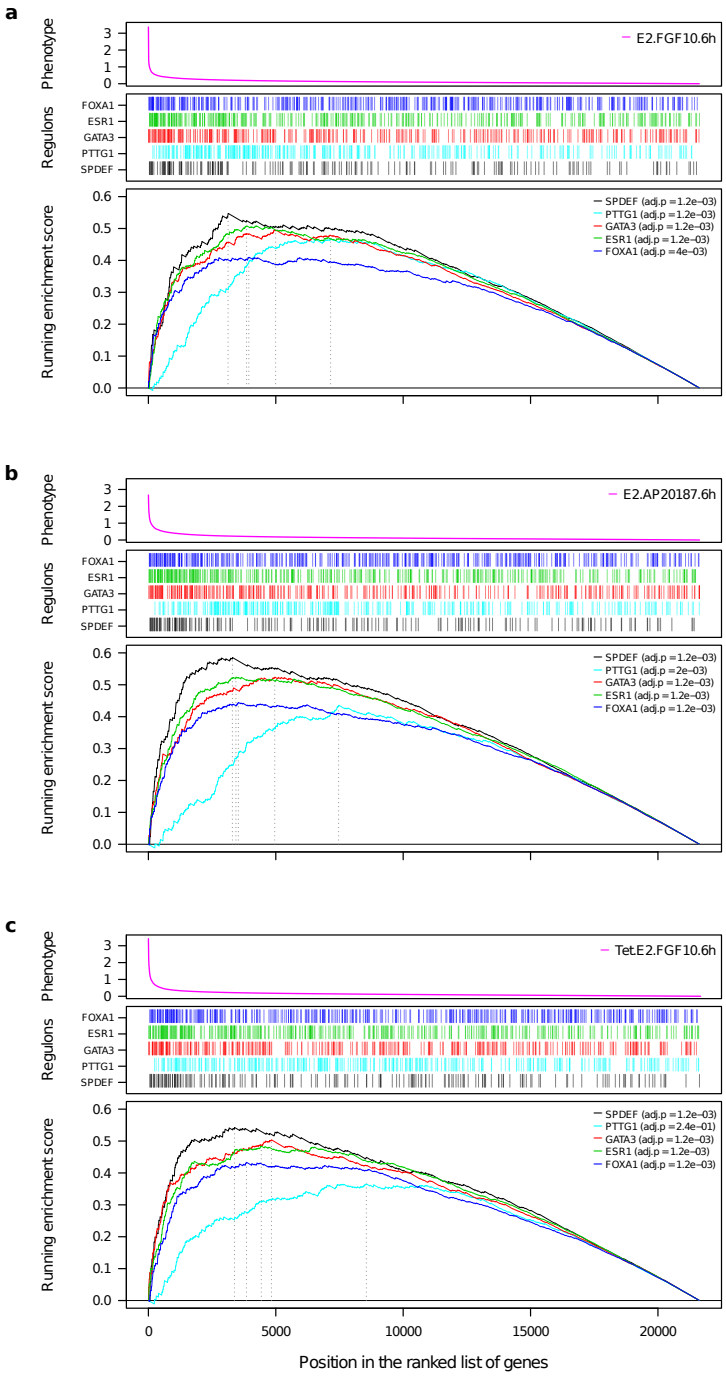

Figure 9: **GSEA of the genes in each of the 5 MR regulons.** Regulons are ranked by their response to FGFR2 signalling (phenotype) using the expression signatures Exp1 (a), Exp2 (b) and Exp3 (c).

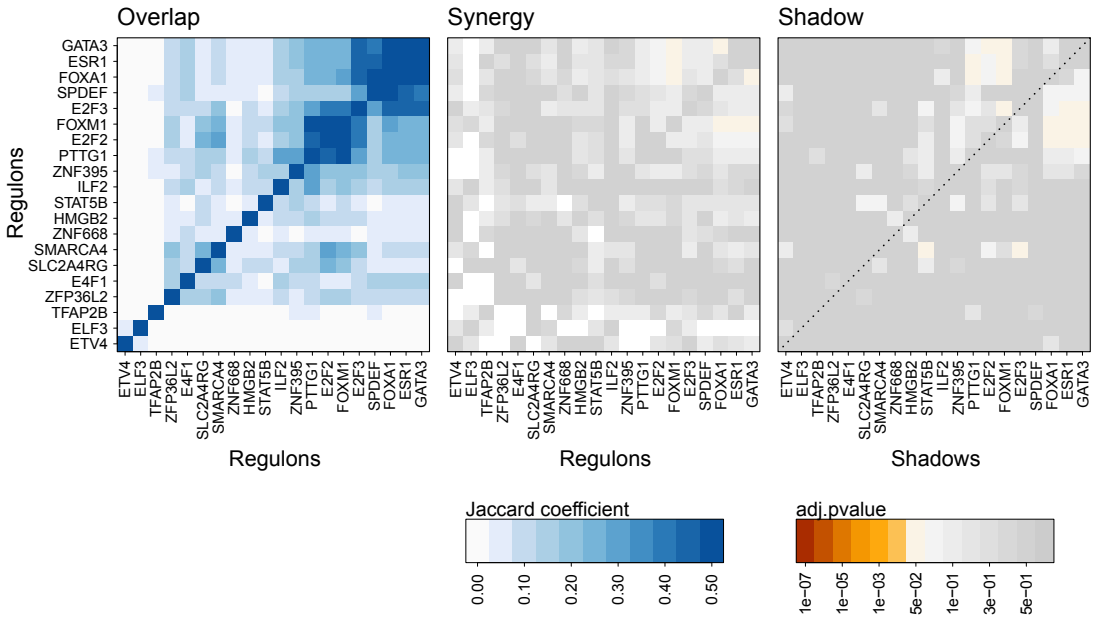

Figure 10: **Statistical analysis of the overlap of regulons computed for the relevance network (RN).** The overlap, synergy and shadowing are depicted (see Fletcher et al. [1] for more details). Shadowing can only be computed for those regulons whose overlap is significant.

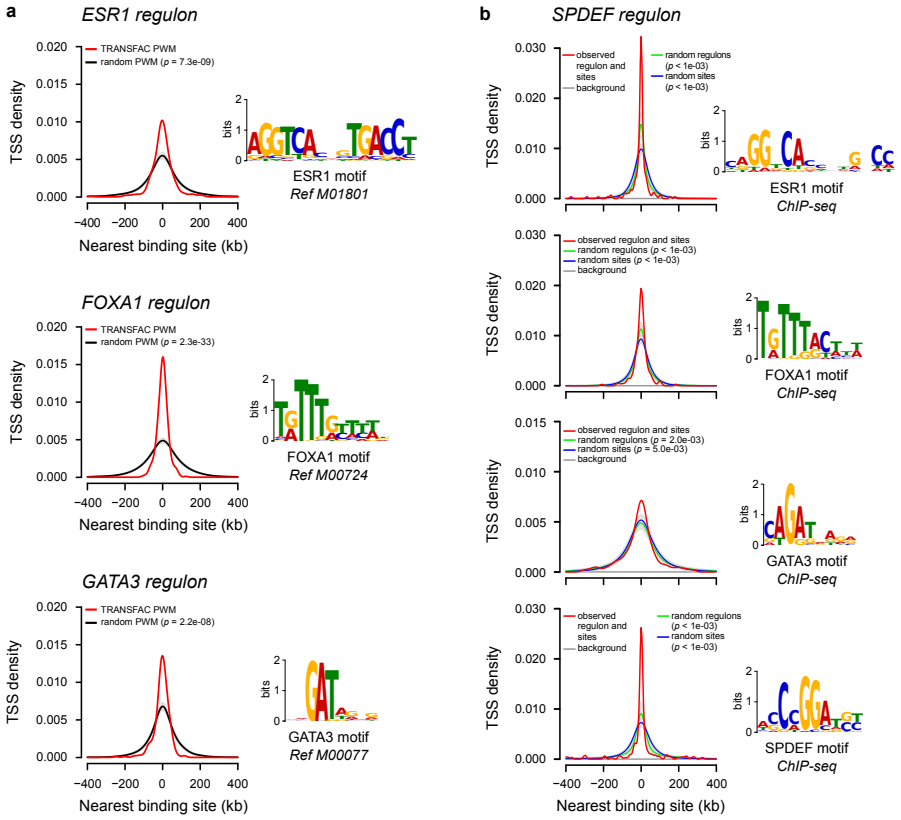

Figure 11: **Validation of regulons.** (a) Enrichment of known binding motifs for ESR1, FOXA1 and GATA3 in each of their inferred regulons. The occurrence of motif sites is shown as the distance between the TSS of the genes in each regulon and the nearest motif encountered (red line). This was compared to the occurrence of random sites of the same length in the same regulons derived for a random motif (black line). Motifs are taken from Transfac. (b) Enrichment of binding sites of the ESR1, FOXA1, GATA3 and SPDEF regulons in SPDEF ChIP-seq data obtained in MCF-7 cells. A background distribution is shown as a reference line (grey line) and represents the distance between the TSS and a random peak placed in the same chromosome.

13 Session information

```
R version 3.2.0 (2015-04-16)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: Ubuntu 14.04.2 LTS

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets
[7] methods   base

other attached packages:
[1] Fletcher2013b_1.4.0 RedeR_1.16.0        RTN_1.6.0
[4] igraph_0.7.1        VennDiagram_1.6.9   biomaRt_2.24.0
[7] Fletcher2013a_1.4.0 limma_3.24.0

loaded via a namespace (and not attached):
 [1] gtools_3.4.2         reshape2_1.4.1       splines_3.2.0
 [4] lattice_0.20-31      colorspace_1.2-6     snow_0.3-13
 [7] stats4_3.2.0         mgcv_1.8-6           chron_2.3-45
[10] XML_3.98-1.1         corrgram_1.7         nloptr_1.0.4
[13] DBI_0.3.1            BiocGenerics_0.14.0  RColorBrewer_1.1-2
[16] foreach_1.4.2        plyr_1.8.1           stringr_0.6.2
[19] caTools_1.17.1       codetools_0.2-11     Biobase_2.28.0
[22] ff_2.2-13            IRanges_2.2.0        SparseM_1.6
[25] seriation_1.0-14     GenomeInfoDb_1.4.0   quantreg_5.11
[28] pbkrtest_0.4-2       parallel_3.2.0       AnnotationDbi_1.30.0
[31] Rcpp_0.11.5          KernSmooth_2.23-14   gdata_2.13.3
[34] S4Vectors_0.6.0      bit_1.1-12           lme4_1.1-7
[37] gplots_2.16.0        gclus_1.3.1          tools_3.2.0
[40] bitops_1.0-6         RCurl_1.95-4.5       RSQLite_1.0.0
[43] cluster_2.0.1        car_2.0-25           MASS_7.3-40
[46] Matrix_1.2-0         data.table_1.9.4     minet_3.26.0
[49] minqa_1.2.4          iterators_1.0.7      TSP_1.1-0
[52] nnet_7.3-9           nlme_3.1-120
```

References

[1] Michael NC Fletcher, Mauro AA Castro, Suet-Feung Chin, Oscar Rueda, Xin Wang, Carlos Caldas, Bruce AJ Ponder, Florian Markowetz, and Kerstin B Meyer. Master regulators of FGFR2 signalling and breast cancer risk. *Nature Communications*, 4:2464, 2013.

[2] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, and et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346–352, 2012.

[3] Pieter Van Vlierberghe, Alberto Ambesi-Impiombato, Arianne Perez-Garcia, J. Erika Haydu, Isaura Rigo, Michael Hadler, Valeria Tosello, Giusy Della Gatta, Elisabeth Paietta, Janis Racevskis, Peter H. Wiernik, Selina M. Luger, Jacob M. Rowe, Montserrat Rue, and Adolfo A. Ferrando. Etv6 mutations in early immature human t cell leukemias. *The Journal of Experimental Medicine*, 208(13):2571–2579, 2011.

[4] Adam Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Favera, and Andrea Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[6] Maria Stella Carro, Wei Keat Lim, Mariano Javier Alvarez, Robert J. Bollo, Xudong Zhao, Evan Y. Snyder, Erik P. Sulman, Sandrine L. Anne, Fiona Doetsch, Howard Colman, Anna Lasorella, Ken Aldape, Andrea Califano, and Antonio Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, 01 2010.

[7] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[8] Celine Lefebvre, Presha Rajbhandari, Mariano J Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, Pavel Sumazin, Manjunath Kustagi, Brygida C Bisikirska, Katia Basso, Pedro Beltrao, Nevan Krogan, Jean Gautier, Riccardo Dalla-Favera, and Andrea Califano. A human b-cell interactome identifies myb and foxm1 as master regulators of proliferation in germinal centers. *Mol Syst Biol*, 6, 06 2010.

[9] Patrick Meyer, Frederic Lafitte, and Gianluca Bontempi. minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9(1):461, 2008.

[10] Adam A Margolin, Kai Wang, Wei Keat Lim, Manjunath Kustagi, Ilya Nemenman, and Andrea Califano. Reverse engineering cellular networks. *Nat. Protocols*, 1(2):662–671, 07 2006.

[11] Mauro AA Castro, Xin Wang, Michael NC Fletcher, Kerstin B Meyer, and Florian Markowetz. Reder: R/bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. *Genome Biology*, 13(4):R29, 2012.

[12] V. Matys, E. Fricke, R. Geffers, E. GÃűÃ§ling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. MÃijnch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.

[13] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[14] David Venet, Jacques E. Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol*, 7(10):e1002240, 10 2011.