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

October 19, 2016

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:

```
> library(Fletcher2013b)
> data(rtni1st)
> data(rtni2nd)
> data(rtniNormals)
> data(rtniTALL)
```

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:

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

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 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 1).

```
> 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'!!!

4 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 2).

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

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

5 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 3 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*).

6 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 4)

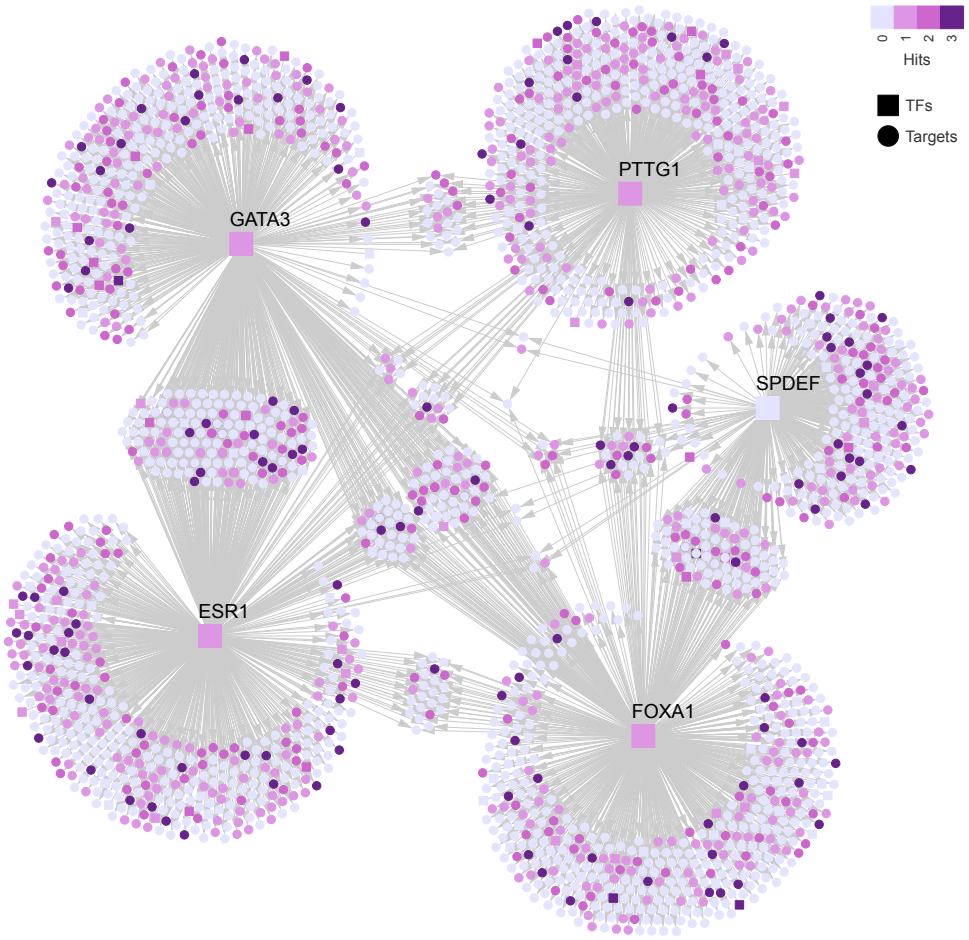

Figure 1: **Breast cancer transcriptional network (TN) enriched for the FGFR2 responsive genes.** The network shows the 5 MRs, each one comprising one TF (square nodes) and all inferred targets (round nodes) applying a DPI threshold of 0.01.

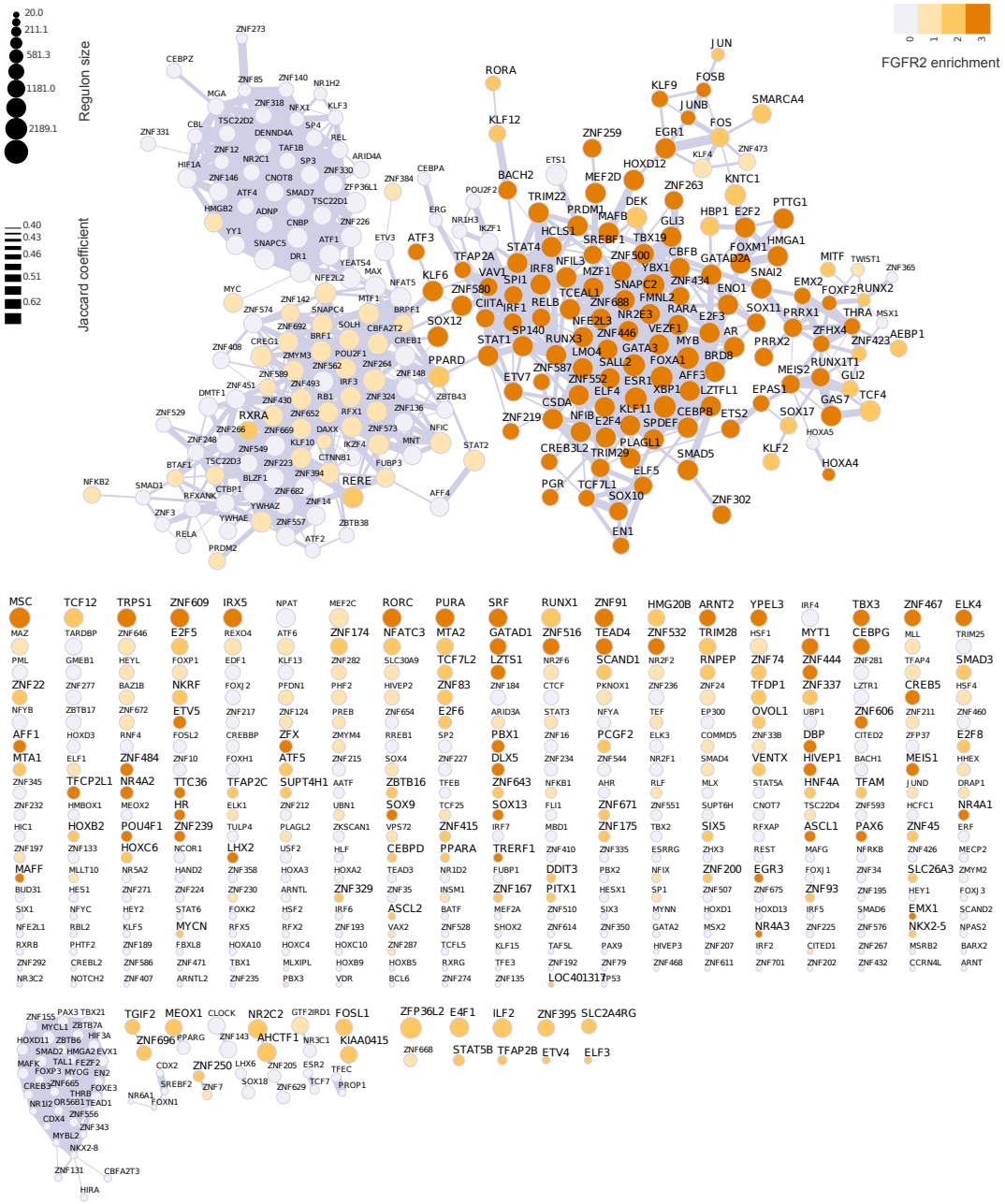

Figure 2: **Enrichment map derived from the relevance network in breast cancer.** Edge width depicts the overlap of regulons, and shades of orange indicate degree of enrichment of a regulon in at least one of the three FGFR2 gene signatures.

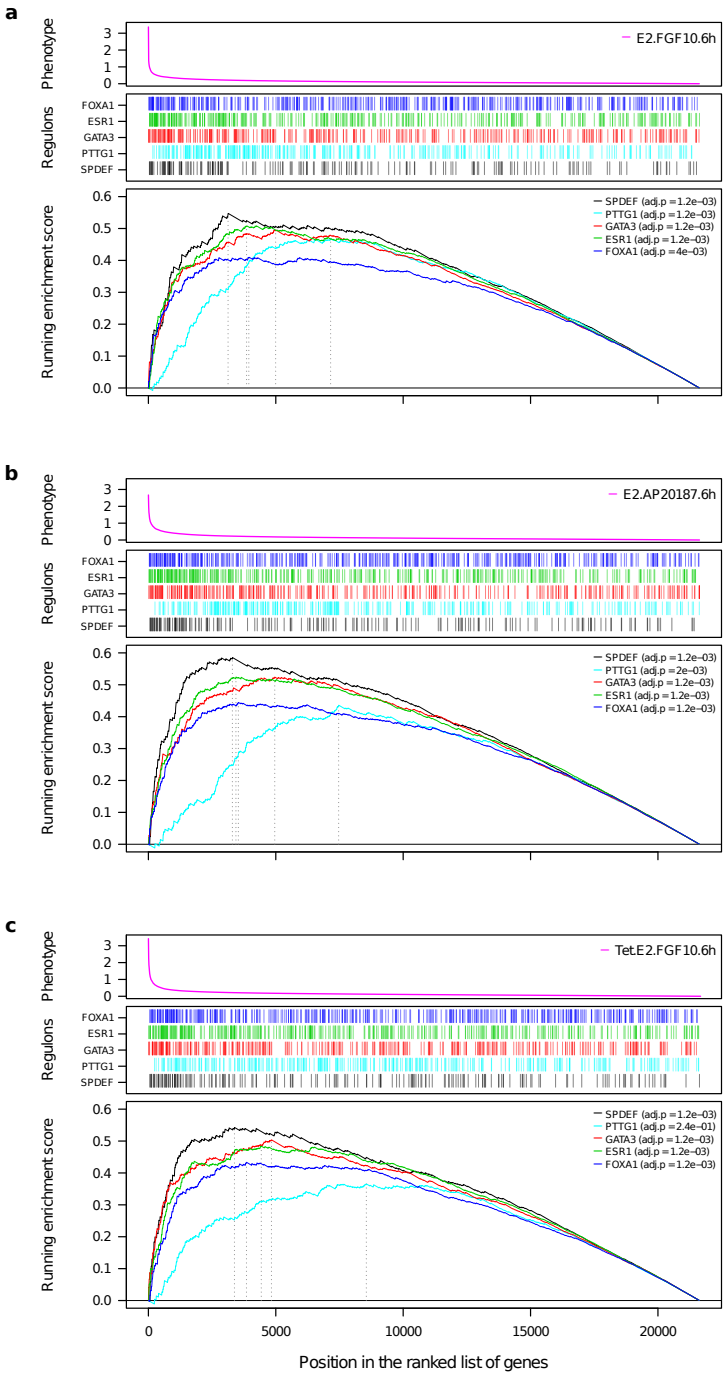

Figure 3: **GSEA of the genes in each of the 5 MR regulons.** Regulons are ranked by their response to FGFR2 signalling (phenotype) using the expression signatures Exp1 (a), Exp2 (b) and Exp3 (c).

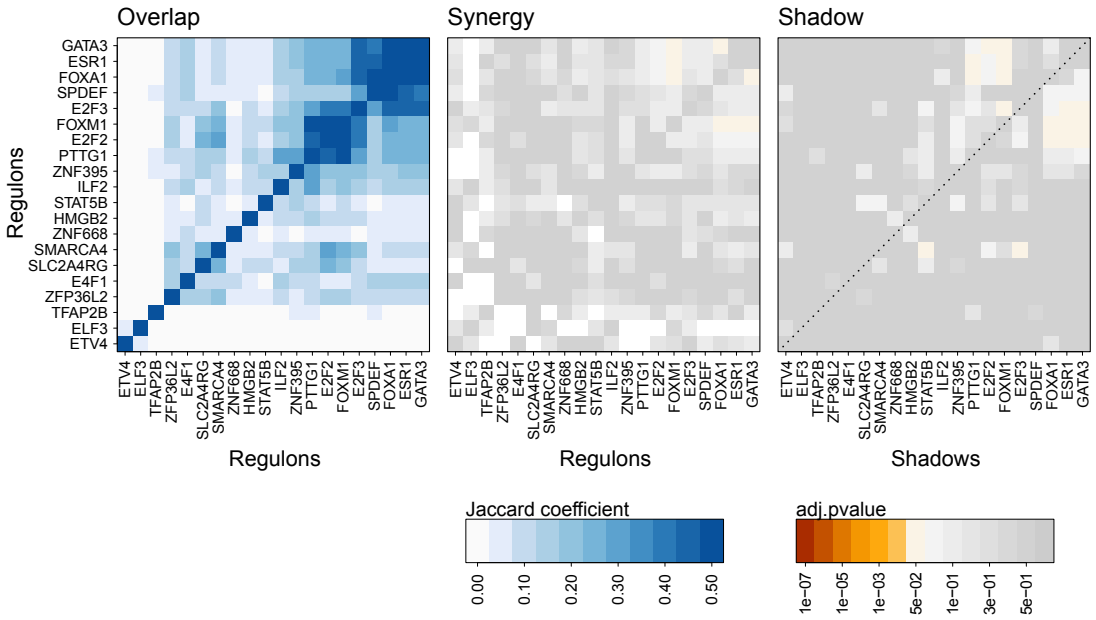

Figure 4: **Statistical analysis of the overlap of regulons computed for the relevance network (RN).** The overlap, synergy and shadowing are depicted (see Fletcher et al. [1] for more details). Shadowing can only be computed for those regulons whose overlap is significant.

7 Session information

```
R version 3.3.1 (2016-06-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.1 LTS

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

loaded via a namespace (and not attached):
[1] tools_3.3.1
```

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.