# Using ssmarina, a package for single-sample Master Regulator Analysis

Mariano J. Alvarez and Andrea Califano
Department of Systems Biology, Columbia University, New York, USA

September 27, 2013

## 1 Overview of MARINa

Phenotypic changes effected by pathophysiological events are now routinely captured by gene expression profile (GEP) measurements, determining mRNA abundance on a genome-wide scale in a cellular population[8, 9]. In contrast, methods to measure protein abundance on a proteome-wide scale using arrays[11] or mass spectrometry[10] technologies are far less developed, covering only a fraction of proteins, requiring large amounts of tissue, and failing to directly capture protein activity. Furthermore, mRNA expression does not constitute a reliable predictor of protein activity, as it fails to capture a variety of post-transcriptional and post-translational events that are involved in its modulation. Even reliable measurements of protein abundance, for instance by low-throughput antibody based methods or by higher-throughput methods such as mass spectrometry, do not necessarily provide quantitative assessment of functional activity. For instance, enzymatic activity of signal transduction proteins, such as kinases, ubiquitin ligases, and acetyltransferases, is frequently modulated by post-translational modification events that do not affect total protein abundance. Similarly, transcription factors may require post-translationally mediated activation, nuclear translocation, and co-factor availability before they may regulate specific repertoires of their transcriptional targets. Finally, most target-specific drugs affect the activity of their protein substrates rather than their protein or mRNA transcript abundance.

The ssMARINa (single-sample MAster Regulator Inference algorithm) allows computational inference of transcription factor protein activity, on an individual sample basis, from gene expression profile data. It uses the expression of genes that are most directly regulated by a given protein, such as the targets of a transcription factor (TF), as an accurate reporter of its activity.

We have shown that analysis of TF targets inferred by the ARACNe algorithm[1, 13], using the Master Regulator Inference algorithm (MARINa)[4], is effective in identifying drivers of specific cellular phenotypes. Several MARINa-inferred Master Regulators (MRs) were experimentally validated and shown to have significant differential protein activity, without significant changes in their mRNA expression[4, 6]. We have recently enhanced the MARINa algorithm by incorporating the mode of regulation (MoR, including activation, repression or non-determined), regulator-target gene interaction confidence, and target pleiotropic regulation[12]. We have found that enrichment of both activated and repressed targets of a TF (i.e., the TF regulon) in genes that are differentially expressed in a specific sample (e.g., a population of drug perturbed cells) is an accurate predictor of its differential activity at the protein level, even when the TF gene expression is unaffected[12].

The ssMARINa algorithm is the extension of this approach to rank relative TF protein activity on a sample-by-sample basis, thus allowing transformation of a typical gene expression matrix (i.e. multiple mRNA profiled across multiple samples) into a TF protein activity matrix, representing the relative activity of each TF in each sample.

The *ssmarina* package implements MARINa and ssMARINa algorithms in R. The *bcellExample* data package provides some example datasets and a small B-cell context-specific transcriptional regulatory net-

Table 1: Regulatory networks available from figshare.

| Title | Figshare citation |
|---|---|
| Human B-cell transcriptional network | http://dx.doi.org/10.6084/m9.figshare.680885 |
| Human B-cell transcriptional network | http://dx.doi.org/10.6084/m9.figshare.680888 |
| Human glioma transcriptional network | http://dx.doi.org/10.6084/m9.figshare.680887 |
| MCF7 human breast carcinoma cell line transcriptional network | http://dx.doi.org/10.6084/m9.figshare.680889 |

work, representing 172,240 inferred regulatory interactions between 621 TFs and 6,249 target genes. Additional networks can be obtained from figshare (Table 1) and from the author's web site (`http://wiki.c2b2.columbia.edu/califanolab/index.php/Software`).

# 2 Citation

Alvarez, M. J., Shen, Y, Ding, B. B., Ye, B. H. & Califano, A. Inferring global protein activity profiles by network based analysis of gene expression signatures. (Manuscript in preparation).

# 3 Generating the *regulon* object

As described under 'Overview of MARINa' (section 1), MARINa and ssMARINa require a gene expression signature and an appropriate cell context-specific regulatory network. This regulatory network is provided in the format of a class *regulon* object. Regulon objects can be generated from networks reverse engineered with the ARACNe algorithm [1]. This is performed by the function *aracne2regulon*, which takes two arguments as input: the ARACNe output *.adj* file, and the expression data-set used by ARACNe to reverse engineer the network. As an example, the package *bcellExample* provides a subset of the ARACNe output file containing the network for 20 TF regulators (*bcellaracne.adj* file). For convenience, the full network is also provided, as a *regulon* class object, together with the gene expression data used to reverse engineer it. The B-cell expression data contains 211 samples representing several normal and tumor human B-cell phenotypes profiled on Affymetrix H-GU95Av2 (Gene Expression Omnibus series GSE2350)[1]. The provided dataset was generated from custom probe-clusters obtained by the the cleaner algorithm[3] and MAS5[2] normalization.

The following lines are an example for the use of *aracne2regulon* function to generate the *regulon* object from the ARACNe output data and the expression data:

```
> data(bcellExample)
> adjfile <- file.path(find.package("bcellExample"), "aracne", "bcellaracne.adj")
> regul <- aracne2regulon(adjfile, dset)

> print(regul)

Object of class regulon with 20 regulators, 3758 targets and 6013 interactions
```

# 4 Master Regulator Analysis (MARINa)

To illustrate this section, we are going to analyze part of the expression data from [1], consitent on 5 naive human B-cell, 5 memory B-cell, 5 centroblast and 5 centrocytes B-cell samples profiled on Affymetrix H-GU95Av2 gene arrays. The complete dataset is available from Gene Expression Omnibus (GSE2350), and here for convenience, we have included the 'cleaner'[3] processed and MAS5[2] normalized samples in the *bcellExample* package.

## 4.1 Generating the gene expression signatures

Lets assume that we are interested in identifying transcriptional regulators associated with the Germinal Center (GC) reaction. GC are the peripheral lymphoid organs where antigen-driven somatic hypermutation of the genes encoding the immunoglobulin variable region occurs, and are the main source of memory B cells and plasma cells that produce high-affinity antibodies. Centroblasts and centrocytes are the main B-cell phenotypes present in the GCs, they are derived from antigen-stimulated peripheral blood B-cells, and represent the most proliferative cellular physiologic phenotypes of the adult human body. Thus, we can obtain a gene expression signature for the GC formation by comparing GC (centroblasts and centrocytes) against naive B-cells. So we first select these two groups of samples:

```
> naiveBcell <- which(colnames(dset) %in% normalSamples[["N"]])
> GCBcell <- which(colnames(dset) %in% c(normalSamples[["CB"]], normalSamples[["CC"]]))
```

The *ssmarina* package includes the function *rowTtest* that efficiently performs Student's t-test by comparing two matrixes row-by-row. The *rowTtest* function takes two matrixes as arguments, the first one containing the 'test' samples and the second the 'reference' samples, and produces a list object containing the Student's t-statistic (`statistic`) and the test's p-value (`p.value`), that by default is estimated by a 2-tailed test.

```
> signature <- rowTtest(dset[, GCBcell], dset[, naiveBcell])
```

While we could define the GES by using the t-statistic, to be consistent with the z-score based null model for MARINa (see section 4.2), we will estimate z-score values for the GES:

```
> signature <- (qnorm(signature$p.value/2, lower.tail=F) * sign(signature$statistic))[, 1]
```

## 4.2 NULL model by sample permutations

A uniform distribution of the targets on the GES is not a good prior for MARINa. Given the high degree of co-regulation in transcriptional networks, the assumption of statistical independence of gene expression is unrealistic an can potentially lead to p-value underestimates. To account for the correlation structure between genes, we define a null model for MARINa by using a set of signatures obtained after permuting the samples at random. The function *ttestNull* performs such process by shuffling the samples among the 'test' and 'reference' sets, according to the re-sampling mode and number of permutations indicated by the parameters *repos* and *per*, respectively.

```
> nullmodel <- ttestNull(dset[, GCBcell], dset[, naiveBcell], per=1000, repos=T)
```

As output, the *ttestNull* function produces a numerical matrix of z-scores, with genes/probes in rows and permutation iterations in columns, than can be used as null model for the MARINa analysis.

## 4.3 MARINa

The last element required by MARINa that we are still missing is an apropriate cell context-specific regulatory network. We have included a B-cell regulatory network in the *bcellExample* package, and additional networks for human B-cell, glioma and breast carcinoma can be obtained from figshare (Table 1).

```
> regulon

Object of class regulon with 621 regulators, 6249 targets and 172240 interactions
```

The MARINa analysis is performed by the *marina* function. It requires a *GES*, *regulon object* and *null model* as arguments, and produces an object of class 'marina', containing the GES, regulon and estimated enrichment, including the Normalized Enrichment Score (NES) and p-value, as output.

```
> mrs <- marina(signature, regulon, nullmodel)
```

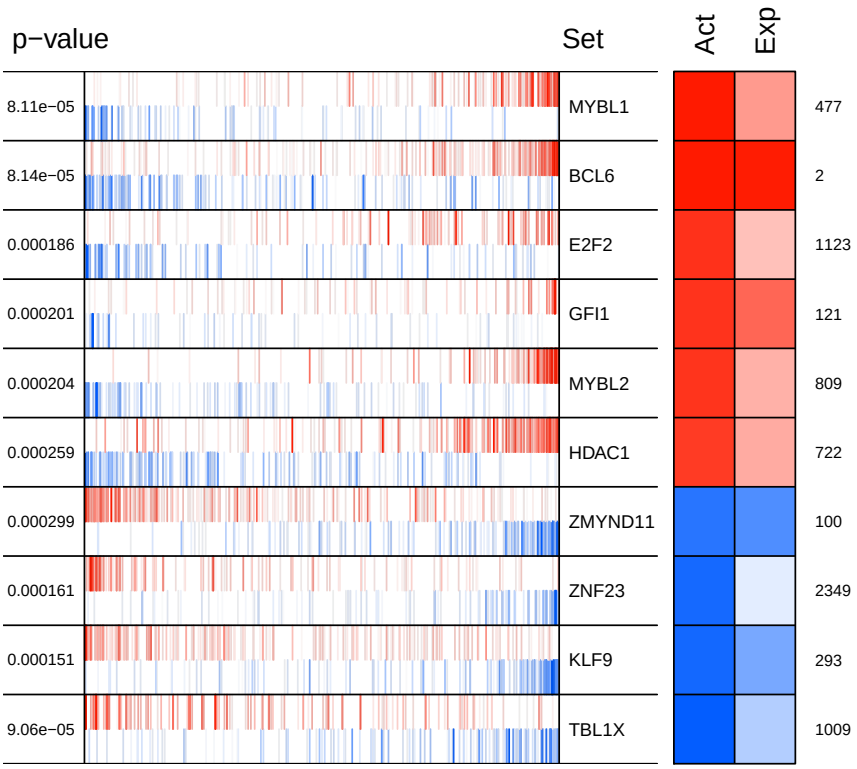
The reults can be summarized by the generic function *summary*, which takes the marina object and either the number of top regulators to report or a specific set of regulators to list. The default for this parameter is the top 10 master regulators (MRs).

```
> summary(mrs)
```

```
        Regulon Size   NES  p.value    FDR
MYBL1     MYBL1  251  3.94 8.11e-05 0.0153
BCL6       BCL6  401  3.94 8.14e-05 0.0153
E2F2       E2F2  214  3.74 1.86e-04 0.0153
GFI1       GFI1  143  3.72 2.01e-04 0.0153
MYBL2     MYBL2  240  3.71 2.04e-04 0.0153
HDAC1     HDAC1  360  3.65 2.59e-04 0.0153
ZMYND11 ZMYND11  452 -3.62 2.99e-04 0.0153
ZNF23     ZNF23  221 -3.77 1.61e-04 0.0153
KLF9       KLF9  337 -3.79 1.51e-04 0.0153
TBL1X     TBL1X  256 -3.91 9.06e-05 0.0153
```

A graphics representation of the results (MARINa plot) can be obtained by the generic function *plot*. It takes the *marina* object and either, the number of top differentially active regulators, or the names of the regulators to include in the plot as arguments. The default behavior is to plot the top 10 most differentially active MRs.

```
> plot(mrs, cex=.7)
```




The MARINa plot shows the projection of the negative (repressed, shown in blue color) and positive (activated, shown in red color) targets for each TF (vertical lines resembling a bar-code) on the GES (*x-axis*), where the genes in the GES were rank-sorted from the one most down-regulated to the one most upregulated in the 'test' vs 'reference' conditions. The optional two-columns heatmap displayed on the right side of the figure shows the inferred differential activity (first column) and differential expression (second column), with the rank of the displayed genes in the GES (shown on the right).

### 4.3.1 Leading-edge analysis

MARINa infers the relative activity of a regulatory gene based on the enrichment of its most closely-regulated targets on a given GES, but does not identify which are the target genes enriched in the GES. Subramanian et al. [14] proposed a method called leading-edge analysis to identify the genes driving the enrichment of a gene-set on a GES based on Gene Set Enrichment Analysis (GSEA). We implemented the leading-edge analysis in the *ledge* function of the *ssmarina* package. The function only has a 'marina' class object as argument and generates an updated 'marina' object that now includes a 'ledge' slot.

```
> mrs <- ledge(mrs)
> summary(mrs)
```

```
        Regulon Size   NES  p.value    FDR
MYBL1     MYBL1  251  3.94 8.11e-05 0.0153
```

```
BCL6       BCL6  401  3.94 8.14e-05 0.0153
E2F2       E2F2  214  3.74 1.86e-04 0.0153
GFI1       GFI1  143  3.72 2.01e-04 0.0153
MYBL2     MYBL2  240  3.71 2.04e-04 0.0153
HDAC1     HDAC1  360  3.65 2.59e-04 0.0153
ZMYND11 ZMYND11  452 -3.62 2.99e-04 0.0153
ZNF23     ZNF23  221 -3.77 1.61e-04 0.0153
KLF9       KLF9  337 -3.79 1.51e-04 0.0153
TBL1X     TBL1X  256 -3.91 9.06e-05 0.0153
                                            Ledge
MYBL1   BCL6, ENTPD1, PDE8A, RAPGEF5, + 140 genes
BCL6      KIF14, BUB1, DLGAP4, GINS1, + 217 genes
E2F2       SMARCA4, MCM7, RYK, CHAF1A, + 87 genes
GFI1       NDC80, DYRK2, ICAM2, CENPF, + 54 genes
MYBL2   SMARCA4, MCM7, TRIP13, GINS1, + 138 genes
HDAC1    BCL6, BTN3A2, DLGAP4, TAGLN, + 190 genes
ZMYND11 ANKRD26, EXTL2, IGFBP4, CTSC, + 234 genes
ZNF23     FKBP1A, PRUNE, UPF2, ZNF44, + 103 genes
KLF9        IFIT1, LPAR1, NID1, STOM, + 158 genes
TBL1X       CHUK, NSF, CCT3, C7orf44, + 121 genes
```

# 5 Beyond MARINa

## 5.1 Bootstrap MARINa

The effect of outlier samples on the gene expression signature can be reduced by the use of resampling techniques[12]. MARINA is capable of performing the analysis with bootstrap if a matrix of bootstraped signatures, instead of a vector, is given as *signature* argument. We implemened the function *bootstrapTtest* in the *ssmarina* package to generate this kind of bootstraped GES matrixes from the 'test' and 'reference' datasets. The function produces 100 bootstrap interactions by default.

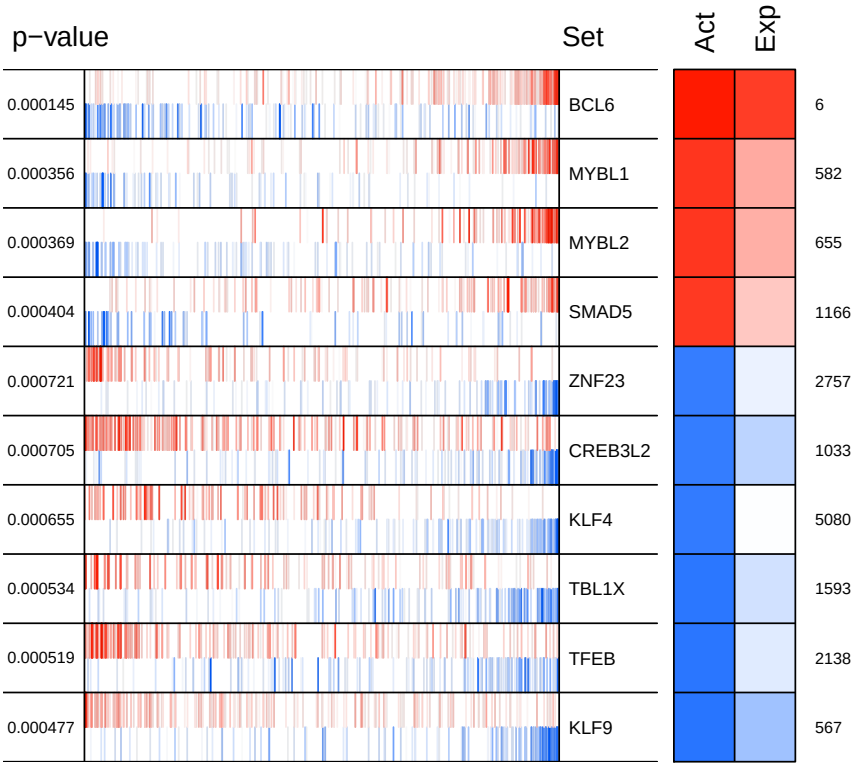
```
> signature <- bootstrapTtest(dset[, GCBcell], dset[, naiveBcell], per=100)
> mrs <- marina(signature, regulon, nullmodel)
```

By default, *marina* integrates the regulator activity results across all bootstraped iteration using the average, but this can be easily modified to use the median or mode values by the *bootstrapMarina* function:

```
> mrs <- bootstrapMarina(mrs, "mode")
```

Bootstraped marina results can be displayed in the same way as non-bootstraped results:

```
> plot(mrs, cex=.7)
```




## 5.2 Shadow analysis

A regulator may appear to be significantly activated based on its regulon's analysis, simply because several of its targets may also be regulated by a *bona fide* activated TF (shadow effect)[4, 5]. This constitutes a significant confounding issue, since transcriptional regulation is highly pleotropic, with individual targets being regulated by many TFs. MARINa and ssMARINa (section 6) address this issue by penalizig the contribution of the pleotropically regulated targets to the enrichment score. However, a post-hoc shadow analysis, as described in [4] can still be applied to the marina results with the function *shadow*. This function takes a class 'marina' object, and performs a shadow analysis on a selected number of top MRs indicated by the argument *pval*, which can be used to indicate either the enrichment p-value cutoff, the number of top MRs, or the names of the MRs to consider in the analysis.

```
> mrshadow <- shadow(mrs, pval=25)
```

As output, the *shadow* function produces an updated 'marina' object. The summary of it, generated by the *summary* function, will lists now not only the top MRs, but also the shadow pairs, in the form: $MR_1 -> MR_2$, indicating that part of the inferred $MR_2$ ativity is due to co-regulation of $MR_2$ target genes by $MR_1$.

```
> summary(mrshadow)
```

```
$MARINA.results
        Regulon Size   NES  p.value    FDR
MYBL1     MYBL1  251  3.58 0.000344 0.0501
MYBL2     MYBL2  240  3.51 0.000453 0.0501
BCL6       BCL6  358  3.46 0.000547 0.0501
TOP2A     TOP2A  749  3.36 0.000774 0.0501
GFI1       GFI1  143  3.23 0.001240 0.0501
ZNF23     ZNF23  182 -3.19 0.001410 0.0501
IRF5       IRF5  174 -3.21 0.001310 0.0501
KLF9       KLF9  298 -3.24 0.001190 0.0501
ZNF274   ZNF274  160 -3.28 0.001050 0.0501
ZMYND11 ZMYND11  452 -3.33 0.000866 0.0501

$Shadow.pairs
 [1] "BCL6 -> SMAD5"     "BCL6 -> TBL1X"     "BCL6 -> KLF4"
 [4] "BCL6 -> MEF2B"     "BCL6 -> ATF5"      "BCL6 -> JUN"
 [7] "MYBL2 -> TBL1X"    "MYBL2 -> KLF4"     "MYBL2 -> ZNF23"
[10] "MYBL2 -> MEIS2"    "MYBL2 -> ZNF211"   "MYBL2 -> ZNF185"
[13] "KLF9 -> TBL1X"     "KLF9 -> KLF4"      "KLF9 -> JUN"
[16] "TFEB -> ZNF185"    "TBL1X -> ATF5"     "CREB3L2 -> MEF2B"
[19] "CREB3L2 -> E2F2"   "ZNF23 -> ZNF211"   "ZMYND11 -> ZNF32"
[22] "ZMYND11 -> NR2F6"  "ZMYND11 -> MEIS2"  "ZMYND11 -> E2F2"
[25] "TOP2A -> MEIS2"    "TOP2A -> MEF2B"    "TOP2A -> E2F2"
[28] "TOP2A -> ZNF211"   "TOP2A -> ZNF185"   "TOP2A -> ATF5"
[31] "TOP2A -> JUN"      "TOP2A -> HDAC1"    "JUN -> HDAC1"
[34] "FOXJ2 -> WHSC1"    "FOXJ2 -> BCL6"     "KLF9 -> SMAD5"
[37] "TOP2A -> SMAD5"    "MEF2B -> SMAD5"    "JUN -> SMAD5"
[40] "HDAC1 -> SMAD5"    "TFEB -> KLF9"      "ZNF274 -> KLF9"
[43] "KLF4 -> TFEB"      "TOP2A -> TFEB"     "ZNF274 -> TFEB"
[46] "FOXJ2 -> TFEB"     "CREB3L2 -> TBL1X"  "TOP2A -> TBL1X"
[49] "ZNF274 -> TBL1X"   "ZMYND11 -> KLF4"   "FOXJ2 -> KLF4"
[52] "TOP2A -> CREB3L2"  "ZNF274 -> CREB3L2" "ZNF32 -> ZNF23"
[55] "ZNF274 -> ZNF23"   "FOXJ2 -> ZNF32"    "WHSC1 -> ZNF32"
[58] "ZNF274 -> NR2F6"   "FOXJ2 -> NR2F6"    "FOXJ2 -> MEIS2"
[61] "FOXJ2 -> MEF2B"    "WHSC1 -> MEF2B"    "ZNF274 -> E2F2"
[64] "FOXJ2 -> E2F2"     "WHSC1 -> E2F2"     "FOXJ2 -> ATF5"
[67] "FOXJ2 -> HDAC1"
```

## 5.3 Synergy analysis

To predict synergistic interactions between regulators we first compute the enrichment of co-regulons, defined as the intersection between regulons. We expect that a combination of regulators will synergistically regulate a gene expression signature if their co-regulon show a significantly higher enrichment on the signature than the union of the corresponding regulons[6]. Co-regulon analysis was implemented in the *ssmarina* package by the *marinaCombinatorial* function. It takes a 'marina' object as argument and computes the enrichment of all co-regulons, generated from a selected number of MRs (indicated by the *regulators* parameter), on the GES. As an usage example, we are going to compute the enrichment of the co-regulons for the top 25 regulators,

```
> mrs <- marinaCombinatorial(mrs, regulators=25)
```

The comparison between the enrichment of the co-regulon versus the union of the corresponding regulons (synergy analysis) is implemented by the function *marinaSynergy*, which requires only a 'marina' object

generated by *marinaCombinatorial* and the number of permutations used to compute the p-values, which default is 1,000:

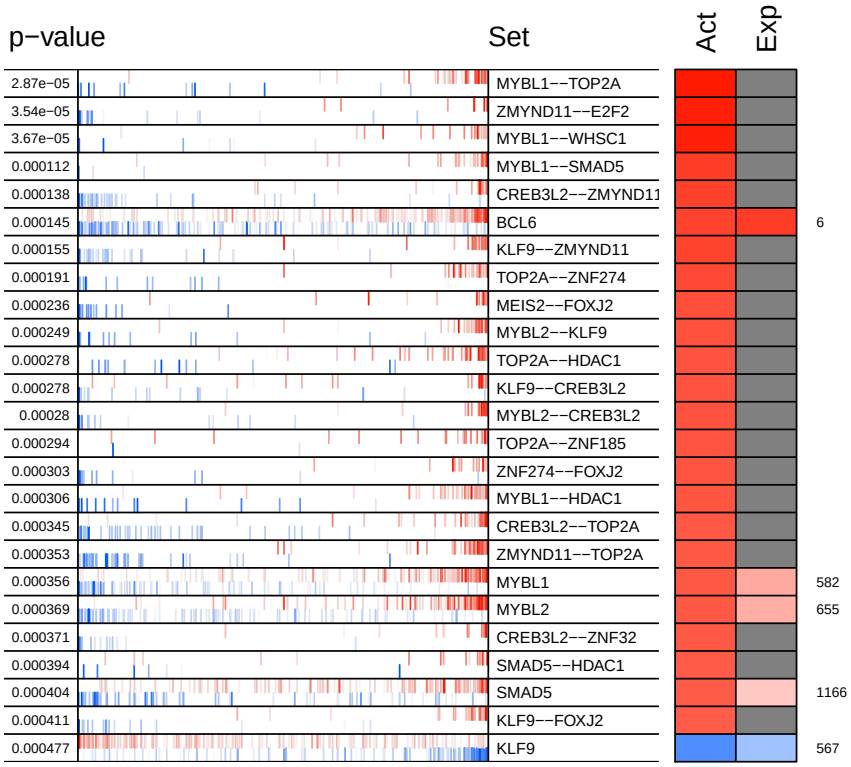
```
> mrs <- marinaSynergy(mrs)
```

The output of *marinaSynergy* is still a class 'marina' object and hence the *plot* and *summary* methods can be applied to it. The output of *summary* will include in this case the enrichment results for the co-regulons and the p-value for the predicted synergistic effect.

```
> summary(mrs)
```

| | Regulon | Size | NES | p.value | FDR | Synergy |
|---|---|---|---|---|---|---|
| MYBL1--TOP2A | MYBL1--TOP2A | 56 | 4.18 | 2.87e-05 | 0.00793 | 9.64e-05 |
| ZMYND11--E2F2 | ZMYND11--E2F2 | 30 | 4.14 | 3.54e-05 | 0.00793 | 9.71e-04 |
| MYBL1--WHSC1 | MYBL1--WHSC1 | 26 | 4.13 | 3.67e-05 | 0.00793 | 5.23e-02 |
| MYBL1--SMAD5 | MYBL1--SMAD5 | 34 | 3.86 | 1.12e-04 | 0.01110 | 2.40e-06 |
| CREB3L2--ZMYND11 | CREB3L2--ZMYND11 | 61 | 3.81 | 1.38e-04 | 0.01110 | 1.91e-03 |
| BCL6 | BCL6 | 401 | 3.80 | 1.45e-04 | 0.01110 | NA |
| KLF9--ZMYND11 | KLF9--ZMYND11 | 64 | 3.78 | 1.55e-04 | 0.01110 | 4.54e-07 |
| TOP2A--ZNF274 | TOP2A--ZNF274 | 44 | 3.73 | 1.91e-04 | 0.01110 | 3.85e-06 |
| MEIS2--FOXJ2 | MEIS2--FOXJ2 | 29 | 3.68 | 2.36e-04 | 0.01110 | 1.88e-05 |
| MYBL2--KLF9 | MYBL2--KLF9 | 44 | 3.66 | 2.49e-04 | 0.01110 | 2.55e-05 |

```
> plot(mrs, 25, cex=.7)
```




# 6 Single-sample MAster Regulator INference algorithm (ssMARINa)

ssMARINa is the extension of MARINa to single sample-based analysis. It effectively transforms a gene expression matrix to a regulators' activity matrix. The simplest implementation of ssMARINa is based on single-sample gene expression signatures obtained by scaling the probes or genes (subtracting the mean and dividing by the standard devition of each row). A gene expression matrix and appropriate regulatory network are the minimum set of parameters required to perform a ssMARINa analysis with the function *ssmarina*.

```
> vpres <- ssmarina(dset, regulon)
```

The *ssmarina* function generates a matrix of regulator's activity, containing 621 regulators x 211 samples in our example.

```
> dim(vpres)
```

```
[1] 621 211
```

The differential activity of regulators between groups of samples, for example between germinal center B-cell and Naive B-cells, can be obtained by any hypothesis testing statistical method, like for example the Student's t-test:

```
> tmp <- rowTtest(vpres[, GCBcell], vpres[, naiveBcell])
> data.frame(Gene=rownames(tmp$p.value), t=round(tmp$statistic, 2),
+ "p-value"=signif(tmp$p.value, 3))[order(tmp$p.value)[1:10], ]
```

| | Gene | t | p.value |
|---|---|---|---|
| ZMYND11 | ZMYND11 | -19.40 | 5.58e-11 |
| TOP2A | TOP2A | 18.96 | 7.44e-11 |
| MYBL2 | MYBL2 | 18.80 | 8.26e-11 |
| JUN | JUN | -17.90 | 1.53e-10 |
| MYB | MYB | 17.81 | 1.64e-10 |
| MYBL1 | MYBL1 | 17.70 | 1.76e-10 |
| ZNF274 | ZNF274 | -17.69 | 1.77e-10 |
| ZNF23 | ZNF23 | -17.58 | 1.92e-10 |
| ZNF32 | ZNF32 | -17.13 | 2.65e-10 |
| TSC22D3 | TSC22D3 | -16.91 | 3.12e-10 |

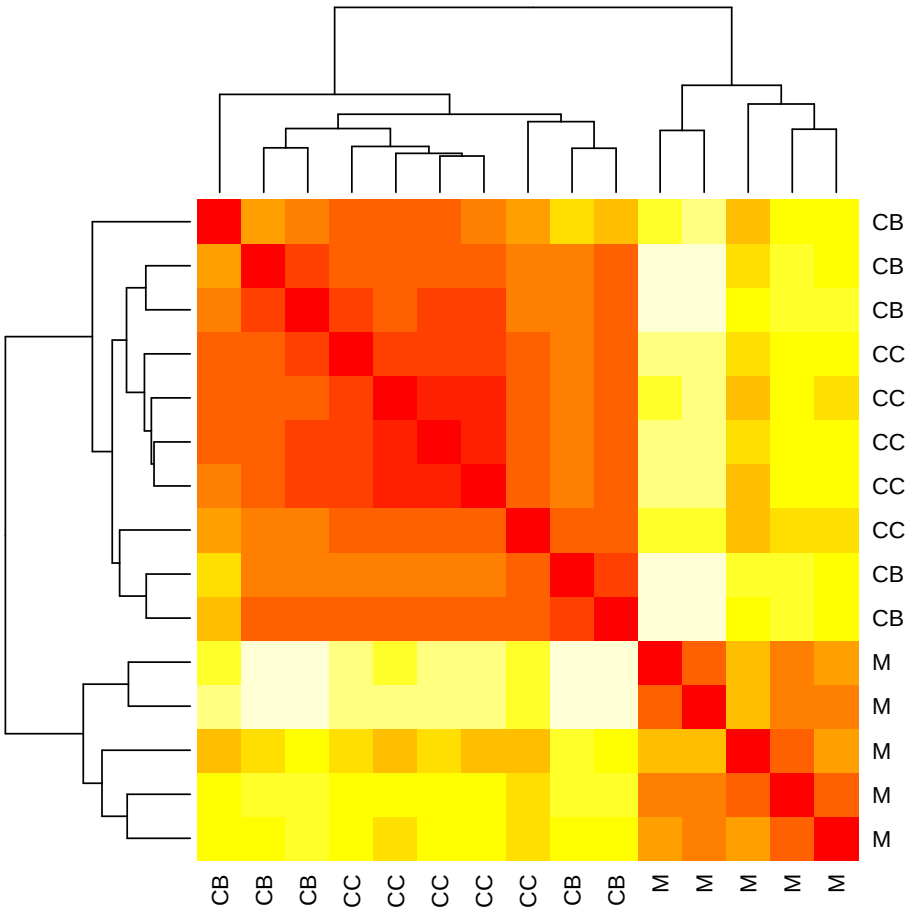
## 6.1 Running ssmarina with a null model

ssMARINa computes the normalized enrichment score (NES) analytically based on the assumption that in the null situation, the target genes are uniformly distributed on the gene expression signature. Because the extensive co-regulation of gene expression taking place in the cell, this assumption never holds true, and this is the reason why a null model based on sample permutations is used in MARINa to estimate NES. The same approach can also be used for ssMARINa, given that a set of samples is used as reference for the analysis. We can generate a set of GESs based on a set of reference samples, and the corresponding null model based on sample permutations, with the function *ssmarinaSignature*. It takes two matrixes as arguments, the first one containing the expression data for all the 'test' samples, and the second corresponding to the 'reference' samples. The number of permutations for the null model can be defined by the *per* argument, whose default value is 1,000.

```
> vpsig <- ssmarinaSignature(dset[, -naiveBcell], dset[, naiveBcell])
> vpres <- ssmarina(vpsig, regulon)
```

Because after ssMARINa analysis, the activity of different regulators is expressed in the same scale (normalized enrichment score), euclidean distance is an appropriate meassure of similarity between samples and we can, for example, perform an unsupervised hierarchical cluster analysis of the samples in a similar way we would do it in the case of gene expression data:

```
> d1 <- vpres[, match(unlist(normalSamples[names(normalSamples) != "N"]), colnames(vpres))]
> colnames(d1) <- rep(c("M", "CB", "CC"), c(5, 5, 5))
> dd <- dist(t(d1), method="euclidean")
> heatmap(as.matrix(dd), Rowv=as.dendrogram(hclust(dd, method="average")), symm=T)
```

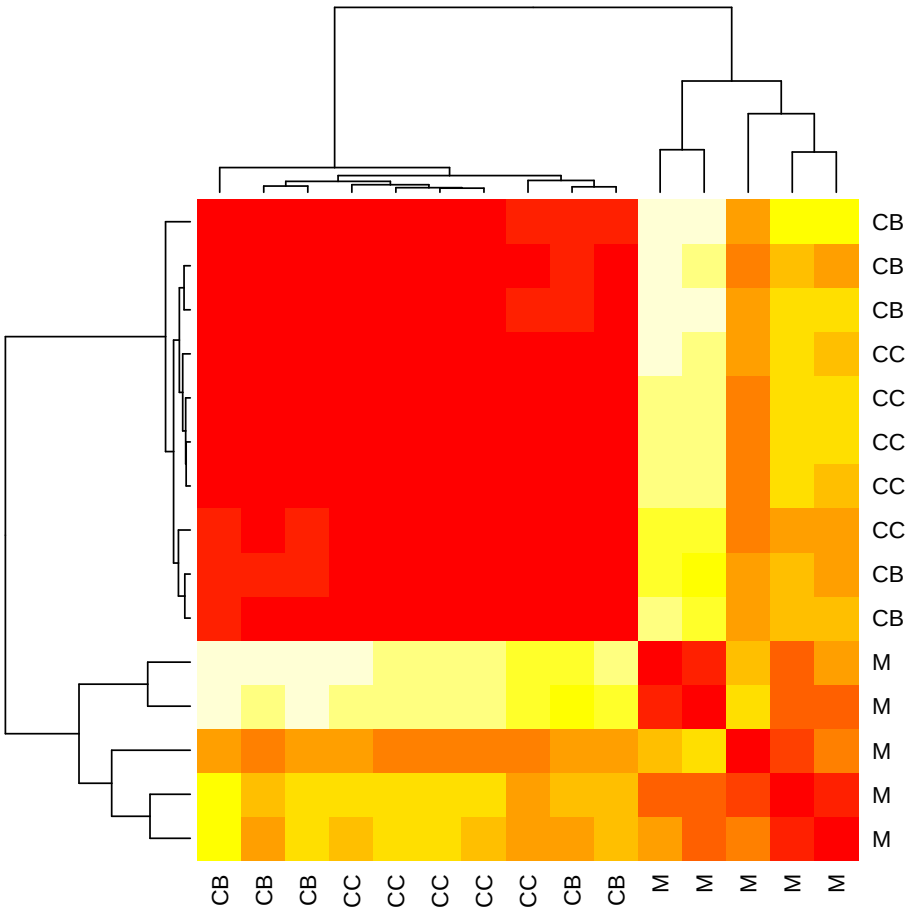



We have developed, and included in the *ssmarina* package, a function to compute the similarity between the columns of a gene expression or ssMARINa-predicted activity matrix. It follows the same concept as the two-tail Gene Set Enrichment Analysis (GSEA)[7], but it is based on the wREA-2 algorithm[12]. The *signatureDistance* function takes an expression or activity matrix as input, and generates a matrix of similarity scores between sample pairs, in the form of a 'similarityDistance' class object. In our example, because we want to compute the similarity between samples based on their regulators' relative activity vectors, we set the *scale.* argument to FALSE.

```
> dd <- signatureDistance(d1, scale.=F)
```

We can use the generic function *scale* to 'scale' the similary matrix in the rage [-1; 1], and the resulting matrix will be analogous to a correlation matrix. In this case, identical signatures will produce a similarity score equal to 1, while perfectly reversed signatures will produce similarity scores equal to -1. Orthogonal signatures will be characterized by similarity scores close to zero. As for other matrixes of similarity, the 'signatureDistance' class object can be transformed into a 'distance' class object with the method *as.dist*, which in turn can be used to perform, for example, cluster analysis of the samples.

```
> heatmap(as.matrix(as.dist(dd)), Rowv=as.dendrogram(hclust(as.dist(dd),
+ method="average")), symm=T)
```




## References

[1] Basso, K. et al. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 37, 382-90 (2005).

[2] Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. 2004. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 3 (Feb. 2004), 307-315.

[3] Alvarez, M. J., Sumazin, P., Rajbhandari, P. & Califano, A. Correlating measurements across samples improves accuracy of large-scale expression profile experiments. Genome Biol. 10, R143 (2009).

[4] Lefebvre, C. et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol. Syst. Biol. 6, 377 (2010).

[5] Jiang, Z. & Gentleman, R. Extensions to gene set enrichment. Bioinformatics (Oxford, England) 23, 306-13 (2007).

[6] Carro, M. S. et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature 463, 318-25 (2010).

[7] Julio, M. K. -d. et al. Regulation of extra-embryonic endoderm stem cell differentiation by Nodal and Cripto signaling. Development 138, 3885-3895 (2011).

[8] Klein, U. et al. Transcriptional analysis of the B cell germinal center reaction. Proc. Natl. Acad. Sci. USA. 100, 2639-44 (2003).

[9] Tothill, R. W. et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin. Cancer Res. 14, 5198-208 (2008).

[10] Bozovic, A. & Kulasingam, V. Quantitative mass spectrometry-based assay development and validation: From small molecules to proteins. Clin. Biochem. 46, 444-455 (2012).

[11] Wolf-Yadlin, A., Sevecka, M. & MacBeath, G. Dissecting protein function and signaling using protein microarrays. Curr. Opin. Chem. Biol. 13, 398-405 (2009).

[12] Alvarez, M. J. et al. Inferring global protein activity profiles by network based analysis of gene expression signatures. Manuscript in preparation.

[13] Margolin, A. A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1, S7 (2006).

[14] Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545-50 (2005).