Bioinformatics, 2018, 1–4
doi: 10.1093/bioinformatics/bty623
Advance Access Publication Date: 13 July 2018
Applications Note

Gene expression

GECO: gene expression correlation analysis after genetic algorithm-driven deconvolution

Jamil Najafov and Ayaz Najafov*

Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on March 6, 2018; revised on June 21, 2018; editorial decision on July 9, 2018; accepted on July 11, 2018

Abstract

Motivation: Large-scale gene expression analysis is a valuable asset for data-driven hypothesis generation. However, the convoluted nature of large expression datasets often hinders extraction of meaningful biological information.

Results: To this end, we developed GECO, a gene expression correlation analysis software that uses a genetic algorithm-driven approach to deconvolute complex expression datasets into two subpopulations that display positive and negative correlations between a pair of queried genes. GECO's mutational enrichment and pairwise drug sensitivity analyses functions that follow the deconvolution step may help to identify the mutational factors that drive the gene expression correlation in the generated subpopulations and their differential drug vulnerabilities. Finally, GECO's drug sensitivity screen function can be used to identify drugs that differentially affect the subpopulations.

Availability and implementation: http://www.proteinguru.com/geco/ and http://www.proteinguru.com/geco/codes/

Contact: ayaz_najafov@hms.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012), Genomics of Drug Sensitivity in Cancer database (Garnett *et al.*, 2012; Yang *et al.*, 2013) and COSMIC (GDSC/COSMIC) (Forbes *et al.*, 2015) provide a vast collection of gene expression and mutation data across ∼1000 cancer cell lines. Functional relationships between genes can be discovered by using gene expression correlation analysis as a hypothesis-generation tool. However, most biological databases are complex and convoluted, clouding clarity of such correlation analyses. To address this issue we developed GECO—a software for gene expression correlation analysis that uses a genetic algorithm-driven method for deconvoluting complex databases. Following deconvolution, GECO performs mutational enrichment and differential drug sensitivity analyses to facilitate data-driven hypothesis generation. While GECO uses gene expression, mutation and drug sensitivity data from the aforementioned cell line databases to demonstrate the application of its features, other types of relevant datasets can also be deconvoluted and processed through GECO.

2 Materials and methods

GECO was implemented using R (RStudio, http://www.r-project.org/), the app was implemented using Shiny (http://shiny.rstudio.com/) and hosted at Shinyapps.io (http://shinyapps.io). The following packages were employed: shiny, gridExtra, ggplot2, ggrepel, splitstackshape, dplyer, shinyjs and DT (available from http://cran.rstudio.com/ or http://bioconductor.org/).

The gene expression and mutation datasets were downloaded from CCLE (http://www.broadinstitute.org/ccle/) and GDSC/COSMIC (http://www.cancerrxgene.org/). Drug sensitivity database for 265 drugs was downloaded from GDSC (Supplementary Table S1). While GECO uses expression and mutation information from the aforementioned databases, its open-source R codes can be readily adapted to analyze any other such databases.

Stringency of deconvolution determined the number of times the genetic algorithm cycles was to be repeated for the population to be deconvoluted. In other words, for stringency level 1, all 1036 cell lines were passed through 'mutagenesis' and selection once and this was repeated twice for stringency level 2, and so on (Supplementary Figs S1 and S2). The efficiency of the stringency and the deconvolution algorithm was tested by pooling two dummy datasets A and B (500 data points each, Supplementary Table S2) and applying the GECO's deconvolution function to re-identify the dataset subpopulations. Three randomly generated, distinct A and B datasets were analyzed.

Pearson's, Kendall's Tau and Spearman's rank correlation tests were performed for the correlation tests. Spearman's rank correlation test was used for sorting (used in the genome-wide correlation section) and deconvolution steps of the app, due to non-normal distribution of the gene expression data.

3 Results

The basic function of GECO is to generate binary (i.e. gene1 versus gene2, without deconvolution) gene expression correlation analyses using Pearson's, Kendall's Tau and Spearman's rank correlation tests across ~1000 human cancer cell lines, generating a visual output and a statistical analysis summary (Supplementary Fig. S3).

The second function of GECO produces rapid genome-wide gene expression correlation analyses for single-gene queries (i.e. gene1 versus genome, without deconvolution), across the ~1000 cancer cell lines. Here, the output files with the statistical analysis summaries are instantaneously generated, using GECO's large (31 GB) pre-computed database of gene expression correlation information across the genome. Top 20 strongest positive/negative gene expression correlation matches for a queried gene are displayed and the complete analysis is given as a downloadable CSV file (Supplementary Fig. S4).

The third and main functionality of GECO addresses the convoluted and often bimodal nature of gene expression datasets. To identify cancer cell line subpopulations where gene expression correlation between two queried genes is more significant than in the initial population (~1000 cell lines), we developed a genetic algorithm that performs cyclic population re-sampling iterations to deconvolute such complex datasets at a chosen stringency level.

During genetic algorithm-driven deconvolution, GECO generates two subpopulations where the queried gene pair presents strong positive versus negative expression correlations (Fig. 1 and Supplementary Figs S5 and S6, Movies S1 and S2). As a result, gene pairs, expression of which may not have strongly correlated across the initial population, are found to strongly correlate in a smaller subpopulation of cell lines identified by such deconvolution analysis. Post-deconvolution, GECO extracts the statistically significant (Fisher's exact test) differential mutational enrichment information found in the generated two subpopulations, allowing rapid identification of mutations that may contribute to the positive/negative correlations between the two queried genes.

To illustrate the accuracy of the deconvolution algorithm, we analyzed two artificial populations that were designed to have either strong positive or strong negative correlation patterns (Supplementary Table S2). These two populations were mixed and processed through GECO's deconvolution algorithm and the percentage of accurately re-identified population members ('strong positive' population members versus 'strong negative' population members) were determined, as a function of GECO deconvolution stringency (Supplementary Fig. S6B). The algorithm was 82% accurate at stringency level 5 and 93% accurate at stringency level 10.

To demonstrate the effectiveness of GECO, we mapped differential enrichment of the reported TP53 gain-of-function mutations (Supplementary Table S3) (Muller and Vousden, 2014; Olive *et al.*, 2004) in cell lines that show positive versus negative correlation between TNF-α expression and the expression of four transcriptional targets of the TNF-α/NF-κB pathway (Pahl, 1999), since these mutations have been reported to influence the activity of this pathway (Cooks *et al.*, 2013; Di Minin *et al.*, 2014). This analysis revealed an interesting differential clustering of various TP53 gain-of-function mutations in cell lines that contribute to positive versus negative correlation between TNF-α expression and expression of the four targets (Supplementary Fig. S7).

The second, post-deconvolution feature of GECO allows the user to perform drug sensitivity comparison between the two subpopulations, for any given drug-of-interest among the 265 drugs employed in the GDSC database. This function allows identification of relationships between the positive/negative correlation trends between a pair of genes that may drive sensitivity/resistance to the tested drugs and binary visualization of the differential drug sensitivity between the subpopulations. For example, the post-deconvolution subpopulation of cell lines that show a strong positive correlation between the expression of AXL and GAS6, are more sensitive to the chemotherapy agent Docetaxel, than those that show a strong negative expression correlation between these genes (Fig. 2A—left panel). Similar analysis applied to TP53 and CCND1 gene pair reveals a strong differential sensitivity to the HSP90 inhibitor AUY922 (Fig. 2A—right panel).

Thus, this post-deconvolution differential drug sensitivity analysis function can reveal potentially useful information about vulnerability of the subpopulations to various drugs in the context of positive/negative correlation between the queried genes, shedding light on the relationship between the genes and the drugs of interest.

The third post-deconvolution function of GECO allows performing virtual screens, to identify drugs to which the two subpopulations are differentially sensitive (Fig. 2B and Supplementary Figs S8 and S9, Movies S3 and S4). This analysis consists of iterative *t*-tests across the 265 drugs between the two subpopulations and presents a summary volcano plot depicting the fold (subpopulation1/subpopulation2) of mean subpopulation IC50 values versus the *t*-test *P*-values.

To illustrate this functionality, we performed the screen for the TNF-α/TNFAIP3 gene pair and found a strong differential sensitivity between the generated subpopulations to EGFR inhibitor Erlotinib, PI3Kγ inhibitor AS605240, ALK inhibitor TAE684, CDK inhibitor Roscovitine and IKKα/β inhibitor BMS345541 (Fig. 2B). Remarkably, EGFR and ALK have been implicated in positive regulation of the NF-κB pathway (Kuo *et al.*, 2007; Shostak and Chariot, 2015), while AS605240, Roscovitine and BMS345541 have been previously shown to inhibit the TNF-α/NF-κB pathway (Burke *et al.*, 2003; Dey *et al.*, 2008; Dutra *et al.*, 2011; Wei *et al.*, 2010).

A similar drug screen analysis for TP53 and its transcriptional target BAX revealed a differential sensitivity of the subpopulations to AMPK activator AICAR and Survivin inhibitor YM155 (Supplementary Fig. S8). In addition, this analysis for TNF and its transcriptional target MMP9 revealed differential sensitivity to the chemotherapy agents Doxorubicin and Mitomycin C (Supplementary Fig. S9).

In summary, GECO uses a genetic algorithm-driven data deconvolution process to address the complexity and heterogeneity of the

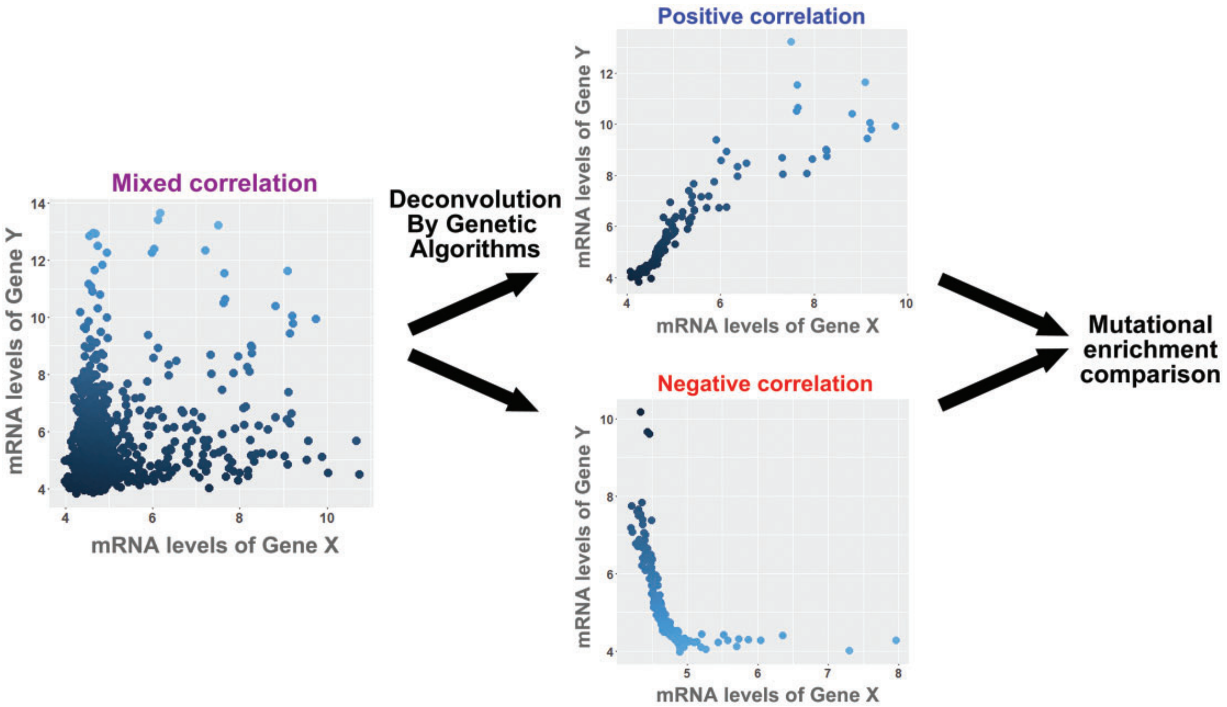


Fig. 1. GECO employs genetic algorithms to deconvolute complex expression datasets and reveal subpopulations with strong positive/negative gene expression correlations

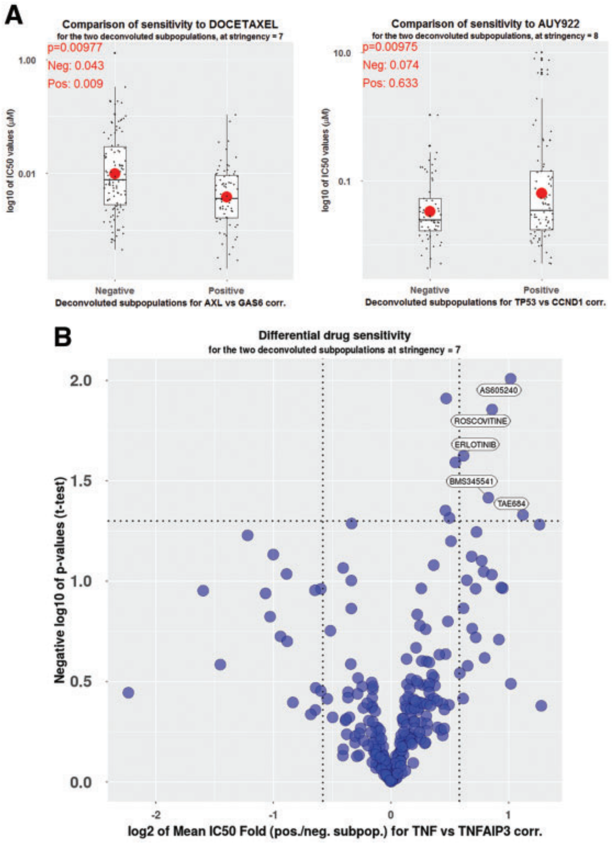


Fig. 2. The two differential drug sensitivity functions of GECO. (A) Two examples of the GECO's pairwise analysis of differential drug sensitivity between the two subpopulations generated from the deconvolution. The *t*-test *P*-values and mean values are indicated. (B) An example of GECO's differential drug sensitivity screen output

cancer cell line expression databases and identifies subpopulations where gene expression correlation for a queried gene pair is significant. GECO's mutational enrichment function that follows the deconvolution step may help to identify the mutational factors that drive the gene expression correlation in the generated subpopulations. The differential drug sensitivity comparison and differential drug sensitivity screen functions of GECO can be useful for studying cell signaling and for identification of drug vulnerabilities of cancer cells.

4 Discussion

We have developed GECO as an application for streamlined gene expression correlation analysis, with additional features, such as genome-wide expression correlation and a genetic algorithm-driven deconvolution of the expression datasets into subpopulations that contribute to positive/negative correlation between a pair of genes. Following the deconvolution step, GECO performs comparison of these subpopulations in the context of mutational enrichment and differential drug sensitivity. Finally, GECO's drug screen function reveals the drugs to which the subpopulations are differentially sensitive. These gene expression correlation, mutational enrichment and differential drug sensitivity analysis features of GECO facilitate the process of hypothesis generation.

By making an assumption that the mRNA expression levels of an upstream signaling regulator X (such as a receptor agonist or a transcription factor) should correlate positively and strongly with the mRNA expression levels of its downstream effector Y (which has been established to be transcriptionally regulated by X) if a cell line subpopulation has a highly activated X–Y pathway, data-driven hypothesis generation for subpopulations-specific pathway activity levels and drug resistance can be performed. For example, the subpopulations that have a strong TNF/TNFAIP3 gene expression correlation are likely to have a strongly activated TNF-α/NF-κB pathway, since TNFAIP3 is a transcriptional target of this pathway.

This would also explain why this subpopulation is more resistant to the inhibitors of this pathway—AS605240, Roscovitine and BMS345541. However, caution needs to be executed when making this assumption, especially in the light of heterogeneity and complexity of the genetic backgrounds of the subpopulations.

The virtual drug sensitivity screen analysis allows identification of drug vulnerabilities of the subpopulations, shedding light on how certain gene expression profiles may render cancer cells vulnerable to certain drugs. This, in turn, could help to understand cancer resistance to drugs—a clinically critical issue (Garraway and Jänne, 2012; Holohan *et al.*, 2013; Konieczkowski *et al.*, 2018).

The GECO's novel genetic algorithm-based deconvolution method for extraction of meaningful information from complex datasets and its derivatives can potentially be applied to other types of complex datasets. All of the analyses performed by GECO generate downloadable CVS files, for the tables, and high-resolution PDF files, for the graph outputs, simplifying the data storage, analysis and publication. GECO app (http://www.proteinguru.com/geco/) and its R source codes (http://www.proteinguru.com/geco/codes/) are freely available.

Acknowledgements

The authors thank Prof. Junying Yuan and the Harvard Medical School Research Computing for the access to the O2 cluster, and RStudio, Inc. Shinyapps.io for hosting the GECO frontend.

Author contributions

J.N. assisted with the software development. A.N. conceived the project, performed the R script development, compiled the GECO databases and supervised the software development. J.N and A.N. tested the software and wrote the manuscript.

Conflict of Interest: none declared.

References

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Burke,J.R. *et al.* (2003) BMS-345541 is a highly selective inhibitor of I kappa B kinase that binds at an allosteric site of the enzyme and blocks NF-kappa B-dependent transcription in mice. *J. Biol. Chem.*, **278**, 1450–1456.

Cooks,T. *et al.* (2013) Mutant p53 prolongs NF-κB activation and promotes chronic inflammation and inflammation-associated colorectal cancer. *Cancer Cell*, **23**, 634–646.

Dey,A. *et al.* (2008) R-Roscovitine simultaneously targets both the p53 and NF-kappaB pathways and causes potentiation of apoptosis: implications in cancer therapy. *Cell Death Differ.*, **15**, 263–273.

Di Minin,G. *et al.* (2014) Mutant p53 reprograms TNF signaling in cancer cells through interaction with the tumor suppressor DAB2IP. *Mol. Cell*, **56**, 617–629.

Dutra,R.C. *et al.* (2011) Inhibitor of PI3Kγ ameliorates TNBS-induced colitis in mice by affecting the functional activity of CD4 + CD25 + FoxP3 + regulatory T cells. *Br. J. Pharmacol.*, **163**, 358–374.

Forbes,S.A. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.

Garnett,M.J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Garraway,L.A. and Jänne,P.A. (2012) Circumventing cancer drug resistance in the era of personalized medicine. *Cancer Discov.*, **2**, 214–226.

Holohan,C. *et al.* (2013) Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer*, **13**, 714–726.

Konieczkowski,D.J. *et al.* (2018) A convergence-based framework for cancer drug resistance. *Cancer Cell*, **33**, 801–815.

Kuo,A.H. *et al.* (2007) Recruitment of insulin receptor substrate-1 and activation of NF-kappaB essential for midkine growth signaling through anaplastic lymphoma kinase. *Oncogene*, **26**, 859–869.

Muller,P.A.J. and Vousden,K.H. (2014) Mutant p53 in cancer: new functions and therapeutic opportunities. *Cancer Cell*, **25**, 304–317.

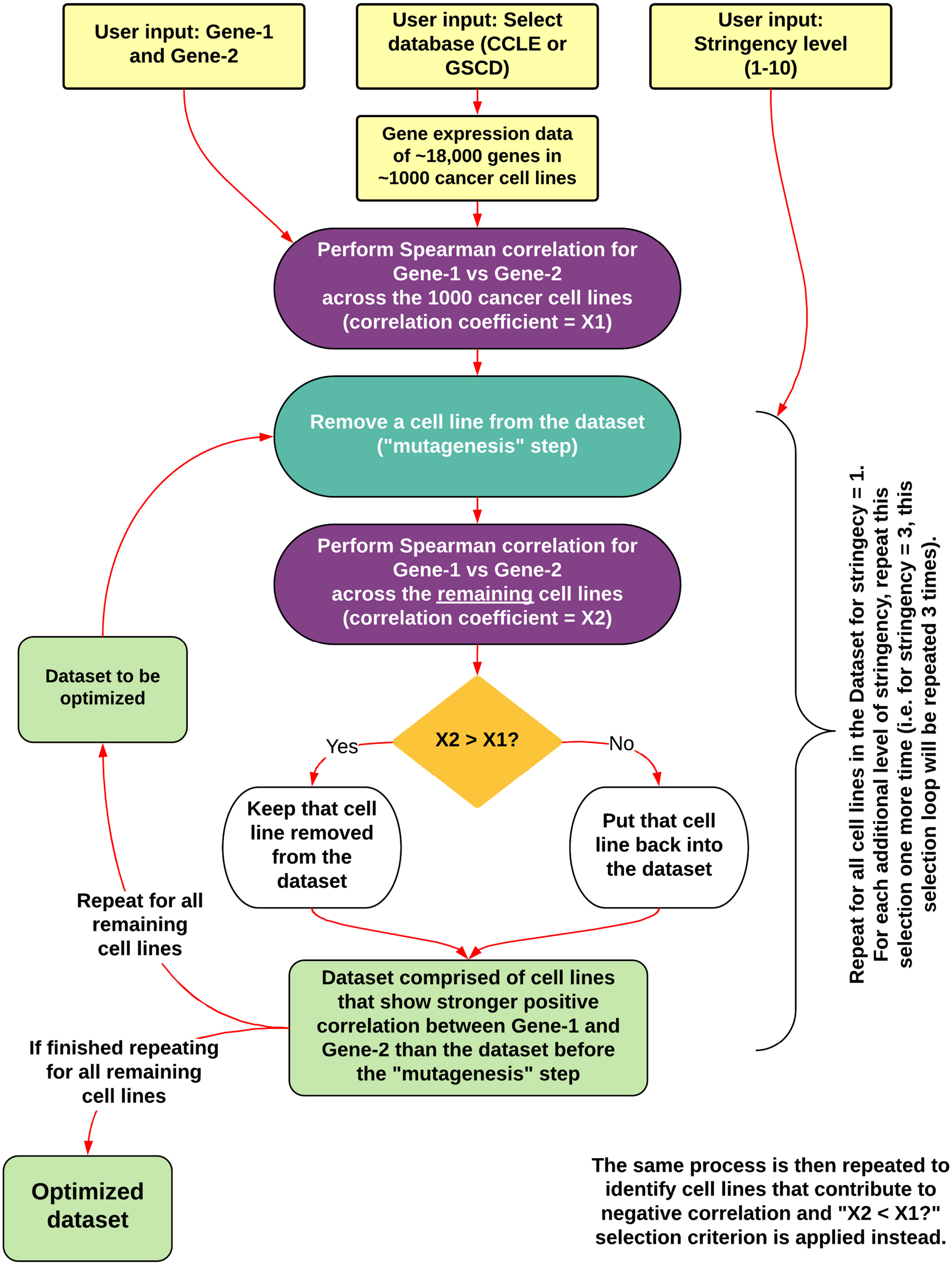
Olive,K.P. *et al.* (2004) Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome. *Cell*, **119**, 847–860.

Pahl,H.L. (1999) Activators and target genes of Rel/NF-kappaB transcription factors. *Oncogene*, **18**, 6853–6866.

Shostak,K. and Chariot,A. (2015) EGFR and NF-κB: partners in cancer. *Trends Mol. Med.*, **21**, 385–393.

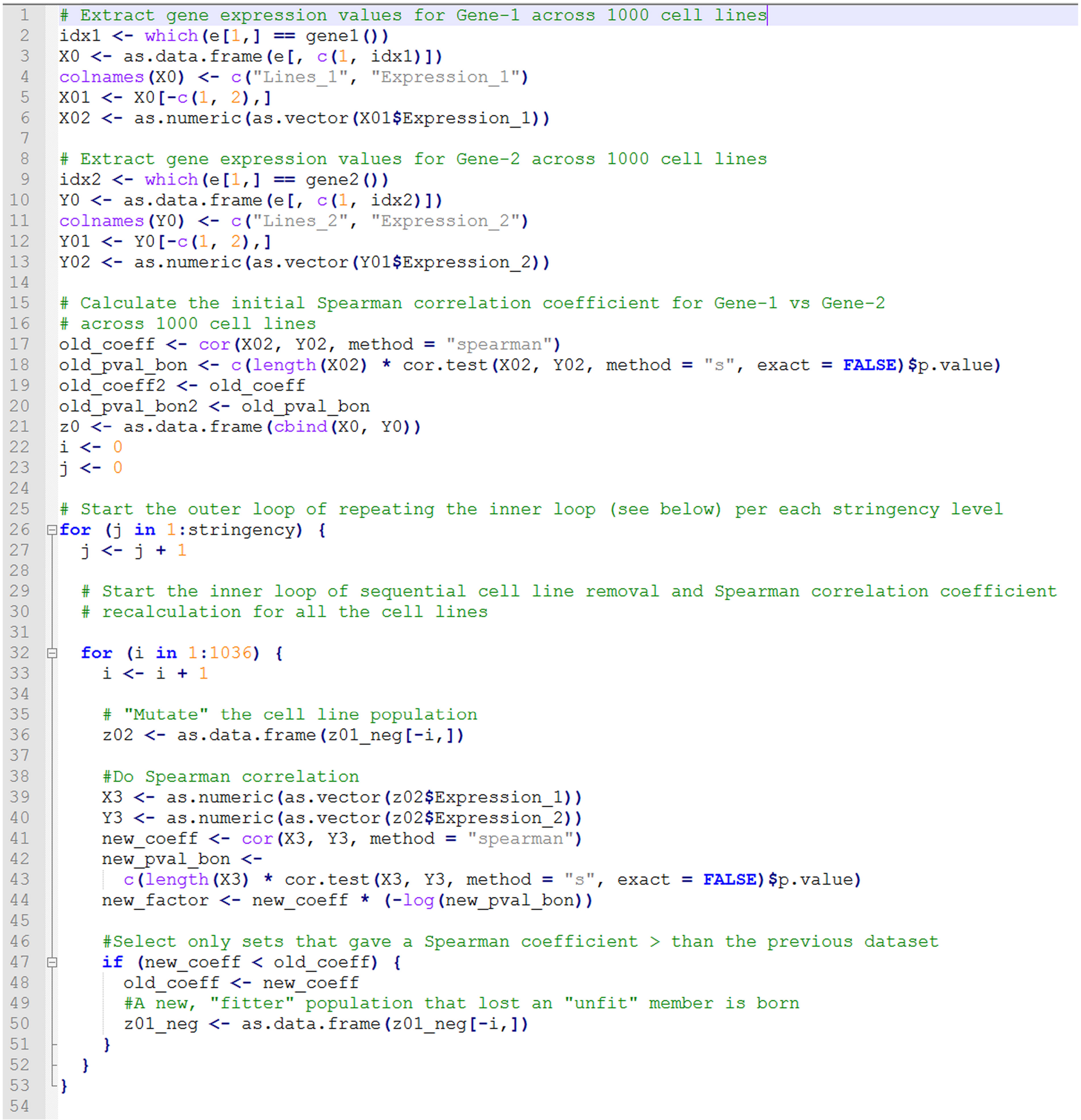
Wei,X. *et al.* (2010) A phosphoinositide 3-kinase-gamma inhibitor, AS605240 prevents bleomycin-induced pulmonary fibrosis in rats. *Biochem. Biophys. Res. Commun.*, **397**, 311–317.

Yang,W. *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–D961.


User input: Gene-1 and Gene-2
User input: Select database (CCLE or GSCD)
User input: Stringency level (1-10)
Gene expression data of ~18,000 genes in ~1000 cancer cell lines
Perform Spearman correlation for Gene-1 vs Gene-2 across the 1000 cancer cell lines (correlation coefficient = X1)
Remove a cell line from the dataset ("mutagenesis" step)
Perform Spearman correlation for Gene-1 vs Gene-2 across the remaining cell lines (correlation coefficient = X2)
X2 > X1?
Yes
No
Keep that cell line removed from the dataset
Put that cell line back into the dataset
Dataset comprised of cell lines that show stronger positive correlation between Gene-1 and Gene-2 than the dataset before the "mutagenesis" step
Repeat for all remaining cell lines
Dataset to be optimized
If finished repeating for all remaining cell lines
Optimized dataset
Repeat for all cell lines in the Dataset for stringecy = 1.
For each additional level of stringency, repeat this selection one more time (i.e. for stringency = 3, this selection loop will be repeated 3 times).

The same process is then repeated to identify cell lines that contribute to negative correlation and "X2 < X1?" selection criterion is applied instead.

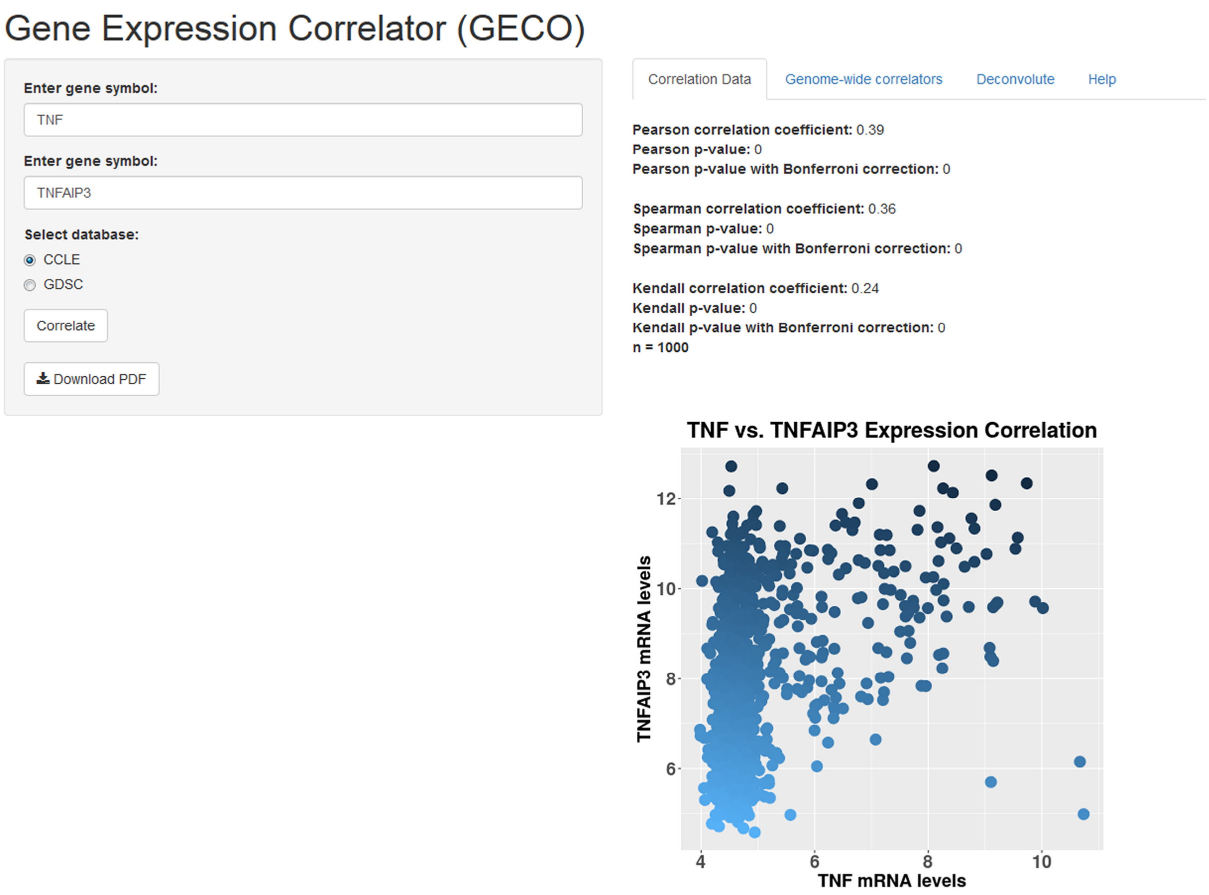
Supplementary Figure 1



```r
# Extract gene expression values for Gene-1 across 1000 cell lines
idx1 <- which(e[1,] == gene1())
X0 <- as.data.frame(e[, c(1, idx1)])
colnames(X0) <- c("Lines_1", "Expression_1")
X01 <- X0[-c(1, 2),]
X02 <- as.numeric(as.vector(X01$Expression_1))

# Extract gene expression values for Gene-2 across 1000 cell lines
idx2 <- which(e[1,] == gene2())
Y0 <- as.data.frame(e[, c(1, idx2)])
colnames(Y0) <- c("Lines_2", "Expression_2")
Y01 <- Y0[-c(1, 2),]
Y02 <- as.numeric(as.vector(Y01$Expression_2))

# Calculate the initial Spearman correlation coefficient for Gene-1 vs Gene-2
# across 1000 cell lines
old_coeff <- cor(X02, Y02, method = "spearman")
old_pval_bon <- c(length(X02) * cor.test(X02, Y02, method = "s", exact = FALSE)$p.value)
old_coeff2 <- old_coeff
old_pval_bon2 <- old_pval_bon
z0 <- as.data.frame(cbind(X0, Y0))
i <- 0
j <- 0

# Start the outer loop of repeating the inner loop (see below) per each stringency level
for (j in 1:stringency) {
  j <- j + 1

  # Start the inner loop of sequential cell line removal and Spearman correlation coefficient
  # recalculation for all the cell lines

  for (i in 1:1036) {
    i <- i + 1

    # "Mutate" the cell line population
    z02 <- as.data.frame(z01_neg[-i,])

    #Do Spearman correlation
    X3 <- as.numeric(as.vector(z02$Expression_1))
    Y3 <- as.numeric(as.vector(z02$Expression_2))
    new_coeff <- cor(X3, Y3, method = "spearman")
    new_pval_bon <-
      c(length(X3) * cor.test(X3, Y3, method = "s", exact = FALSE)$p.value)
    new_factor <- new_coeff * (-log(new_pval_bon))

    #Select only sets that gave a Spearman coefficient > than the previous dataset
    if (new_coeff < old_coeff) {
      old_coeff <- new_coeff
      #A new, "fitter" population that lost an "unfit" member is born
      z01_neg <- as.data.frame(z01_neg[-i,])
    }
  }
}

```

Supplementary Figure 2

A



Gene Expression Correlator (GECO)

Enter gene symbol: TNF

Enter gene symbol: TNFAIP3

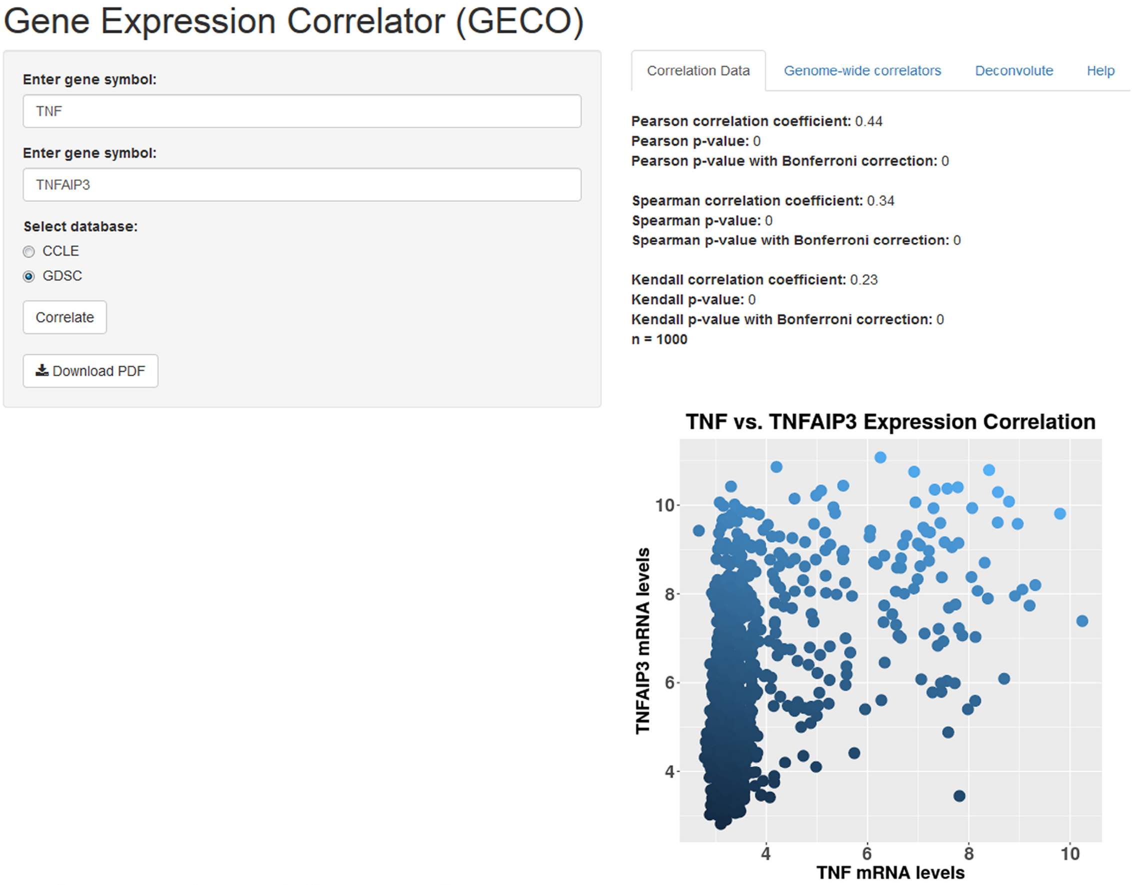
Select database:
● CCLE
○ GDSC

Pearson correlation coefficient: 0.39
Pearson p-value: 0
Pearson p-value with Bonferroni correction: 0

Spearman correlation coefficient: 0.36
Spearman p-value: 0
Spearman p-value with Bonferroni correction: 0

Kendall correlation coefficient: 0.24
Kendall p-value: 0
Kendall p-value with Bonferroni correction: 0
n = 1000

B



Gene Expression Correlator (GECO)

Enter gene symbol: TNF

Enter gene symbol: TNFAIP3

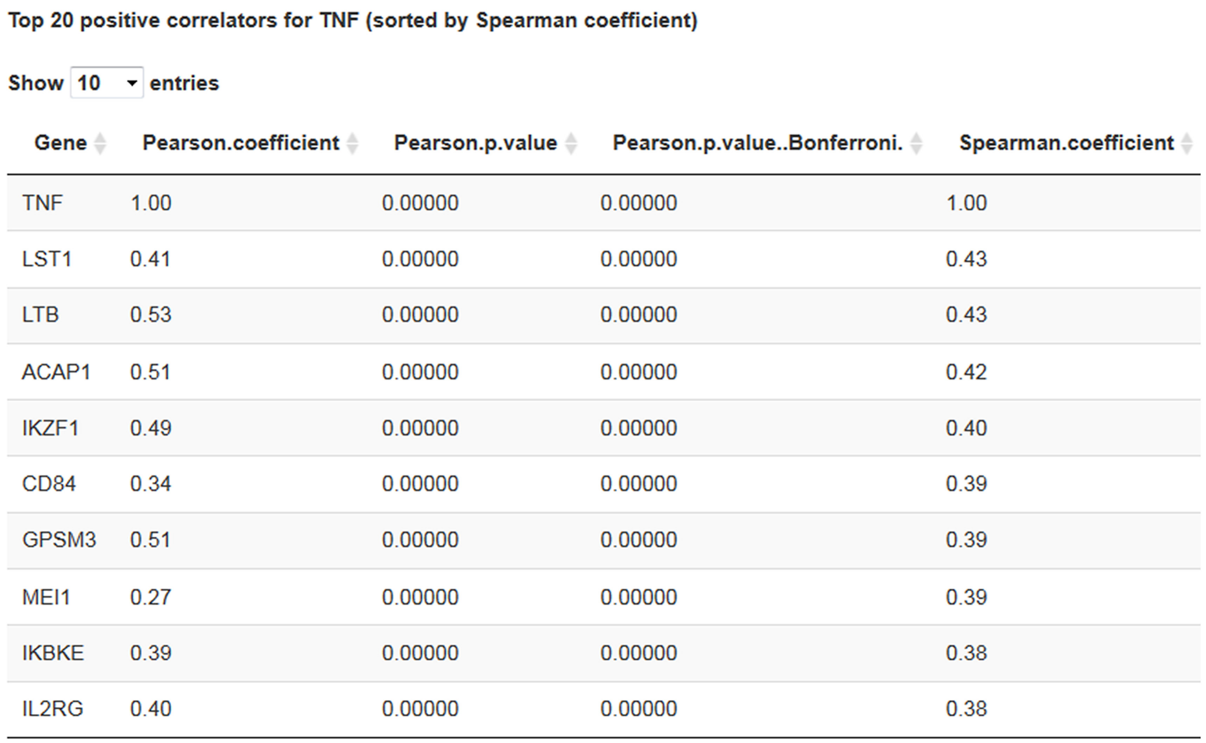
Select database:
○ CCLE
● GDSC

Pearson correlation coefficient: 0.44
Pearson p-value: 0
Pearson p-value with Bonferroni correction: 0

Spearman correlation coefficient: 0.34
Spearman p-value: 0
Spearman p-value with Bonferroni correction: 0

Kendall correlation coefficient: 0.23
Kendall p-value: 0
Kendall p-value with Bonferroni correction: 0
n = 1000

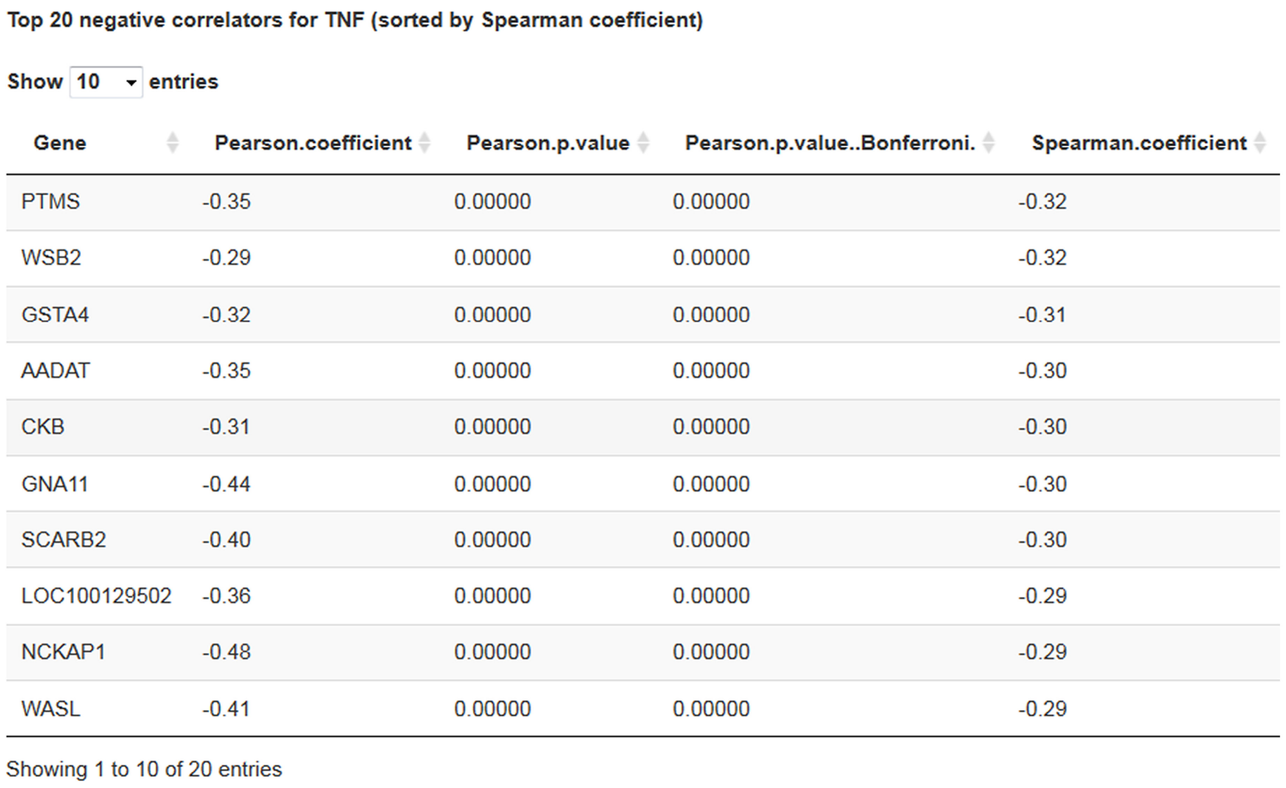
Supplementary Figure 3



Top 20 positive correlators for TNF (sorted by Spearman coefficient)

Show 10 entries

Gene	Pearson.coefficient	Pearson.p.value	Pearson.p.value..Bonferroni.	Spearman.coefficient
TNF	1.00	0.00000	0.00000	1.00
LST1	0.41	0.00000	0.00000	0.43
LTB	0.53	0.00000	0.00000	0.43
ACAP1	0.51	0.00000	0.00000	0.42
IKZF1	0.49	0.00000	0.00000	0.40
CD84	0.34	0.00000	0.00000	0.39
GPSM3	0.51	0.00000	0.00000	0.39
MEI1	0.27	0.00000	0.00000	0.39
IKBKE	0.39	0.00000	0.00000	0.38
IL2RG	0.40	0.00000	0.00000	0.38



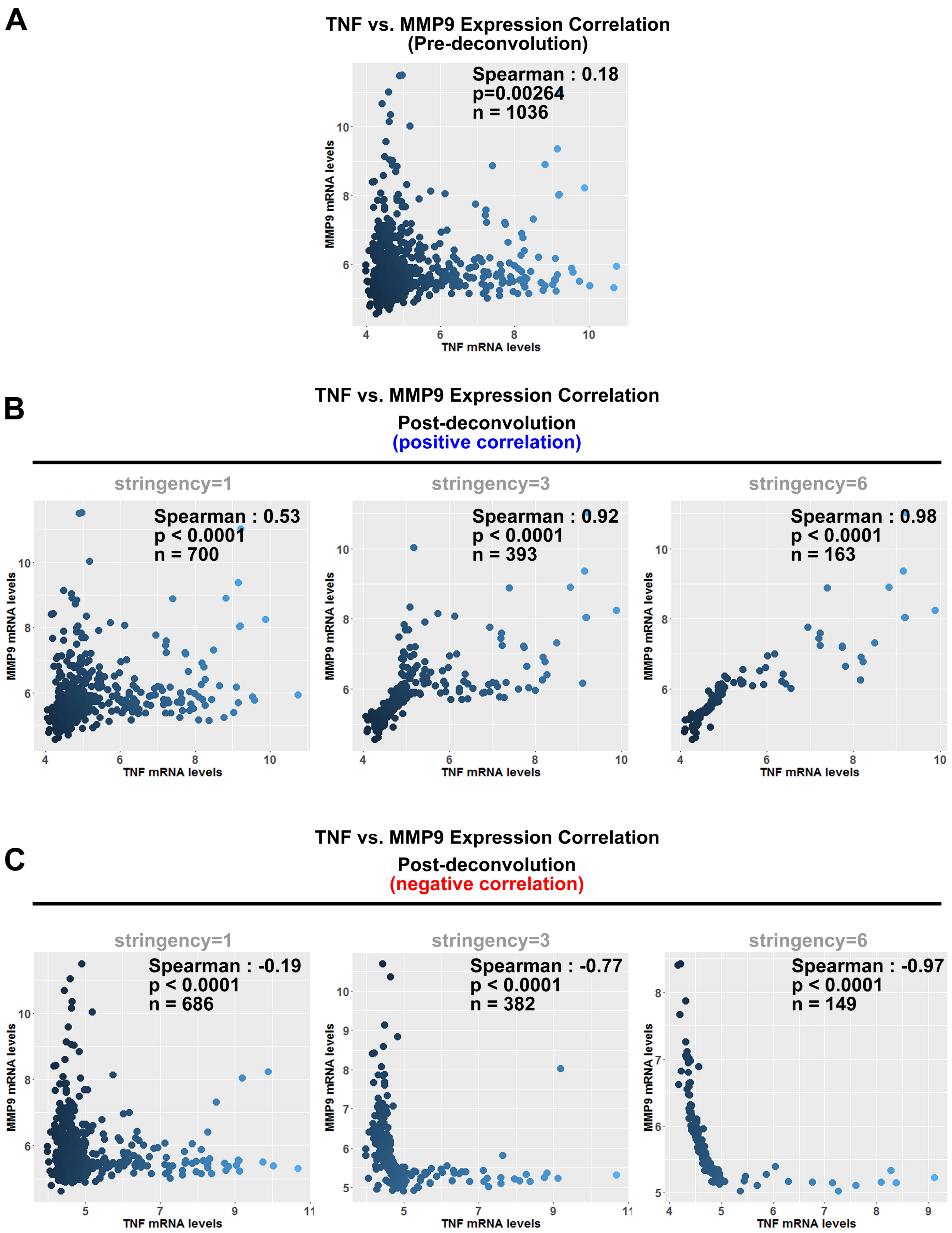
Top 20 negative correlators for TNF (sorted by Spearman coefficient)

Show 10 entries

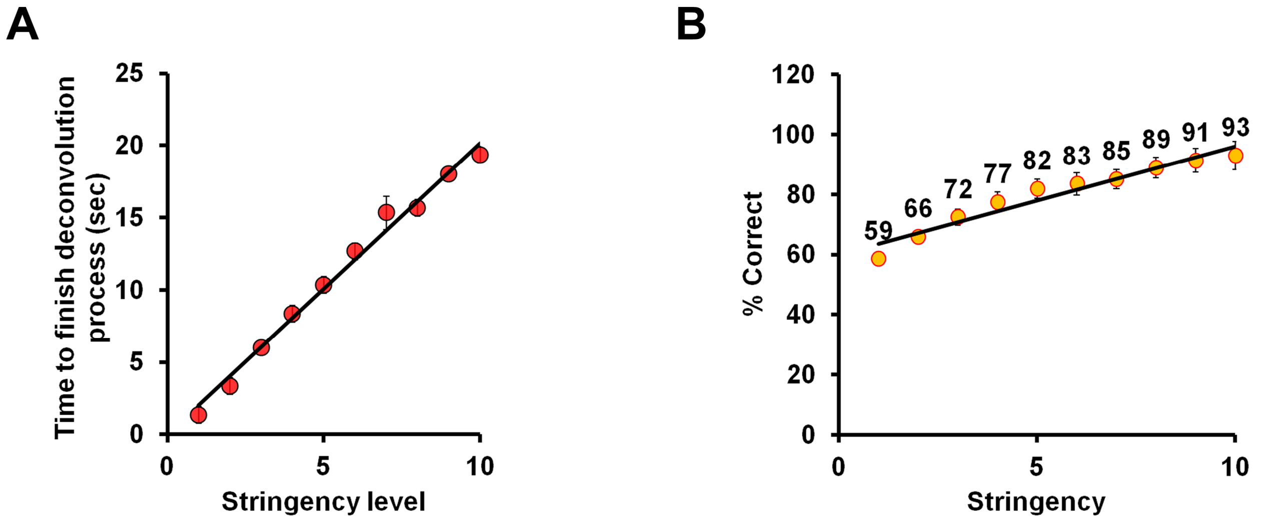
Gene	Pearson.coefficient	Pearson.p.value	Pearson.p.value..Bonferroni.	Spearman.coefficient
PTMS	-0.35	0.00000	0.00000	-0.32
WSB2	-0.29	0.00000	0.00000	-0.32
GSTA4	-0.32	0.00000	0.00000	-0.31
AADAT	-0.35	0.00000	0.00000	-0.30
CKB	-0.31	0.00000	0.00000	-0.30
GNA11	-0.44	0.00000	0.00000	-0.30
SCARB2	-0.40	0.00000	0.00000	-0.30
LOC100129502	-0.36	0.00000	0.00000	-0.29
NCKAP1	-0.48	0.00000	0.00000	-0.29
WASL	-0.41	0.00000	0.00000	-0.29

Showing 1 to 10 of 20 entries

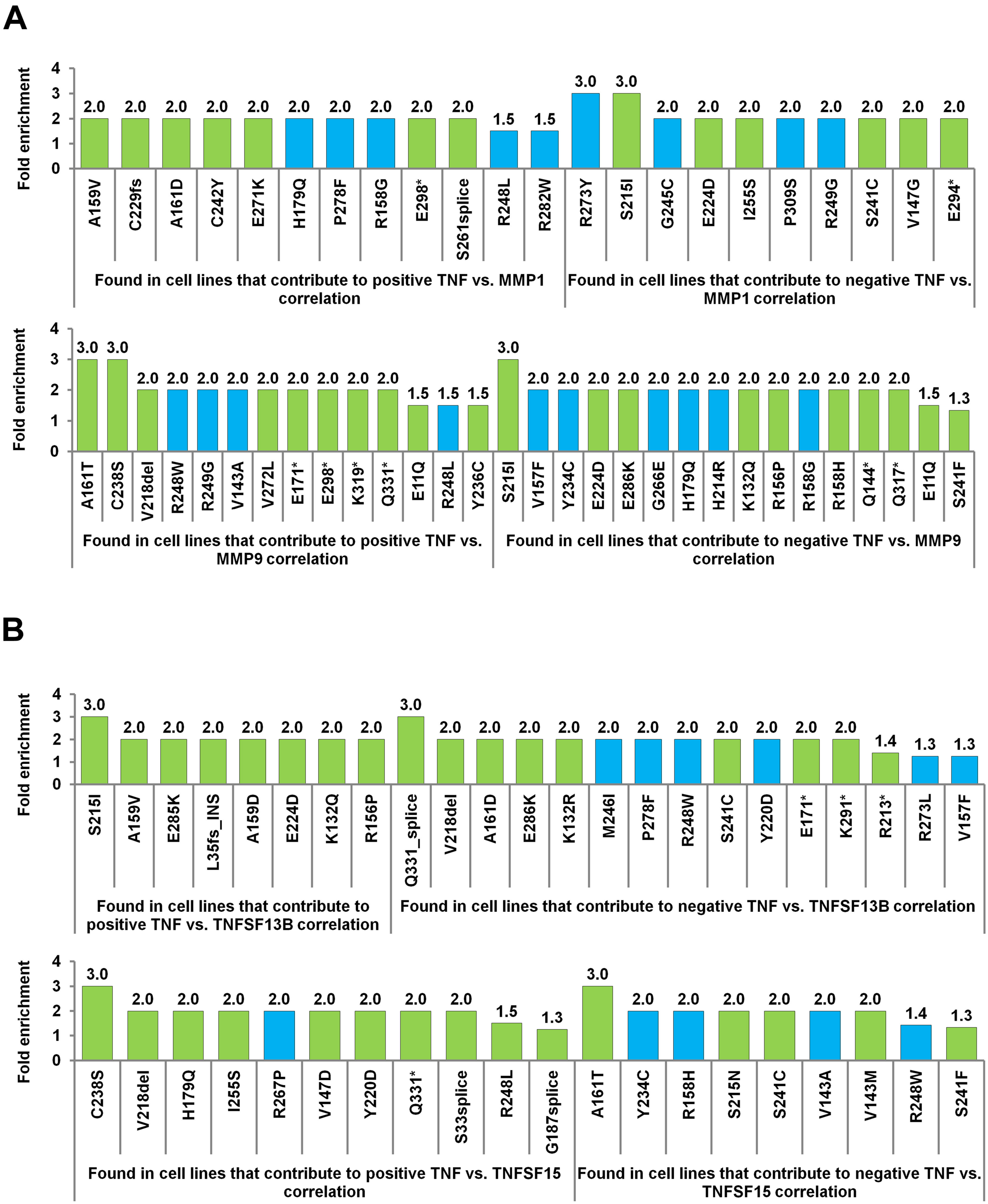
Supplementary Figure 4



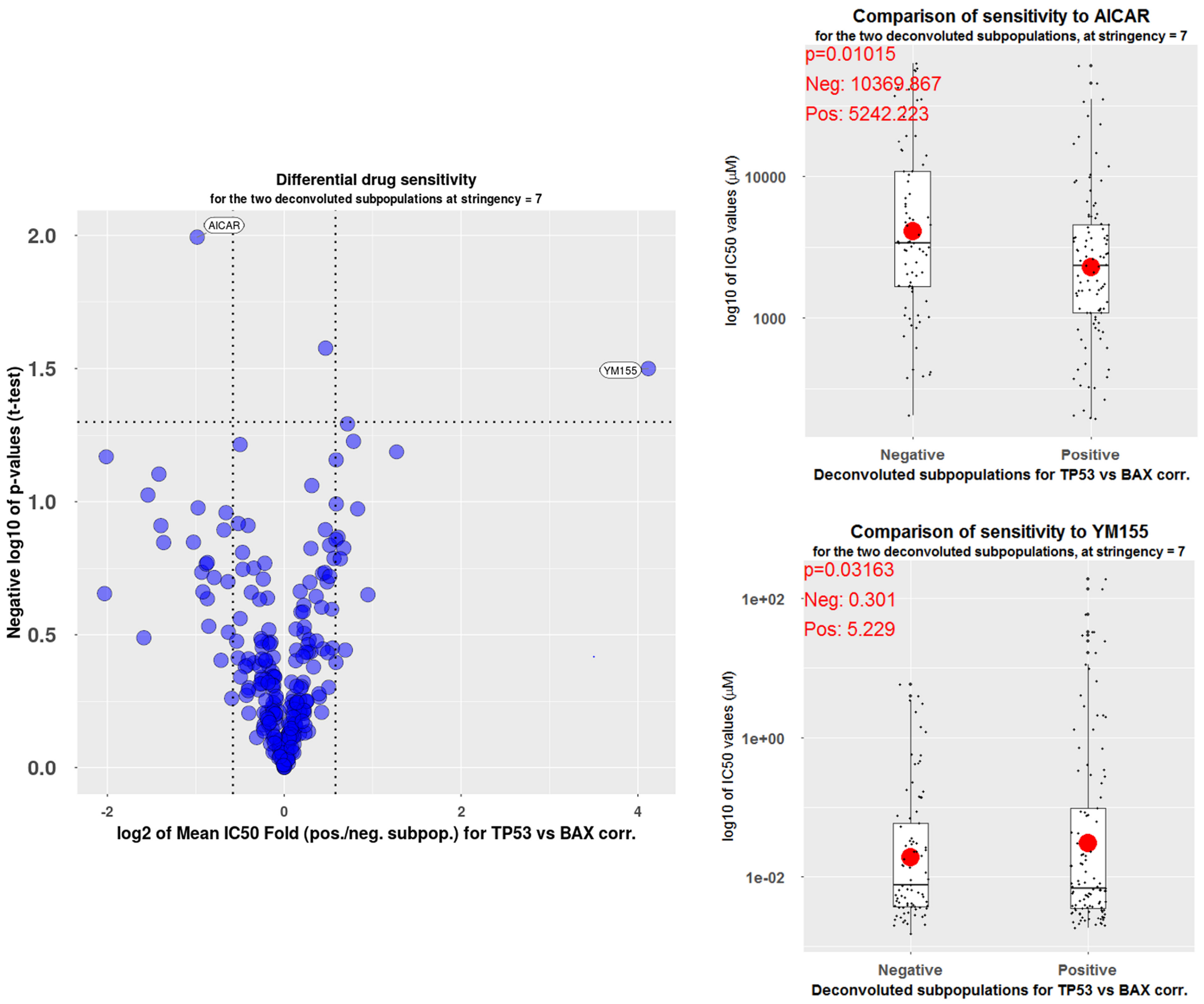
Supplementary Figure 5



Supplementary Figure 6



Supplementary Figure 7

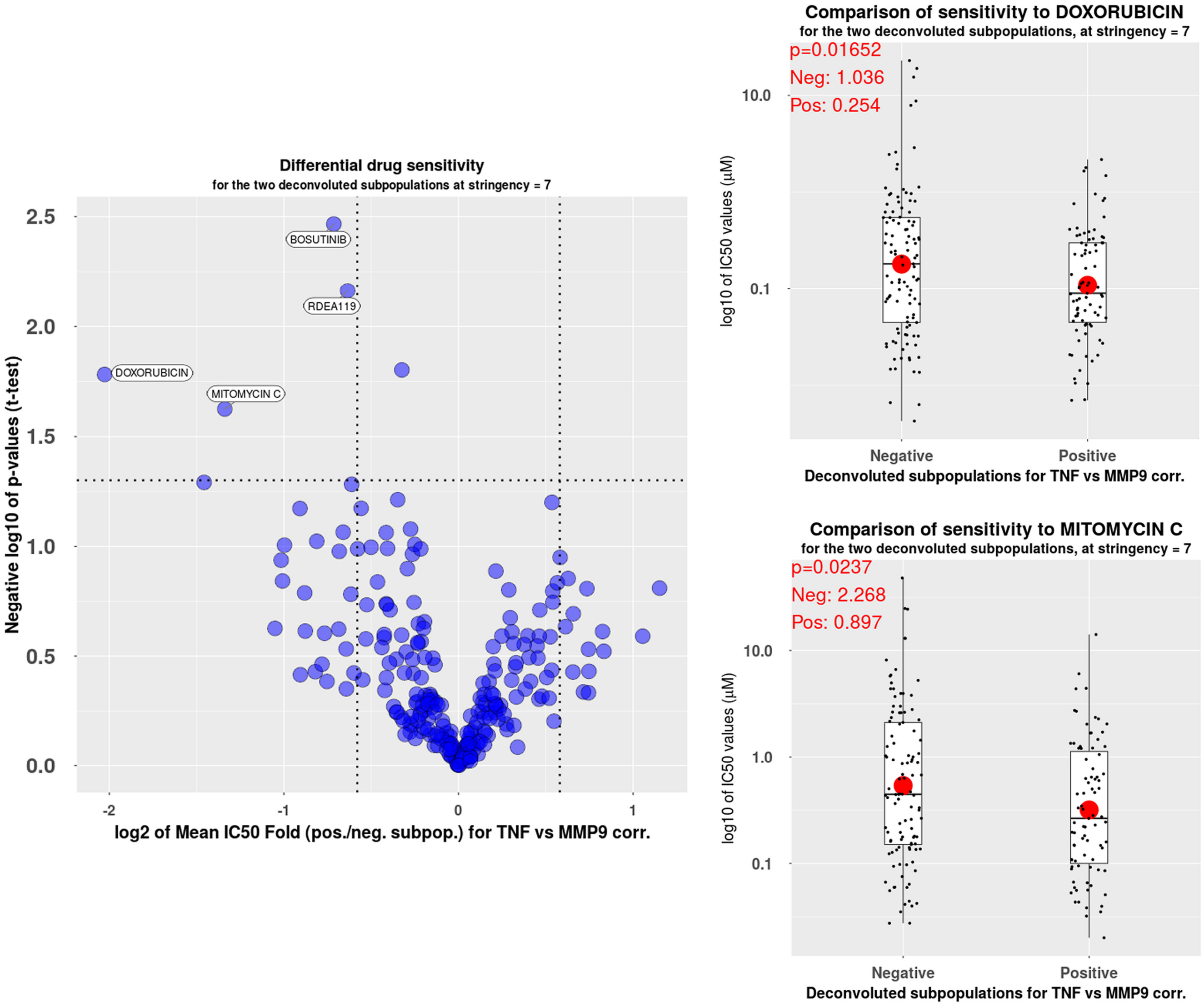


Differential drug sensitivity
for the two deconvoluted subpopulations at stringency = 7

Comparison of sensitivity to AICAR
for the two deconvoluted subpopulations, at stringency = 7

Comparison of sensitivity to YM155
for the two deconvoluted subpopulations, at stringency = 7

Supplementary Figure 8



Supplementary Figure 9

Supplementary Figure 1. Outline of the GECO's genetic algorithms. The initial dataset correlation (X1) is modified by performing cyclic "mutagenesis" of the original cell line population of 1036 cell lines, which is simply put: iterative removal of each cell line and subsequent analysis of the effect of this removal on the gene expression correlation of the queried pair of genes (by the Spearman correlation test) to obtain a new correlation coefficient (X2). If removing a cell line results in a stronger correlation (i.e. X2>X1 for positive correlation selection), the cell line is kept removed, if not, the cell line is not removed. This is repeated for all 1036 cell lines once, when the stringency is set to 1. Stringency level determines how many rounds this process is performed (from 1 to 10). With each round, more and more "outliers" are removed, making the subpopulation more and more "fit" and increasing the correlation coefficient, but also decreasing the subpopulation size. A parallel process is done for the selection of an optimized subpopulation with a strong negative correlation coefficient for the queried pair of genes.

Supplementary Figure 2. The R code for the GECO's deconvolution genetic algorithm. Two "for" loops drive the iterative selection for a "fitter" subpopulation with stronger positive/negative gene expression correlation across the 1036 cell lines, using Spearman correlation coefficient as a marker of selection of "fitness". "Mutagenesis" constitutes removal of each cell lines on each iteration of the selection cycle.

Supplementary Figure 3. GECO's pre-deconvolution binary gene expression correlation analysis output sample from the "Correlation Data" tab. (A) A sample output of the statistical analysis when TNF vs. TNFAIP3 genes are queried using CCLE database. (B) Same as (A), except GDSC database was used. The graphical output can be downloaded as a high-resolution PDF output.

Supplementary Figure 4. GECO's pre-deconvolution genome-wide gene expression correlation analysis sample output from the "Genome-wide correlators" tab. A sample tabular output of the genome-wide statistical analysis, using a large pre-computed gene expression correlation database, is instantaneously presented when a gene symbol (e.g. TNF) is queried. The correlators are sorted by the Spearman correlation coefficient. Top 20 positive correlators (top panel) and top 20 negative correlators (bottom panel) are instantaneously reported. The whole analysis (e. g. TNF vs. genome gene expression correlation analysis) can be downloaded as a CSV file.

Supplementary Figure 5. Comparison of the gene expression correlation between TNF and MMP9 before deconvolution (A) and after deconvolution (B and C). Outputs of three different stringency settings are shown (B and C). While TNF and MMP9 show a weak positive correlation in their expression across the 1036 cell lines before the deconvolution (A), GECO's genetic algorithm-driven deconvolution reveals a subpopulation of cell lines that show a much stronger positive expression correlation (B) and a subpopulation of cell lines that even show a strong negative expression correlation (C).

Supplementary Figure 6. GECO's deconvolution performance. (A) Linearity of the deconvolution performance (seconds) as a function of increased stringency level (1-10). Three gene pairs were queried and averages were plotted with S.D. shown in the error bars. (B) The accuracy of the GECO deconvolution algorithm as a function of stringency level. Two dummy subpopulations (see Supplementary Table 2) with strong positive and strong negative correlation were combined and then deconvoluted using GECO (using the algorithm described in the Supplementary Figure 2). Averages of % correctly re-identified subpopulation members from three independent exercises were plotted with S.D. shown in the error bars.

Supplementary Figure 7. GECO's mutational enrichment from the deconvoluted subpopulations. Mutational enrichment for gain-of-function TP53 mutations (blue) in cell lines that contribute to strong positive vs. negative gene expression correlation between TNF and four NF-κB target genes: MMP1, MMP9, TNFSF13B, and TNFSF15. Gain-of-function TP53 mutations are enriched either (A) in cell lines that contribute to both positive and negative correlation or (B) primarily enriched in cell lines that contribute to negative correlation. Fold enrichment is defined as frequency of a given mutation found in the subpopulation that contributes to the positive/negative correlation between the gene pair (TNF vs. MMP1/MMP9/TNFSF13B/TNFSF15) over the subpopulation that does not contribute.

Supplementary Figure 8. GECO's drug screen reveals differential sensitivity to AMPK activator AICAR and Survivin inhibitor YM155 in the cell line subpopulations that contribute to the positive/negative gene expression correlation between TP53 and BAX. The volcano plot depicts the screen results, where the significant ($p<0.05$) hits labeled. The boxplot depicts the pairwise sensitivity comparison for the indicated subpopulations and drugs. The t-test p-values and mean values are indicated.

Supplementary Figure 9. GECO's drug screen reveals differential sensitivity to the chemotherapy agent Doxorubicin in the cell line subpopulations that contribute to the positive/negative gene expression correlation between TNF and MMP9. The volcano plot depicts the screen results, where the significant ($p<0.05$) hits labeled. The boxplot depicts the pairwise sensitivity comparison for the indicated subpopulations and drugs. The t-test p-values and mean values are indicated.