**Supplementary Information**

MIXTURE: an improved algorithm for immune tumor microenvironment estimation based on gene expression data.

Elmer A. Fernández1,2\*, Yamil D. Mahmoud4, Florencia Veigas4, Darío Rocha2, Mónica Balzarini3, Hugo D. Lujan1, Gabriel A. Rabinovich5,6 and M. Romina Girotti4\*

1 Centro de Investigación y Desarrollo en Inmunología y Enfermedades Infecciosas (CIDIE), CONICET, Catholic University of Cordoba, Argentine.

2 Universidad Nacional de Córdoba, Facultad de Ciencias Exactas; Físicas y Naturales. Argentina,

3 Universidad Nacional de Córdoba, FCA. Estadística y Biometría & CONICET, Argentina

4 Laboratory of Translational Immuno-Oncology, Institute of Biology and Experimental Medicine (IBYME), National Council of Scientific and Technical Investigations (CONICET), C1428 Buenos Aires, Argentina.

5 Laboratory of Immunopathology, Institute of Biology and Experimental Medicine (IBYME), National Council of Scientific and Technical Investigations (CONICET), C1428 Buenos Aires, Argentina.

6 Department of Biological Chemistry, Faculty of Exact and Natural Sciences, University of Buenos Aires, C1428 Buenos Aires, Argentina.

\* Correspondence to: Elmer A Fernández, efernandez@cidie.ucc.edu.ar.

María Romina Girotti, mgirotti@dna.uba.ar

How to install MIXTURE and reproduce the results?

<https://github.com/elmerfer/MIXTURE>.App

Algorithm

Table 1: The MIXTURE Algorithm

|  |  |
| --- | --- |
| Step | MIXTURE Function: Inputs (*Y*, *X* ); Output: |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 | GOTO 3 |
| 9 | (this yields ) |

The ABBAS method was implemented using the source code from the supplementary material from Li et al. Genome Biology (2017) 18:127, which is implemented in TIMER tool9. The current implementation can be seen at <https://github.com/elmerfer/MIXTURE.App>

# Simulated scenarios:

**Scenario 1:**

**S1.1:** Pure cell type: Each of the 22 expression profiles of the cell-type signature matrix from Newman et al.8, was alternatively used as the subject expression profile from which to estimate its mixture proportion. (R code available at <https://github.com/elmerfer/MIXTURE.App/PaperMIXTURE/ScenarioS1.R>, Data has an RData file at https://github.com/elmerfer/MIXTURE.App/Data/LM22.rds)

X = LM22

for k=1:22

Y = select.column(X,k)(i.e column “k” is assigned to the subject mixture)

estimate by CIBERSORT , MIXTURE or ABBAS

end for

R code:

ncores2use <- 3L

#MIXTURE

out.mixture <- MIXTURE(expressionMatrix = LM22, signatureMatrix = LM22, functionMixture = nu.svm.robust.RFE, useCores = ncores2use, verbose = TRUE, iter = 1000, nullDist = "none")

#ABBAS

out.abbas <- MIXTURE(expressionMatrix = M.brca.n, signatureMatrix = LM22, functionMixture = ls.rfe.abbas, useCores = ncores2use, verbose = TRUE, iter = 1000, nullDist = "none")

#ABIS

brca.abis <- MIXTURE(expressionMatrix = M.brca.n, signatureMatrix = LM22, functionMixture = rlm.abis, useCores = ncores2use, verbose = TRUE, iter = 1000, nullDist = "none")

**S1.2:** Mixture of cell types without noise: between 2 and 8 cell-types of the immune cell-type molecular signature *LM22,* were randomly selected to simulate ***Y***, ***X*** being the *LM22* matrix (R code available at <https://github.com/elmerfer/MIXTURE.App/PaperMIXTURE/ScenarioS1.R>, Data has an RData file at https://github.com/elmerfer/MIXTURE.App/Data/betas.list.rds)

R code

set.seed(123)

***#nrep: maximum number of selected pure cell types from LM22 signature matrix***

betas.list <- lapply(1:1000, function(x, Mat, nrep) {

ns <- sample(2:nrep,1)

r <- runif(ns, 0.2,1)

id <- sample(22,ns) ***#random selection of ns (between 2 to 8) cell-types from LM22***

betas <- rep(0,22)

betas[id] <- r/sum(r) ***#random proportions***

A <- Mat %\*% betas ***#the simulated subject cell-type mixture***

list(beta = betas, id = id,A = A)

}, data.matrix(LM22), nrep = 8)

***#Building the simulated subject mixture population***

M.c <- do.call(cbind, lapply(betas.list, function(x) x$A))

out.betas.robust <- MIXTURE(expressionMatrix = M.c, signatureMatrix = LM22, functionMixture = nu.svm.robust.RFE, useCores = ncores2use, verbose = TRUE, iter = 1000, nullDist = "none")

out.betas.abbas <- MIXTURE(expressionMatrix = M.c, signatureMatrix = LM22, functionMixture = ls.rfe.abbas, useCores = ncores2use, verbose = TRUE, iter = 1000, nullDist = "none")

out.betas.abis <- MIXTURE(expressionMatrix = M.c, signatureMatrix = LM22, functionMixture = rlm.abis, useCores = ncores2use, verbose = TRUE, iter = 1000, nullDist = "none")

**S1.3:** Mixture of cell-types with noise: same as Scenario **S1.2** but adding to ***Y***a noise vector composed of a random sample of N values of gene expressions drawn from *LM22* (R code available at <https://github.com/elmerfer/MIXTURE.App/PaperMIXTURE/ScenarioS1.R>, Data has an RData file at https://github.com/elmerfer/MIXTURE.App/Data/betas.noise.list.rds)

…

A <- Mat %\*% betas + matrix(as.vector(data.matrix(Mat))[sample(nrow(Mat))],ncol=1) ***#the simulated subject cell-type mixture plus a random selection of 527 expression values from LM22***

…

**Scenario 2**:

**S2.1:** The Follicular Lymphoma (FL) data set, generated by taking lymph node biopsy samples and enumerating the immune cell sub-types using flow cytometry (Newman et al. 2015). In this case the process identified 3 leukocyte types (B, CD8 and CD4) in various proportions across 14 patient samples. Since LM22 contains 22 cell-types, we summarize the resulted proportions of those leukocytes types according to (Newman et al. 2015) (R code available at <https://github.com/elmerfer/MIXTURE.App/PaperMIXTURE/ScenarioS2.R>, Data has an RData file at <https://github.com/elmerfer/MIXTURE.App/Data/newman_fl.rda> - This data was extracted from the DTANGLE R library )

**S2.2:** The Peripheral Blood Mononuclear Cells (PBMC) data, generated from blood samples from 22 adults where the proportion of nine leukocytes were determined by flow cytometry. (Newman et al. 2015) (R code available at <https://github.com/elmerfer/MIXTURE.App/PaperMIXTURE/ScenarioS2.R>, Data has an RData file at <https://github.com/elmerfer/MIXTURE.App/Data/newman_fl.rda> - This data was extracted from the DTANGLE R library )

**Scenario 3**:

* TCGA data: RNA-seq data was downloaded by means of the TCGA-Assembler R packageS1 in December 2017 and clinical data (survival information) was downloaded from the Broad GDAC Firehose. The RNAseq data was normalized with library size and the expected distribution visually verified. This data was used to feed the deconvolution algorithms.
* Data of HNSCC biopsies: Gene expression data from (Rickman et al. 2008) were obtained from ArrayExpress (Accession code E-TABM-302). Microarray data was normalized with RMA before using for deconvolution analysis.
* Data of melanoma biopsies: Gene expression data from (Riaz et al. 2017) were obtained from their GitHub repository (https://github.com/riazn/bms038\_analysis/tree/master/data). RNAseq count data was normalized to FPKM (fragment per kilobase per million) through the Bioconductor R package DESeq2 1.18.1. The on-treatment biopsy from patient 32 was excluded from further analyses since it presented extreme expression values.

In <https://github.com/elmerfer/MIXTURE.App/PaperMIXTURE/> can be found the source R code, named BRCA\_TCGA\_MIXTURE\_paper.R, melanoma\_script.Rmd and HNSCC\_script.Rmd respectively.

Noise Constrain Threshold Selection

In order to select the noise constrain threshold, we run MIXTURE without constrain on the Simulated scenario S1.2 (Pure mixed samples with noise). The threshold was chosen as the third quantile level of the estimated coefficients without using a noise constraint value (i.e Δ=0)

Please take care of the working directory

source('…/MIXTURE/Utils/MIXTURE.DEBUG\_V0.1.R')

load("…/MIXTURE/Data/LM22.RData")

betas.noise.list <- readRDS(file = "…/MIXTURE/Data/betas.noise.list.rds")

M.pure.noise <- do.call(cbind, lapply(betas.noise.list, function(x) x$A))

out.betas.noise.cib <- MIXTURE(expressionMatrix = M.pure.noise, signatureMatrix = LM22, functionMixture = nu.svm.robust.RFE.old, useCores = 10L)

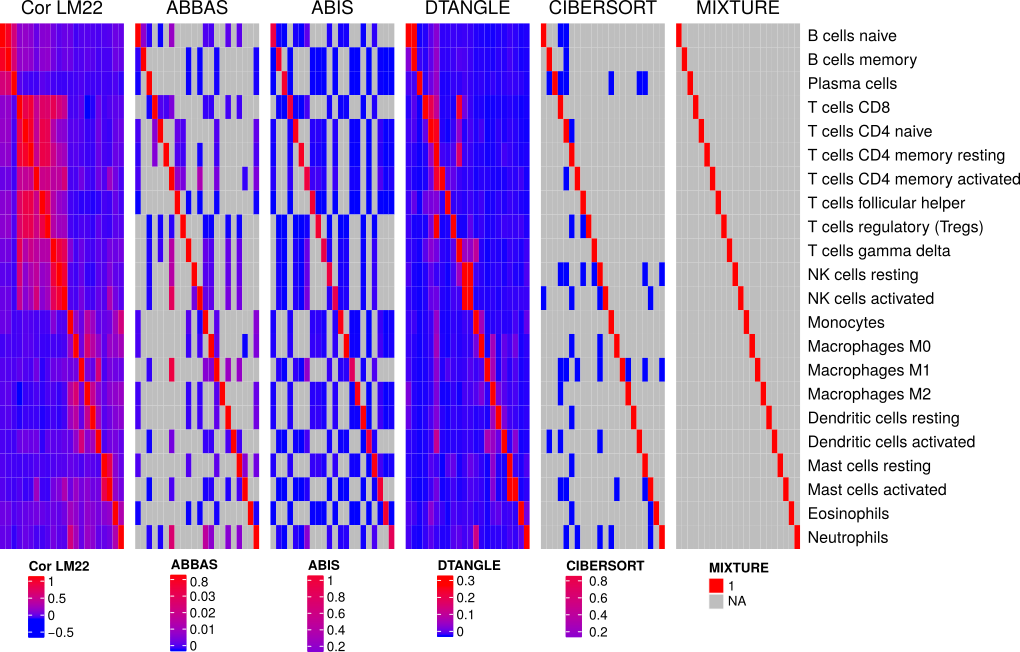
RFE v-SVR, without noise constraint

round(summary(as.vector(GetMixture(out.betas.noise.cib))),3)

# Min. 1st Qu. Median Mean 3rd Qu. Max.

#0.000 0.000 0.000 0.045 0.007 0.804

# Supplementary Figures:



Supplementary Figure 1: Panels from left to right :CorLM22: shows the correlation between cell-types expression profiles, showing high correlation in red and low correlations in blue. Rows and Columns corresponds to the different cell-types depicted as row names on the right. Strong correlation patterns emerge between B cell types, T cells types, NK cells and Mast cells. Right panels: The Estimated cell type proportions where rows represent the simulated cell-type for each sample in scenario S.1 (i.e the pure cell-type mixture) and columns represent the expected cell type proportions (i.e the ). ABBAS: Estimated proportions ( ) by ABBAS method (grey color: null cell-types. i.e cell-type not present or detected). ABIS: Cell type proportion estimated by means of the ABIS method. For these two methods, strong collinearity patterns can be advised resembling correlation patterns in the CorLM22 panel. DTANGLE: Cell type estimated by means of the DTANGLE method. It can be seen that this method does not provide null cell-type detections. Last right panels, CIBERSORT and MIXTURE where it can be seen that only MIXTURE is able to estimate the specific cell-type for each sample in scenario S.1

# References

S1) Weil M. et al. [Bioinformatics.](https://www.ncbi.nlm.nih.gov/pubmed/29272348) 2018 May 1;34(9):1615-1617. doi: 10.1093/bioinformatics/btx812.