iScience, Volume *27*

Supplemental information

An integrative multi-omics approach to characterize interactions between tuberculosis and diabetes mellitus

Caian L. Vinhaes, Eduardo R. Fukutani, Gabriel C. Santana, María B. Arriaga, Beatriz Barreto-Duarte, Mariana Araújo-Pereira, Mateus Maggitti-Bezerril, Alice M.S. Andrade, Marina C. Figueiredo, Ginger L. Milne, Valeria C. Rolla, Afrânio L. Kristki, Marcelo Cordeiro-Santos, Timothy R. Sterling, Bruno B. Andrade, Artur T.L. Queiroz, and for the RePORT Brazil Consortium

Supplementary Tables and Figures

Table S1. Characteristics of the study population, Related to Figure 1.

Characteristic	TB-DM	TB only	DM only	Non-TB/Non-DM controls	P-value
N	24	28	11	13	
Age – years	46.5 (37-55.7)	28.5 (25-40.5)	56 (51-59)	34 (27-50)	**<0.001**
Female – no. (%)	9 (37)	10 (35)	7 (63)	9 (69)	0.10
Race – no. (%)					0.6
White	3 (12)	3 (10)	0	1 (7)	
Black	14 (58)	11 (39)	6 (54)	7 (53)	
Pardo	7 (29)	14 (50)	5 (45)	5 (38)	
BMI (Kg/m²)	22 (18-26)	20 (18-22)	29 (27-32)	29 (24-33)	**<0.001**
HbA1c (%)	8.1 (6.6-11.3)	5.5 (5.2-5.6)	7.1 (6.9-8.4)	5 (4.8-5.3)	**<0.001**
AFB smear grade – no. (%)					0.36
0	5 (17)	1 (4)			
1+	8 (28)	6 (25)			
2+	4 (14)	3 (12)			
3+	11 (39)	14 (58)			
Cavities on chest X-ray – no. (%)	15 (62)	13 (46)			0.2

Data represent medians and interquartile ranges (age, BMI and HbA1c) and frequencies (female sex, race, AFB smear grade and cavities on chest X-ray). The Kruskal-Wallis test was used to compare continuous variables between the groups and the distributions while the Chi-square test was used to compare frequencies. AFB, acid-fast bacilli; BMI, Body Mass Index; HbA1c, Glycated hemoglobin. P-values in bold font are statistically significant.

Table S2. Top 20 variables from the multi-omic database that were identified by the Gini score and mean accuracy to discriminate between all the clinical groups, Related to Figure 1.

Variable	Gini score	Accuracy
AKR1C5P	2.0	0.5
CASKIN1	1.6	0.3
CDC42P3	2.0	0.5
CSMD2	2.4	0.5
DUSP3	2.5	0.2
ESYT3	1.2	0.3
FBXO6	3.5	1.0
HNRNPA1P61	1.2	0.2
IL15	1.6	0.2
IL6	2.7	0.3
KCNH7	1.7	0.25
LINC02009	2.6	1.2
LTE4	3.7	4.8
MIR1256	1.4	0.3
MMP28	2.3	1.3
PGD/LTE4	2.6	0.3
PGDM	8.5	12.6
PGD/TNE	1.7	0.2
SECTM1	3.4	1.6
11dTxB2	4.9	7.3

Table S3. Expression of multiomic markers used in the study according to the clinical groups, Related to Figure 1.

Parameter	Unit	Non-TB/Non-DM controls	DM	TB	TBDM	P-value
FBXO6	VST	6.5 (6.4-6.7)	6.4 (6.2-6.5)	7.2 (6.9-7.7)	7.3 (6.9-7.4)	**<0.001**
SECTM1	VST	8.5 (8.3-8.8)	8.4 (8.2-8.7)	9.8 (9.3-10.5)	9.9 (9.5-10.2)	**<0.001**
MMP28	VST	5.8 (5.7-6.0)	5.5 (5.4-5.6)	5.5 (5.4-5.6)	5.6 (5.5-5.7)	**0.005**
LINC02009	VST	5.4 (5.1-5.6)	5.4 (5.2-5.5)	5.5 (5.4-5.9)	5.9 (5.7-6.1)	**0.002**
PGDM	ng_mg_Cr	0.4 (0.3-0.5)	0.7 (0.5-0.9)	1.8 (1.4-2.5)	4.4 (1.9-13.4)	**<0.001**
11dTxB2	ng_mgCr	0.22 (0.20-0.27)	0.3 (0.2-0.4)	0.9 (0.7-1.6)	0.9 (0.7-2.4)	**<0.001**
LTE4	ng_mgCr1_A	0.07 (0.05-0.09)	0.2 (0.1-1.1)	0.1 (0.08-0.2)	0.2 (0.1-0.9)	**<0.001**

Data represents medians and interquartile ranges. The Kruskal-Wallis test was used to compare the distributions of the multiomic markers between the study groups. P values in bold font are statistically significant. VST, variance stabilizing transformation gene expression; FBOX6, F-Box Protein 6; MMP28, Matrix Metalloproteinase 28; PGDM, Prostaglandin D Metabolite; 11dTxB2, 11-dehydrothromboxane B2; LTE4, Leukotriene E4.

Table S4: Packages used for statistical analyses, Related to STAR Methods.

Objective	R package	Version	Reference
Identify differentially expressed genes	DESeq2	R 4.2.2	1
Feature selection analysis using machine learning	Random Forest	R 4.2.2	2
Cross-validation	Caret	R 4.2.2	3
Heatmap	ComplexHeatmap	R 4.2.2	4
Spearman correlations	Hmisc	R 4.2.2	5

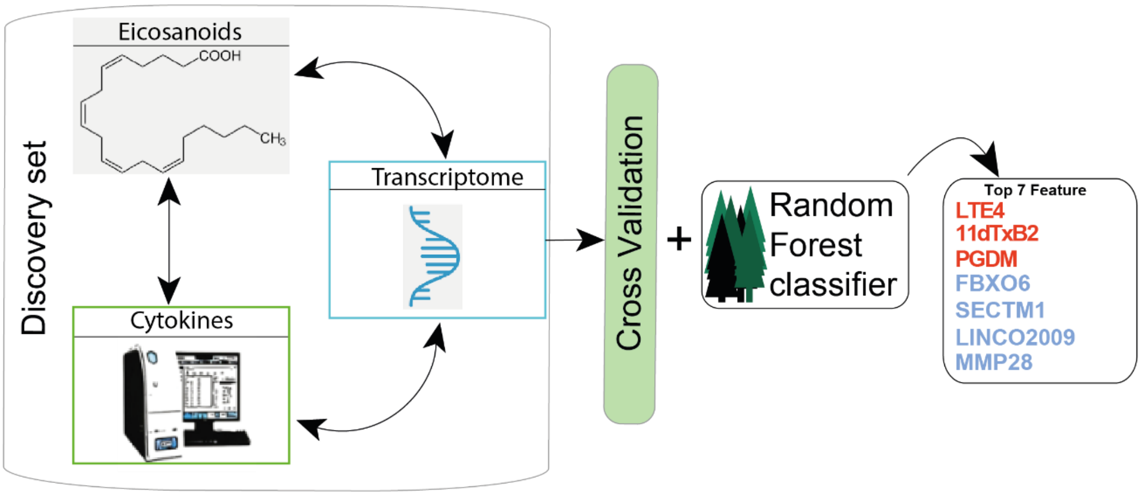

Figure S1. Dimensional data reduction, Related to Figure 1. A graphical abstract from the dimensionality reduction approach: A random forest model was applied to the multiplatform data (Luminex and RNAseq from peripheral blood and eicosanoids from urine) for feature selection based on clinical groups.

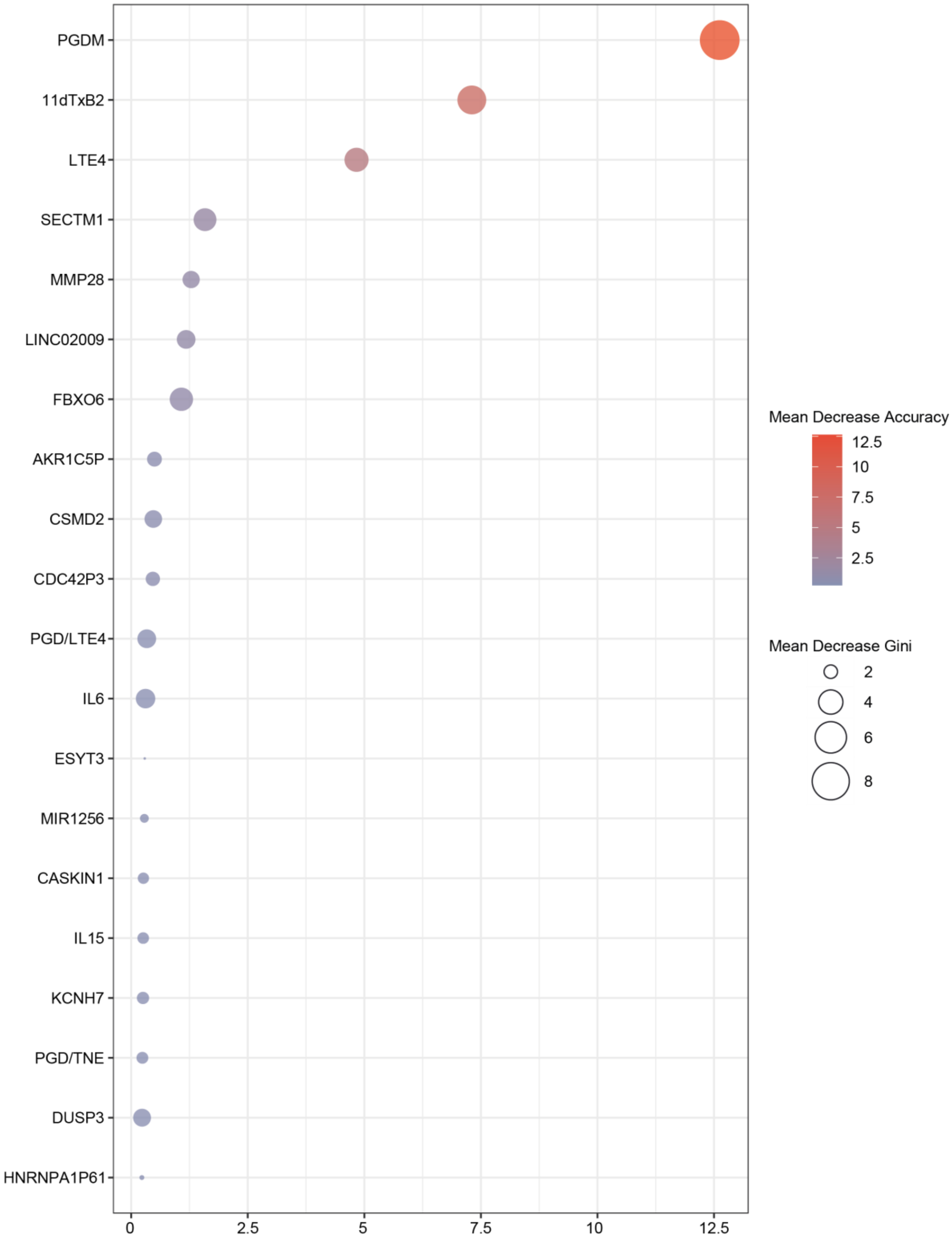

Figure S2. Feature importance of markers, Related to Figure 1. A dot plot to show the feature importance of the top 20 markers used in the randomForest model.

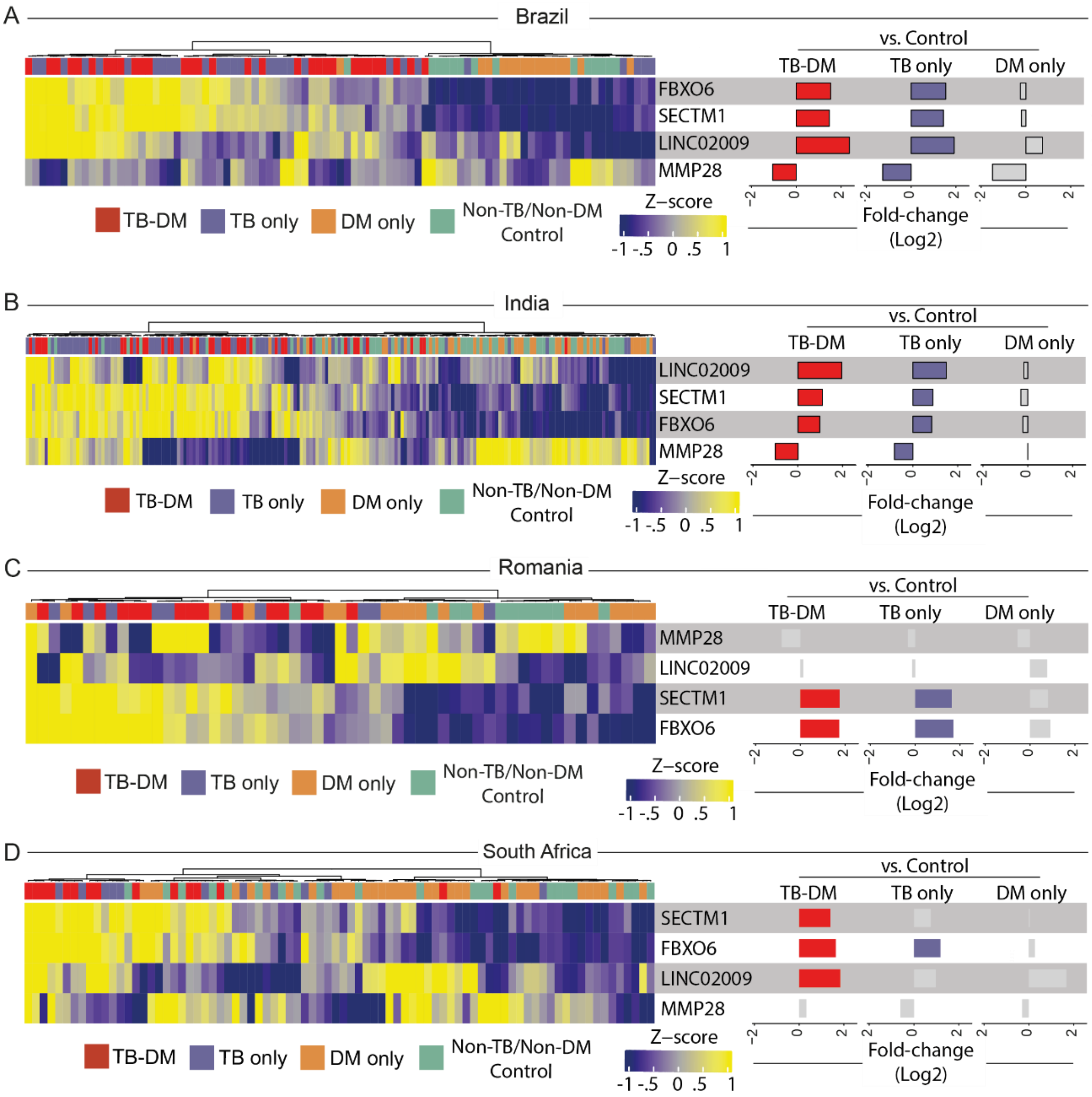

Figure S3. A partial transcriptomic signature to detect TB-DM, Related to Figure 4. A z-score normalized heatmap was employed to depict overall trends in gene expression among the clinical groups on each study site, as indicated. Panels to the right of heatmaps show the average fold-difference between the signature gene expression in the HC group versus TB-DM, TB only and DM only (log-transformed values).

References

1. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550. 10.1186/s13059-014-0550-8.
2. Breiman, L. (2001). Random Forest. Machine Learning. 10.1023/A:1010933404324.
3. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software *28*, 26. 10.18637/jss.v028.i05.
4. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847-2849. 10.1093/bioinformatics/btw313.
5. Harrell, Frank E. Hmisc: A package of miscellaneous R functions. [Internet]. Available from: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://cran.rproject.org/web/packages/Hmisc/Hmisc.pdf