Investigating Random Forest Classification on Publicly Available Tuberculosis Data to Uncover Robust Transcriptional Biomarkers

Carly A. Bobak[1,2], Alexander J. Titus[1,3] and Jane E. Hill[1,2]
[1]*Program in Quantitative Biomedical Sciences, Dartmouth School of Graduate and Advanced Studies, Hanover, NH, U.S.A.*
[2]*Thayer School of Engineering, Dartmouth School of Graduate and Advanced Studies, Hanover, NH, U.S.A.*
[3]*Department of Epidemiology, Dartmouth Geisel School of Medicine, Hanover, NH, U.S.A.*

Keywords: Tuberculosis, Random Forest, Machine Learning, Transcriptional Signatures, Data Integration.

Abstract: There has been increasing concern amongst the scientific community of a reproducibility crisis, particularly in the field of bioinformatics. Often, published research results do not correlate with clinical success. One theory explaining this phenomenon is that findings from homogeneous cohort studies are not generalizable to an inherently heterogeneous population. In this work, we integrate data from 4 distinct tuberculosis (TB) cohorts, for a total of 1164 samples, to find common differentially regulated genes which may be used to diagnose active TB from latent TB, treated TB, other diseases, and healthy controls. We selected 25 genes using random forest to get an AUC of 0.89 in our training data, and 0.86 in our test data. A total of 18 out of 25 genes had been previously associated with TB in independent studies, suggesting that integrating data may be an important tool for increasing micro-array research reproducibility.

1 INTRODUCTION

Reproducibility of research findings is paramount to scientific endeavors. However, there is increasing concern in the scientific community that published research findings are frequently irreproducible (Goodman et al., 2016). In fact, in 2016 a survey of 1576 scientific researchers found that 90% of respondents believed there is some degree of a reproducibility crisis, with 52% believing this crisis is 'significant' (Baker, 2016). Biomedical research is not immune to this crisis (Goodman et al., 2016). The National Institute of Health (NIH) noted the failure of many biomedical studies to present reproducible findings, and is leading a variety of interventions to ameliorate this phenomenon (Collins and Tabak, 2014).

Many factors have been implicated in contributing to the reproducibility crisis. An investigation into whether false findings represent the majority of scientific research identified that bias is often introduced in experimental design, data collection, and analysis. The study concluded that scientific results need to be externally validated from many distinct research groups before the findings can be considered truth (Ioannidis, 2005). Other studies have implicated imperfect animal and cell models as causes to the low correlation between research findings and clinical success (Begley and Ellis, 2012; Mestas and Hughes, 2004; Sweeney et al., 2016b). The emphasis on achieving statistical significance has been criticized in the literature as well (Sweeney et al., 2016b; Nuzzo, 2014).

While careful experimental design and increased emphasis on external validation play important roles in increasing the reproducibility of published research findings, other solutions include sharing research data (Collins and Tabak, 2014). A recent editorial in *Nature* made recommendations for "improv(ing) the transparency and reproducibility of research by means of data accessibility" (Nature, 2017).

The Khatri Lab of Stanford University has taken this notion one step further; they've proposed a framework for meta-analysis using data from publicly available resources such as the NIH Gene Expression Omnibus (GEO) (Sweeney et al., 2016a). Khatri argues that part of the reason why clinical studies cannot recapitulate biomarkers from research is due to the inevitable heterogeneity in actual populations. By embracing such heterogeneity, we can search for biomarkers which are present above the noise, and that these markers should be more robust in clinical settings (Sweeney et al., 2016a; Titus et al., 2017). Thus, by investigating biomarkers which have discriminatory potential across many studies, the irrepro-

Bobak C., Titus A. and Hill J.
Investigating Random Forest Classification on Publicly Available Tuberculosis Data to Uncover Robust Transcriptional Biomarkers.
DOI: 10.5220/0006752406950701
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF 2018)*, pages 695-701
ISBN: 978-989-758-281-3

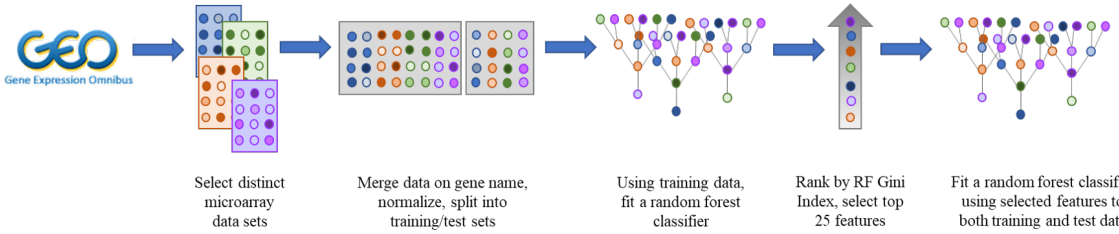


Figure 1: A brief overview of the data analysis methodology.

ducibility due to study specific observations is mitigated. This meta-analysis framework has been successfully applied on a variety of clinical settings, including tuberculosis (Sweeney et al., 2016a).

Tuberculosis is a top 10 cause of death worldwide, partly attributable to a significant gap in TB diagnosis (WHO, 2016). In 2015, only 6.1 million cases of TB were reported to the WHO, leaving a 4.3 million gap between incident and reported cases. Moreover, only 57% of cases were bacterially-confirmed, most other cases were clinically diagnosed given the symptoms presented (WHO, 2016). As such, there has been significant scientific interest in alternative diagnostic tools for TB, particularly in the realm of transcriptional biomarkers, to allow for faster diagnosis and treatment (Sweeney et al., 2016a; Prada-Medina et al., 2017; Sambarey et al., 2017).

In this work, we use Khatri's idea of integrating publicly shared datasets from previous TB research to identify a transcriptional signature which is robust to variations in study population. Instead of applying the Khatri meta-analysis framework, we investigate the potential of random forest to classify patients with active TB from healthy controls, latent TB, and other diseases. We hypothesize that by integrating publicly available transcriptomic data sets and using machine learning algorithms, we will be able fully leverage data from individual samples in such a way that we can capture more nuanced patterns for feature selection. These patterns may provide important insight into the pathogenesis of TB. The work presented here is a proof-of-concept exercise and the results will be used to guide the design of a more comprehensive study.

2 METHODS

Figure 1 gives an overview of the data analysis methodology. Data was collected from the NIH GEO depository (https://www.ncbi.nlm.nih.gov/gds) using both 'tuberculosis' and 'TB' as search terms, and limited to studies using human subjects that collected expression array data. Eligibility for inclusion required that the dataset include healthy controls, have at least 100 samples, and was from distinct institutions. The studies accompanying these datasets can be found in (Firszt and Vickery, 2011; Maertzdorf et al., 2011; Bloom et al., 2013; Blankley et al., 2016). A brief summary of the samples available in these datasets is presented in Table 1. The datasets included patients aged 16-87 from from the UK, South Africa, The Gambia, and France. This particular group of data does not include any HIV+ patients. The other diseases included in these studies are Streptococcal Pharyngitis, Staphylococcus infection, Still's disease, Systemic Lupus Erythematosus, Sarcoidosis, Pneumonia, and Lung Cancer. In total, 1164 samples are included in this analysis.

For the purposes of developing a diagnostic transcriptional signature for TB, a "1 vs. The Rest" algorithm was trained to separate active TB from latency, healthy controls, treated TB, and other diseases. A binary classification model (active TB vs. the rest) is most relevant in a clinical context.

All expression and phenotype sets were downloaded to R using the 'MetaIntegrator' package (Haynes et al., 2016). Expression sets were checked to ensure that they had been log_2 transformed, but were otherwise used as deposited. GSE19491, GSE42834, and GSE83456 were all analyzed using Illumina systems, but GSE28623 was analyzed using an Agilent system. Integrating datasets requires a unique identifier which can be used to match information from each set. Since data could not be merged on Probe ID, we first matched probes to their gene annotation, calculated the median expression value for

Table 1: Dataset by number of available samples. "LTB" denotes latent TB, "treated" denotes treated TB, "HC" denotes healthy controls, and "OD" denotes other diseases.

GSE	Number of Samples				
	TB	LTB	Treated	HC	OD
19491	89	69	14	133	193
28623	46	25	-	37	-
42834	65	-	-	143	148
83456	92	-	-	61	49

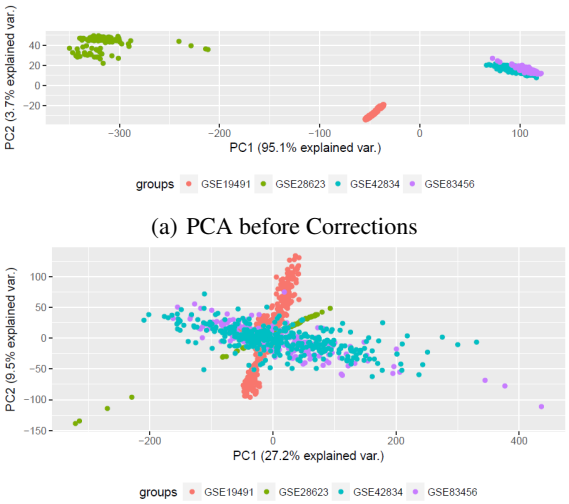
each gene, and then combined datasets based on the unique gene annotation.

To adjust for batch effect, COmbat CO-Normalization Using conTrols (COCONUT) was applied to the data. This method was implemented in R using the 'COCONUT' package (Sweeney et al., 2016c). COCONUT normalizes the data in an unbiased way while maintaining the distribution of genes both within and between studies. To achieve this, control samples are normalized using ComBat empiric Bayes normalization method and the parameters from the control samples are then applied to the diseased components.

A random forest model was fit using the 'randomForest' package in R (Liaw and Wiener, 2002). Random forest algorithms generate many classification trees, using randomly selected samples. A bootstrapped sample of the data is selected, and each split of the tree considers a random subset of candidate variables. Features are selected based on which variables best divide the data according to class at each split. By averaging over many classification trees, the method demonstrates low bias and variance. This method has proven to be particularly resilient for the classification of microarray data (Diaz-Uriarte and de Andres, 2006).

The data was split into a training ($\frac{2}{3}$) and test set ($\frac{1}{3}$) at random. Five hundred trees were used, at which point the classification error had stabilized. We chose to use 25 genes as our cut off for initial analysis and comparison to the current literature. While previous methods have emphasized using a small gene signature, we intend to examine a larger set to drive hypotheses regarding TB pathogenesis in future work. The top 25 genes were selected from this model using the Gini Impurity Index. We then refit a model using only those features on our training data, and used the refit model to predict classification on our test set.

To assess model performance, Area Under the Receiving Operator Characteristic (AUROC) curves are examined, as well as sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) of the resulting model. A heat map with hierarchical clustering using Jaccard's index, a distance metric which has demonstrated to be robust to noise, is visualized to assess unsupervised clustering among the selected features. (Toldo and Fusiello, 2008). Second, an unsupervised dimensionality reduction using a t-distributed Stochastic Nearest Estimate (t-SNE) is conducted, which allows for a non-parametric and non-linear mapping of the features to a reduced dimensional latent space (Maaten and Hinton, 2008). While features were selected using a "One vs. the Rest" algorithm, the unsupervised clustering methods are coloured using all categories in order to better visualize how the selected features discriminate between active TB and each subcategory.



(a) PCA before Corrections

(b) PCA after Corrections

Figure 2: Applying COCONUT to adjust for batch affect.

3 RESULTS

RNA was analyzed using three different platforms: Illumina HumanHT-12 V3.0 expression beadchip, llumina HumanHT-12 V4.0 expression beadchip, Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Feature Number version). This, in conjunction with other within study similarities (such as differences in normalization techniques between studies) introduced considerable batch effect with the merged data. This is demonstrated using a PCA in Figure 2. Note that data sets GSE42834 and GSE83456 both were analyzed using llumina HumanHT-12 V4.0 expression beadchip, and hence are clustered together. Data was adjusted using COCONUT, the results of which are shown in Figure 2. Note that while more of our data overlaps, we can still see distinct projections from each unique platform.

The top 25 genes returned from the initial random forest model, ordered by ranked importance, are: FCGR1A, GBP6, ANKRD22, VAMP5, C1QB, GBP2, FOXO1, ANKRD46, MEF2D, FCR1B, WDFY1, ETV7, PSME2, TXNDC12, BATF2, GBP5, TAP1, BLK, ZNF395, SOCS1, ICOS, PSMB9, SERTAD2, CTRL, and AIM2.

The Out-Of-Bag (OOB) error from the model fit with the 25 selected genes is $\sim$12%, with a total misclassification rate of $\sim$8%. The model has specificity of $\sim$94%. However, sensitivity was low, at a $\sim$69%.

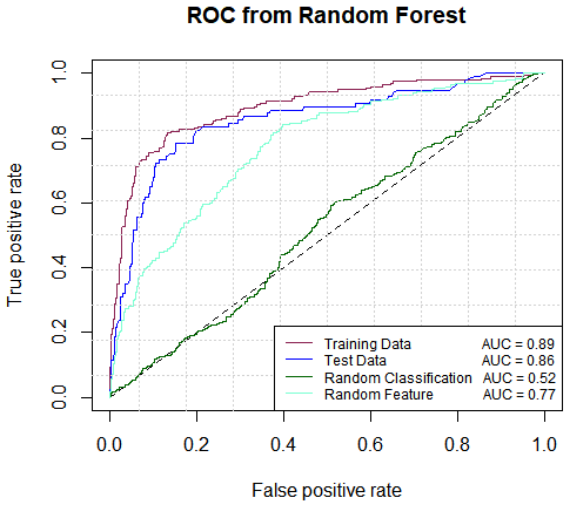


Figure 3: ROC curves from random forest model classifying active TB vs. the rest.

The positive predictive value (PPV) and negative predictive value (NPV) are ∼80% and ∼90% respectively. The AUC for the training set is ∼0.89 (95% CI:0.87-0.92), and ∼0.86 (0.81-0.90) for the test set.

The ROC curves from this model are shown in Figure 3. For reference to a null model, the ROC curves from randomly assigning our cases and controls as well as from fitting a random forest model on a randomly selected set of 25 genes are also shown. The AUCs from these null models are ∼ 0.52 (0.49-0.57) and ∼ 0.77 (0.74-0.81) respectively.

Figure 4 shows a heat map with hierarchical clustering using the selected features. Clustering still predominantly occurs based on study, but given the results of Figure 2 this is unsurprising. However, within studies, active TB consistently clusters away from healthy controls, with some mixing between active TB and latent TB, as well as active TB and other diseases.

The data was projected into a 3-D space using a t-SNE embedding, the results of which are visualized in Figure 5. Similar to Figure 4, Figure 5 shows clustering of active TB with separate clusters of TB related to original study. As well, Figure 5 demonstrates mixing between active TB cases and other diseases, highlighting that distinguishing these two classes remains a complicated problem to be tackled in future work.

4 DISCUSSION

Our results demonstrate that data integration across heterogenous studies is possible, and may lead to identification of important biomarkers which are consistent in patients with active TB. Unsurprisingly, classifying TB from other diseases and LTB is more difficult. The Khatri lab has shown that including additional data sets improves the classification accuracy of their models (Sweeney et al., 2016b). Similarly, further inclusion of more studies, particularly those with additional LTB and other disease samples, may improve classification accuracy of machine learning models built on integrated data. Future model iterations will consider weighting a classification penalty higher for misclassifications between TB and LTB, or TB and other diseases. Balancing between cases and controls may also improve classification accuracy, as cross validation techniques and more sophisticated feature selection.

The results from our unsupervised heat map and t-SNE suggest that despite using COCONUT, the integrated data clusters by study to some degree. Future work will investigate other methods for accounting for study-specific batch effect. As well, integrating raw data where applicable may negate the effect of some study-specific similarities.

The current analysis presented here has some limitations. In order to use COCONUT, healthy controls need to be included in the dataset. This limitation prevented the inclusion of TB datasets collected on youths and infants, as well as those which include HIV+ co-infection. There is some suggestion that instead of viewing datasets as distinct batches, each distinct platform could be considered a batch. This would allow the inclusion of datasets without healthy controls, as long as another dataset with healthy controls has been processed on the same platform (Sweeney et al., 2016c).

The Khatri meta-analysis framework uses a DerSimonian Laird random effects model in order to

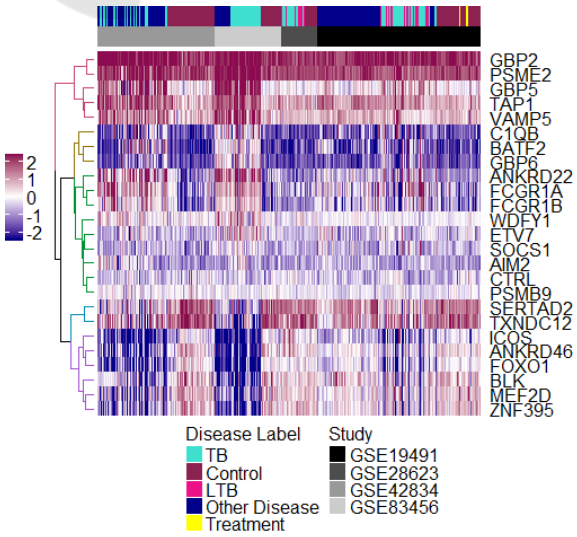


Figure 4: Heat map using selected genes.

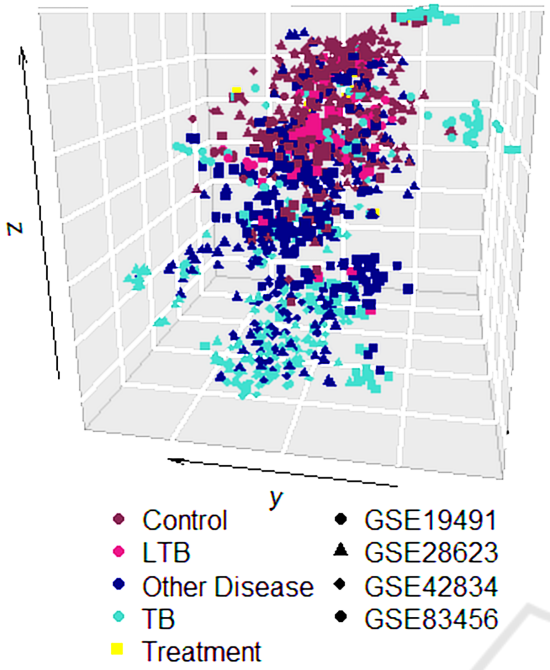


Figure 5: a t-SNE of the selected gene features.

compare gene expression from multiple studies which extracted total RNA from whole blood (Sweeney et al., 2016b; Sweeney et al., 2016a). Genes are identified in a discovery set, and later tested against independent validation sets. In order to reduce the number of features, genes are first filtered based on presence in all studies in the discovery set, FDR adjusted p-value, and effect size. They then reduce this set further by using a greedy forward search to maximize AUC, and built a gene score based on a geometric mean (Sweeney et al., 2016b; Sweeney et al., 2016a).

While this method has shown promising results, we hypothesized that model performance can be improved by employing other data analysis machinery. The pre-filtering step of the Khatri meta-analysis framework may exclude genes with small but important differences for discrimination between active TB and other classes. Moreover, the Khatri meta-analysis framework emphasizes selecting a small number of genes which limits the ability to explore biological mechanisms through pathway analysis.

Random forest models have a variety of strengths which lend them well to biomarker discovery from microarray data. For instance, random forest algorithms are robust to feature sets where there are many more potential predictor variables than there are outcomes (ie, $p \gg n$). As well, random forest incorporates interactions between predictor variables, which can often be a concern with microarray data. Perhaps most importantly, random forest models generate a ranked list of important features which can illuminate biological significance of features, but can also be leveraged for feature selection (Diaz-Uriarte and de Andres, 2006).

The reported mean AUC across all datasets in Khatri's model was 0.9 as compared to our mean AUC of 0.89 in our training data, and 0.86 in our test data (Sweeney et al., 2016a). Both models had lower accuracy when classifying active TB cases from other diseases, suggesting more data may be needed to identify TB specific gene responses.

In comparison to the results Khatri's work presented in (Sweeney et al., 2016a), both models selected GBP5. Two datasets assessed here overlap with those in Khatri's discovery set (used for feature selection). However, the Khatri's analysis included patients who were HIV+ as well as very young children (Sweeney et al., 2016a). One dataset he used did not include healthy controls, and hence was not eligible for this analysis. This may, in part, explain why out selected features excluded his findings of DUSP3 and KLF2 (Sweeney et al., 2016a). Future work will aim to include infants, youths, and HIV+ cases as part of the analysis. As well, future models will be tested against independent validation sets in order to demonstrate generalizability of the transcription signature.

Notably, of the 25 selected genes, 18 have been previously linked to TB in other studies. FCGR1A was associated with TB in (Prada-Medina et al., 2017; Jenum et al., 2016), GBP6 in (Kim et al., 2011), ANKRD22 in (Matsumiya et al., 2014), VAMP5 in (Sambarey et al., 2013), C1QB in (Cai et al., 2014; Sambarey et al., 2017), GBP2 in (Sambarey et al., 2017), FOXO1 in (Liu et al., 2013; Lu and Huang, 2011), FCGR1B (Prada-Medina et al., 2017), ETV7 in (Matsumiya et al., 2014), PSME2 in (Maji et al., 2015), BATF2 in (Prada-Medina et al., 2017), GBP5 in (Matsumiya et al., 2014; Sweeney et al., 2016a), TAP1 in (Fang et al., 2017), ZNF395 in (Matsumiya et al., 2014), SOCS1 in (Masood et al., 2012), ICOS in (Moguche et al., 2015), PSMB9 in (Sambarey et al., 2017), and AIM2 in (Prada-Medina et al., 2017). The remaining genes will be investigated as potential new findings for future TB research.

Importantly, all these studies are distinct from the data used to obtain our results. In a single integrated model, we have managed to reproduce at least partial results from 14 discrete studies. Not only does this address the issue of reproducibility in expression array studies of TB, but it is our belief that these results will add to the validity of current TB knowledge. Many of theses genes presented here have been previously implicated in adaptive immunity, and further pathway analysis may illuminate the biological mechanisms present in TB infection.

5 CONCLUSIONS

The initial analysis of integrated data shown here provides evidence that feature selection and model training based on heterogeneous integrated datasets is a potential tool to address the reproducibility crisis of array expression experiments. We intend to thoroughly investigate a variety of classification techniques to explore the possibility of using data integration to develop robust disease biomarkers. Combining datasets in this way should ameliorate much of the reproducibility problem in diagnostic research, and lead to a greater correlation between academic research and clinical success. It is our hope that this direction in research will not only lead to future diagnostic development, but to advancements in drug and vaccine development as well.

ACKNOWLEDGEMENTS

Dartmouth College holds an Institutional Program Unifying Population and Laboratory Based Sciences award from the Burroughs Wellcome Fund, and C. Bobak was supported by this grant (Grant#1014106). A. Titus was supported by the Office of the U.S. Director of the National Institutes of Health under award number T32LM012204. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454.

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533.

Blankley, S., Graham, C. M., Turner, J., Berry, M. P. R., Bloom, C. I., Xu, Z., Pascual, V., Banchereau, J., Chaussabel, D., Breen, R., Santis, G., Blankenship, D. M., Lipman, M., and O'Garra, A. (2016). The transcriptional signature of active tuberculosis reflects symptom status in extra-pulmonary and pulmonary tuberculosis. *PLOS ONE*, 11(10):e0162220.

Bloom, C. I., Graham, C. M., Berry, M. P. R., Rozakeas, F., Redford, P. S., Wang, Y., Xu, Z., Wilkinson, K. A., Wilkinson, R. J., Kendrick, Y., Devouassoux, G., Ferry, T., Miyara, M., Bouvry, D., Dominique, V., Gorochov, G., Blankenship, D., Saadatian, M., Vanhems, P., Beynon, H., Vancheeswaran, R., Wickremasinghe, M., Chaussabel, D., Banchereau, J., Pascual, V., pei Ho, L., Lipman, M., and O'Garra, A. (2013). Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PLoS ONE*, 8(8):e70630.

Cai, Y., Yang, Q., Tang, Y., Zhang, M., Liu, H., Zhang, G., Deng, Q., Huang, J., Gao, Z., Zhou, B., Feng, C. G., and Chen, X. (2014). Increased complement c1q level marks active disease in human tuberculosis. *PLoS ONE*, 9(3):e92340.

Collins, F. S. and Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, 505(7485):612–613.

Diaz-Uriarte, R. and de Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3.

Fang, K., Liu, F., Wen, J., Liu, H., Xiao, S., and Li, X. (2017). Association of tap1 and tap2 polymorphisms with risk and prognosis of pediatric spinal tuberculosis. *INTERNATIONAL JOURNAL OF CLINICAL AND EXPERIMENTAL MEDICINE*, 10(3):5769–5777.

Firszt, R. and Vickery, B. (2011). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Pediatrics*, 128(Supplement 3):S145–S146.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12.

Haynes, W. A., Vallania, F., Liu, C., Bongen, E., Tomczak, A., Andres-Terre, M., Lofgren, S., Tam, A., Deisseroth, C. A., Li, M. D., Sweeney, T. E., and Khatri, P. (2016). Empowering multi-cohort gene expression analysis to increase reproducibility. *bioRxiv*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Jenum, S., Bakken, R., Dhanasekaran, S., Mukherjee, A., Lodha, R., Singh, S., Singh, V., Haks, M. C., Ottenhoff, T. H. M., Kabra, S. K., Doherty, T. M., Ritz, C., and Grewal, H. M. S. (2016). BLR1 and FCGR1a transcripts in peripheral blood associate with the extent of intrathoracic tuberculosis in children and predict treatment outcome. *Scientific Reports*, 6(1).

Kim, B.-H., Shenoy, A. R., Kumar, P., Das, R., Tiwari, S., and MacMicking, J. D. (2011). A family of IFN--inducible 65-kD GTPases protects against bacterial infection. *Science*, 332(6030):717–721.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Liu, Y., Jiang, J., Wang, X., Zhai, F., and Cheng, X. (2013). miR-582-5p is upregulated in patients with active tuberculosis and inhibits apoptosis of monocytes by targeting FOXO1. *PLoS ONE*, 8(10):e78381.

Lu, H. and Huang, H. (2011). FOXO1: A potential target for human diseases. *Current Drug Targets*, 12(9):1235–1244.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Maertzdorf, J., Ota, M., Repsilber, D., Mollenkopf, H. J., Weiner, J., Hill, P. C., and Kaufmann, S. H. E. (2011). Functional correlations of pathogenesis-driven gene

expression signatures in tuberculosis. *PLoS ONE*, 6(10):e26938.

Maji, A., Misra, R., Mondal, A. K., Kumar, D., Bajaj, D., Singhal, A., Arora, G., Bhaduri, A., Sajid, A., Bhatia, S., Singh, S., Singh, H., Rao, V., Dash, D., Shalini, E. B., Michael, J. S., Chaudhary, A., Gokhale, R. S., and Singh, Y. (2015). Expression profiling of lymph nodes in tuberculosis patients reveal inflammatory milieu at site of infection. *Scientific Reports*, 5(1).

Masood, K. I., Rottenberg, M. E., Carow, B., Rao, N., Ashraf, M., Hussain, R., and Hasan, Z. (2012). SOCS1 gene expression is increased in severe pulmonary tuberculosis. *Scandinavian Journal of Immunology*, 76(4):398–404.

Matsumiya, M., Harris, S. A., Satti, I., Stockdale, L., Tanner, R., O'Shea, M. K., Tameris, M., Mahomed, H., Hatherill, M., Scriba, T. J., Hanekom, W. A., McShane, H., and Fletcher, H. A. (2014). Inflammatory and myeloid-associated gene expression before and one day after infant vaccination with MVA85a correlates with induction of a t cell response. *BMC Infectious Diseases*, 14(1).

Mestas, J. and Hughes, C. C. W. (2004). Of mice and not men: Differences between mouse and human immunology. *The Journal of Immunology*, 172(5):2731–2738.

Moguche, A. O., Shafiani, S., Clemons, C., Larson, R. P., Dinh, C., Higdon, L. E., Cambier, C., Sissons, J. R., Gallegos, A. M., Fink, P. J., and Urdahl, K. B. (2015). ICOS and bcl6-dependent pathways maintain a CD4 t cell population with memory-like properties during tuberculosis. *The Journal of Experimental Medicine*, 212(5):715–728.

Nature (2017). Empty rhetoric over data sharing slows science. *Nature*, 546(7658):327–327.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506(7487):150–152.

Prada-Medina, C. A., Fukutani, K. F., Kumar, N. P., Gil-Santana, L., Babu, S., Lichtenstein, F., West, K., Sivakumar, S., Menon, P. A., Viswanathan, V., Andrade, B. B., Nakaya, H. I., and Kornfeld, H. (2017). Systems immunology of diabetes-tuberculosis comorbidity reveals signatures of disease complications. *Scientific Reports*, 7(1).

Sambarey, A., Devaprasad, A., Baloni, P., Mishra, M., Mohan, A., Tyagi, P., Singh, A., Akshata, J., Sultana, R., Buggi, S., and Chandra, N. (2017). Meta-analysis of host response networks identifies a common core in tuberculosis. *npj Systems Biology and Applications*, 3(1).

Sambarey, A., Prashanthi, K., and Chandra, N. (2013). Mining large-scale response networks reveals 'topmost activities' in mycobacterium tuberculosis infection. *Scientific Reports*, 3(1).

Sweeney, T. E., Braviak, L., Tato, C. M., and Khatri, P. (2016a). Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. *The Lancet Respiratory Medicine*, 4(3):213–224.

Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P., and Khatri, P. (2016b). Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Research*, 45(1):e1–e1.

Sweeney, T. E., MD, and PhD (2016c). *COCONUT: COmbat CO-Normalization Using conTrols (COCONUT)*. R package version 1.0.1.

Titus, A. J., Way, G. P., Johnson, K. C., and Christensen, B. C. (2017). Deconvolution of DNA methylation identifies differentially methylated gene regions on 1p36 across breast cancer subtypes. *bioRxiv*.

Toldo, R. and Fusiello, A. (2008). Robust multiple structures estimation with j-linkage. In *Lecture Notes in Computer Science*, pages 537–547. Springer Berlin Heidelberg.

WHO (2016). Global tuberculosis report 2016. Technical report, World Health Organization.