Assessment of Imputation Methods for Missing Gene Expression Data in Meta-Analysis of Distinct Cohorts of Tuberculosis Patients

Carly A. Bobak[†], Lauren McDonnell, Matthew D. Nemesure, Justin Lin, and Jane E. Hill[‡]

Thayer School of Engineering, Dartmouth College,
Hanover, NH, USA
[†]*E-mail: Carly.A.Bobak.GR@dartmouth.edu*
[‡]*E-mail: Jane.E.Hill@dartmouth.edu*

The growth of publicly available repositories, such as the Gene Expression Omnibus, has allowed researchers to conduct meta-analysis of gene expression data across distinct cohorts. In this work, we assess eight imputation methods for their ability to impute gene expression data when values are missing across an entire cohort of Tuberculosis (TB) patients. We investigate how varying proportions of missing data (across 10%, 20%, and 30% of patient samples) influence the imputation results, and test for significantly differentially expressed genes and enriched pathways in patients with active TB. Our results indicate that truncating to common genes observed across cohorts, which is the current method used by researchers, results in the exclusion of important biology and suggest that LASSO and LLS imputation methodologies can reasonably impute genes across cohorts when total missingness rates are below 20%.

Keywords: meta-analysis; imputation; multi-cohort analysis; cohort-wide imputation.

1. Introduction

The importance and availability of biological data repositories continues to grow at an astounding rate.[1–5] At the time of writing, the Gene Expression Omnibus (GEO) houses 27,856 datasets with gene expression profiling by array in humans.[6] This wealth of data has allowed many researchers to use meta-analysis of gene expression as a tool for biological discovery and validation. Many methodologies have been proposed for efficient meta-analysis of gene expression array data, including conormalization of genes and workflows for best identifying gene signatures.[1–4] Currently, these analyses restrict to the common set of genes observed across all included studies which will directly impact downstream analyses.[5]

Imputation methods, such as Local Least Squares (LLS) and k-Nearest-Neighbors (k-NN), have been developed to recover missing gene expression data within a single study, however, published attempts to recover missing genes across an entire dataset have been limited.[5,7] Only one study examined imputing genes across two Affymetrix platforms where probe names of one platform were a subset of probe names from the other. This study demonstrated reasonable accuracy in imputing 9986 probes using the information which was common across both platforms with LASSO imputation.[5] Currently, methods have not examined imputing genes across cohorts from many different gene expression platforms.

Tuberculosis (TB), a disease caused by *Mycobacterium tuberculosis*, is a global health crisis. In 2017, there were 10 million reported cases of active TB and an estimated 1.3 million deaths from the disease.[8] In 2014, the WHO developed the Sustainable Development Goals which included 'End TB', a program aiming to eradicate the TB epidemic by 2030.[9] Accomplishing this goal requires the development of improved diagnostics that are less invasive and more reliable. Multiple studies have proposed using a 'gene signature' for the diagnosis of TB using gene expression from whole blood.[2–4,10–20] These signatures seek to summarize TB specific host immune responses.[21] While many gene signatures for the diagnosis of active TB have been proposed, currently none have been approved for use in the clinic and a point-of-care diagnostic.[8,20] This has led some researchers to examine a meta-analysis method, where publicly available data across many cohorts of patients is used for the discovery of a gene signature to diagnose TB.[1–4]

We posit that important biology is missed when studies integrating gene sets only focus on the common gene set and show, given some guidelines, that it is possible to impute genes across an entire dataset with reasonable accuracy. Our analysis targets gene expression data from the host's response to TB. We identify 20 datasets from GEO which evaluate gene expression in human blood of patients with TB disease, and merge these sets using gene symbols common across all studies. We then examine the components of 16 published gene signatures for active TB for their presence or absense in the merged dataset. We compare eight gene expression imputation strategies on the ability to impute genes across entire cohorts, and evaluate the impact of the imputed genes on the biological relevance through differential expression (DE) analysis and pathway enrichment.

2. Methods

All analyses were completed using R 3.5.0 (R Core Team, Vienna, Australia).

2.1. *Data Sources*

Gene expression array data was downloaded from the Gene Expression Omnibus (GEO)[6] database with accession numbers listed in Table 1. Search terms to identify datasets included 'Tuberculosis' and 'TB', and results were limited to studies in human subjects examining gene expression in whole blood. This multi-cohort analysis consists of 20 distinct datasets across 30 different countries and includes patients with active TB, latent TB infections (LTBI), other diseases, as well as healthy controls and treated patients (includes all individuals treated for TB, LTBI and other diseases). In total, the dataset includes 3,096 participants.

2.2. *Data Processing*

Data was processed following previous meta-analyses for gene expression array data.[1–4] Where possible, supplementary files were downloaded from GEO. Raw affy values were normalized using GCRMA.[36] All probe names were converted to gene symbols using the most up-to-date annotation packages from Bioconductor.[37] Where multiple probes mapped to a single gene

Table 1. Summary of all samples included in the study.

GSE	Number of Samples					
	Healthy Controls	**LTBI**	**Other Diseases**	**TB**	**Treated**	**Total**
GSE116014[22]	24	58	-	-	-	82
GSE19491[10]	117	69	193	61	14	454
GSE28623[23]	37	25	-	46	-	108
GSE31348[24]	-	-	-	27	108	135
GSE34608[25]	18	-	18	8	-	44
GSE36238[26]	-	-	-	9	9	18
GSE37250[11]	-	167	175	195	-	537
GSE39939[12]	-	14	108	35	-	157
GSE39940[12]	-	54	169	111	-	334
GSE40553[27]	-	-	-	29	175	204
GSE41055[13]	9	9	-	9	-	27
GSE42834[14]	118	-	108	40	15	281
GSE56153[28]	18	-	-	18	35	71
GSE58411[29]	-	-	-	31	76	107
GSE62147[30]	-	-	-	52	-	52
GSE69581[31]	-	25	10	15	-	50
GSE73408[32]	-	35	39	35	-	109
GSE81746[33]	2	1	-	5	-	8
GSE83456[34]	61	-	49	92	-	202
GSE83892[35]	-	-	17	99	-	116
Total	**404**	**457**	**886**	**917**	**432**	**3096**

symbol, the median expression value was used. Within each cohort, expression values were quantile-normalized, $\log_2$-transformed, mean-centered and scaled by standard deviation. Each dataset was then run through an alias converter in the '*limma*' package[38] ensure the most common alias for each gene was used prior to merging. Cohorts were merged based on gene symbol. To reduce batch effect and ensure gene distributions were common across all datasets, FSQN normalization was applied.[39]

2.3. *Generation of Missing Values*

The data has been processed according to the workflow shown in Figure 1. In short, the set of genes that are common to all patients in the full dataset were selected ('complete subset'). Similar to previous imputation studies, 5% of the genes in the complete subset were randomly selected to be masked as missing values. We considered three different cutoffs of the 'missingness rate', or the proportion of patients with missing genes, at 10%, 20%, and 30%.[7] Unlike previous studies, we selected patients across full cohorts. To do this, we iteratively selected a cohort at random and masked gene k across all patients in that cohort if the sample size was within 30 patients of the cutoff rate. The process was repeated for each of the selected missing genes. Datasets with masked genes are referred to as the 'incomplete subset'.

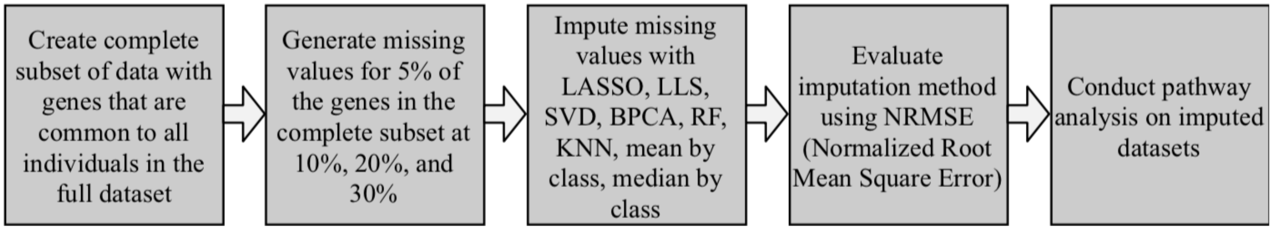
Create complete subset of data with genes that are common to all individuals in the full dataset
Generate missing values for 5% of the genes in the complete subset at 10%, 20%, and 30%
Impute missing values with LASSO, LLS, SVD, BPCA, RF, KNN, mean by class, median by class
Evaluate imputation method using NRMSE (Normalized Root Mean Square Error)
Conduct pathway analysis on imputed datasets

Fig. 1. Workflow for imputing missing gene expression data

2.4. *Evaluation of Biologically Relevant Missing Genes*

Through an extensive literature search, Warsinske et al previously identified 16 gene signatures designed to classify active TB from other clinical groups.[20] We evaluated the individual genes which comprise each of these gene signatures for their presence and absence in the common genes across all 20 datasets in Subsection 2.1. To reduce the number of missing biomarkers, we ran each member of the gene signature through the '*limma*' alias converter used on the merged multi-cohort data. Where probe names were reported with gene names in the 11 studies, we retrieved updated gene symbols using the Bioconductor annotation packages.[37]

2.5. *Imputation methods*

Eight frequently-used imputation methods were implemented on the incomplete subset: Least Absolute Shrinkage and Selection Operator (LASSO), Local Least Squares (LLS), Singular Value Decomposition (SVD), Bayesian Principal Component Analysis (BPCA), Random Forest, k-nearest-neighbors (k-NN), mean (by class) and median (by class).[5,40–44] LASSO imputation builds a LASSO regression model for each gene with missing values, selecting lambda via 10-fold cross validation.[5] RF imputation replaces missing genes with the mean value, then trains a random forest ($ntree$=500) to compute a proximity matrix. Then, using proximity as a weight, it re-imputes the missing values as the weighted average of the non-missing genes and iterates.[40] In k-NN, the k (k=10) nearest neighbors (Euclidean distance) for a missing gene are identified, and the weighted average of these neighbors is the imputed value.[41] LLS imputation selects k (k=10) genes using correlation structure (pearson) and imputes the missing values as a linear combination of the k genes using local least squares regression.[42] SVD imputation finds a low rank singular value decomposition (k=3096) and uses the most significant eigenvectors to linearly impute missing genes.[41] BPCA imputation works by building a probabilistic PCA using Bayes' theorem, and then imputes values using a Bayes estimation algorithm ($nPcs$=2).[43] Mean and median (by class) impute across cohorts by taking the mean or median values of the observed genes by the disease class of the missing patient.[44]

2.6. *Assessment of Imputation Performance*

We evaluated the performance of each imputation method using the Normalized Root Mean Square Error (NRMSE).

$$\text{NRMSE} = \sqrt{\frac{\text{mean}\big((y_{\text{original}} - y_{\text{imputed}})^2\big)}{\text{variance}(y_{\text{original}})}} \tag{1}$$

Here, y_{original} refers to the complete subset and y_{imputed} refers to the imputed subset. Smaller NRMSE values indicate better accuracy.[7] We also computed NRMSE for each imputed gene for finer granularity and assessment of the imputation methods.

2.7. *Differential Expression Analysis and Pathway Enrichment*

To evaluate the differences in downstream analysis across the complete subset, the incomplete subsets, and the imputed subsets, we used *'limma'* to conduct differential expression (DE) analysis. For this analysis we contrasted active TB with any other disease category. Genes which met a Benjamini-Hochberg adjusted p-value < 0.05 and an absolute log fold change (log FC) of at least 0.5 were considered significant. To compare the significantly DE genes, we examined the intersection and set difference across subsets. Lists of significantly DE genes and their subsequent log FC were then sent to the ConsensusPathDB to search for enriched pathways from KEGG, wikipathways, reactome, and SMPDB with an enrichment p-value < 0.01 and a minimum of at least 4 significant genes.[45] The intersection and set difference of pathways were examined and interpreted through literature review.

3. Results

3.1. *Evaluation of the common gene subset across previous TB signatures*

A total of 14,003 unique genes were common to all 20 datasets out of a a total of 22,283 possible genes. Previously, 16 unique gene signatures which classify active TB from other disease classes were identified, for a total of 607 genes associated with active TB disease.[2,10–20] Of those 607 genes, 132 were absent across the common gene subset. Hence, approximately 22% of genes previously proposed as being diagnostic for TB would be excluded from downstream analysis if we restrict to common genes observed across all datasets. The distribution of the missing 132 genes across all 3,096 patients is shown in Fig. 2. Most genes were missing in less than 5% of the total number of patients, indicating that much of this biology is imputable. Notably, only 22 of the 132 missing genes had been reported as part of a TB diagnostic signature in at least two studies. Of these 22 missing genes, 12 are missing in less than 10% of patients included in the multi-cohort, with 16 out of 22 missing in less than $\frac{1}{3}$ of samples. The remaining 6 genes are missing from more than 80% of the samples included in this analysis.

3.2. *Assessment of imputation performance*

To evaluate the impact of imputing genes across whole datasets at varying missingness rates, we used the NRMSE across the imputed and complete subset of genes. We also evaluated the NRMSE in each imputed gene compared to the true expression values in the complete subset. We report the minimum, median, mean, and maximum NRMSE across the imputed genes, and the full data NRMSE across the complete imputed dataset for each model and missingness rate in Fig. 3. We included a color scale varying from blue (small NRMSE) to red (high NRMSE) within each column of the table. The order of models presented in the table is indicative of the median NRMSE value within each missing proportion. When evaluating NRMSE at the imputed gene level, the LASSO and LLS models had the best overall performance, although

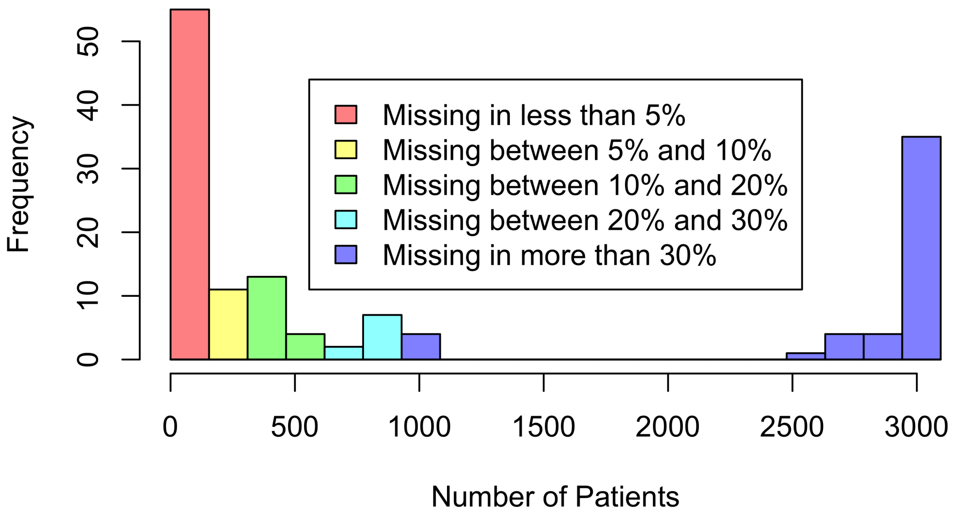

Fig. 2. The frequency of missing genes from previous TB signatures across the number of patients. Most previously reported genes were missing from less than 5% of total patients across the multi-cohort and only 6 genes missing in $> 80\%$ of patients were reported in more than one TB signature. Genes missing in > 2500 patients largely represent genes which were unnamed or open reading frames at the time of the original TB signature publication.

LASSO models took approximately a 1.5 days to impute while LLS imputed in less than 10 minutes.

3.3. *Impact on downstream analysis*

To investigate the impact on downstream analysis across incomplete and imputed datasets at varying levels of missing data, we used *'limma'* to identify significantly DE genes in each generated dataset compared to the complete cohort.[38] To better interpret the impact that differences in significantly DE genes in each dataset have on informing biological understanding of TB mechanisms, we enriched each set of DE genes into pathways using ConsensusPathDB ($p < 0.01$). Gene pathway enrichment can be used to evaluate which predefined biological processes are associated with gene sets of interest.[45] Pathway enrichment of the imputed datasets allows the comparison of how the imputation perturbs interpretation of gene expression results in terms of differences in biological processes. The results from both the DE gene and pathway enrichment analyses are shown in Table 2. Significant genes from the complete dataset were enriched for a total of 45 pathways under this analysis schema. Notably, the incomplete subset missed 11 pathways indicated as significant in the complete dataset, with most imputation methods recovering all but 2 pathways. Critically, one of the pathways missed by the incomplete dataset was 'The human immune response to tuberculosis', and this pathway was recovered all but one of the imputation models.

4. Discussion

Meta-analysis of gene expression data is becoming increasingly common as public repositories of biological data continue to grow.[1–5] While best practices for integrating and analyzing

Missing Rate	Imputation Model	Min NRMSE	Median NRMSE	Mean NRMSE	Max NRMSE	Full Data NRMSE
0.1	LASSO	0.056	0.288	0.275	0.407	0.058
	LLS	0.044	0.306	0.331	0.419	0.061
	BPCA	0.154	0.323	0.315	0.418	0.066
	KNN	0.151	0.325	0.318	0.414	0.066
	RF	0.216	0.326	0.323	0.412	0.067
	Mean	0.261	0.337	0.336	0.424	0.069
	Median	0.261	0.337	0.336	0.424	0.069
	SVD	0.261	0.338	0.337	0.431	0.069
0.2	LASSO	0.154	0.437	0.415	0.581	0.073
	LLS	0.202	0.454	0.437	0.574	0.076
	KNN	0.252	0.476	0.464	0.582	0.079
	BPCA	0.259	0.476	0.467	0.578	0.080
	RF	0.340	0.485	0.480	0.583	0.081
	Mean	0.429	0.500	0.500	0.586	0.084
	SVD	0.431	0.501	0.500	0.583	0.084
	Median	0.431	0.505	0.504	0.585	0.085
0.3	LASSO	0.234	0.552	0.535	0.673	0.080
	LLS	0.280	0.582	0.563	0.705	0.082
	KNN	0.363	0.617	0.603	0.720	0.086
	BPCA	0.377	0.622	0.606	0.748	0.086
	RF	0.469	0.634	0.627	0.739	0.088
	Mean	0.563	0.653	0.652	0.665	0.091
	SVD	0.560	0.655	0.654	0.747	0.091
	Median	0.572	0.657	0.672	0.781	0.092

maximum

median

minimum

Fig. 3. Heat table of the NRMSE values over the individually imputed genes. Color scale varies from lowest NRMSE in the column to highest NRMSE in the column. Single cohort studies report full data NRMSE values between ∼ 0.015 − 0.09.

biological data across distinct cohorts are debated, most researchers have elected to focus on the common subset of genes present across all studies. Here, we extend imputation concepts developed to address missing values of genes within studies to account for missingness across full cohorts of patients.

Approximately 22% of genes which had been previously included in a diagnostic signature for active TB were missing when analyses were restricted to common genes across all datasets, suggesting that important biology is being missed as a result of this restriction. Imputing reduces the impact of the missing genes substantially; 66 out of 132 genes are missing in less than 10% of patients, 83 genes in less than 20% of patients, and 92 genes in less than 30% of patients. Of the genes missing in more than 30% of patients, 28 out of 40 genes represent an open reading frame or do not have an available gene symbol. Hence, in this analysis imputation allows for the recovery of most well-annotated genes associated with active TB.

While the overall NMRSE results examining the impact of existing imputation methods across full cohorts of data are generally reasonably small (approximately 0.06 to 0.1), they are higher than observed in similar studies which have examined imputation within a single cohort (approximately 0.015-0.09).[7] This result is to be expected; the multi-cohort data presented here is heterogeneous, encompassing patients from 30 countries and across 27 diseases. This

Table 2. A summary of the results from downstream analysis with each imputation scheme. All imputation strategies outperformed the incomplete subset in terms of preserving relevant TB biology, with best recovery of genes and pathways observed with LASSO.

Data Subsets	Imputation Model	Missingness Rate	Genes Overlap	Missing Genes	Extra Genes	Pathway Overlap	Missing Pathways	Extra Pathways	TB Pathway
Incomplete	-	0.1	281	17	-	34	11	-	-
		0.2	280	18	-	34	11	-	-
		0.3	281	17	-	34	11	-	-
Imputed	LASSO	0.1	297	1	-	45	-	3	Yes
		0.2	297	1	1	45	-	3	Yes
		0.3	298	-	-	45	-	3	Yes
Imputed	LLS	0.1	295	3	-	45	-	3	Yes
		0.2	295	3	1	45	2	2	Yes
		0.3	296	2	-	45	-	3	Yes
Imputed	KNN	0.1	295	5	2	43	2	3	Yes
		0.2	291	7	1	43	2	-	Yes
		0.3	294	4	1	37	8	-	-
Imputed	BPCA	0.1	294	4	-	43	2	3	Yes
		0.2	293	5	1	43	2	3	Yes
		0.3	294	4	-	42	3	2	Yes
Imputed	RF	0.1	293	5	-	43	2	3	Yes
		0.2	289	9	-	43	2	3	Yes
		0.3	293	5	-	43	2	3	Yes
Imputed	SVD	0.1	291	7	-	43	2	3	Yes
		0.2	286	12	-	42	3	3	Yes
		0.3	290	8	-	41	4	-	Yes
Imputed	Mean	0.1	293	5	2	43	2	3	Yes
		0.2	292	6	5	43	2	-	Yes
		0.3	296	2	4	45	2	-	Yes
Imputed	Median	0.1	293	5	2	43	2	3	Yes
		0.2	294	4	7	43	2	-	Yes
		0.3	296	2	8	45	-	-	Yes

heterogeneity likely makes imputation difficult.

Best performance in terms of NRMSE across full datasets and within imputed genes was observed from LASSO and LLS imputation models. Both methods utilize least squares regression models, perhaps indicating that these models are adept at imputation across cohorts.

Downstream analysis demonstrated all imputed methods outperform the incomplete data, regardless of missingness rate. The incomplete subset, representing the restricted analysis, missed between 17 and 18 DE genes and 11 enriched pathways. Across the different imputation schemas, on average 5 DE genes were absent compared to the complete subset (range between complete recovery and 12 missing DE genes). Most algorithms missed two pathways which were present in the complete subset, and found 3 pathways which were not. Across all imputation methods, there was remarkable consistency in the missed and extra pathways.

Of the 24 imputed data sets, 18 consistently missed the 'Pertussis - Homo sapiens (human)' and 'Complement and coagulation cascades - Homo sapiens (human)' pathways. Pertussis is a respiratory tract infection better known as whooping cough and may share some symptomatic similarities with TB.[46] The 'Complement and coagulation cascades' pathway is closely related to regulation of IFNγ and has previously been linked to TB progression.[47]

The same three extra pathways commonly were enriched with imputed data: 'Generic Transcription Pathway', 'RNA Polymerase II Transcription' and 'Gene expression (Transcription)'. All three of these pathways may be indicative of epigenetic changes or gene silencing occurring during infection. While these pathways did not reach the $p < 0.01$ significant threshold in the complete subset data, they were on the cusp of significance ($p = 0.0101$, $p = 0.0101$, and $p = 0.0127$ respectively). Hence, the inclusion of these pathways is indicative of small perturbations of gene values that were in the complete data subset and likely not a symptom of misleading biology.

4.1. *Limitations*

Given the nature of data missing across full cohorts, it is difficult to evaluate whether missing data is a result of missing (completely) at random (MAR or MCAR) or missing not at random (MNAR). We assume missing data is attributable to platform specificity or prior data analysis.

While LASSO and LLS methods had the strongest observed performance in this study, it is possible that other methods would have superior performance in other applications. Moreover, further study is needed to understand the bounds on the number of datasets with observed data across a gene in order to impute expression with reasonable accuracy, as well as the proportion of observed genes necessary to impute missing genes.

The mapping of probe sets to gene symbols is dynamic; as our understanding of biology deepens many definitions have changed.[5] For example, some reported gene names in TB signatures are missing from 100% of samples in the muti-cohort data. Recovering these genes requires the identification of the original probe names to link to an updated gene symbol.

5. Conclusion

Downstream analysis of imputed data indicates that imputed genes missing in up to 30% of patients did not drastically alter the significantly DE genes or enriched pathways. When no imputation scheme was used, 11 out of 45 pathways were missed compared to the full data. We urge researchers considering a meta-analysis of gene expression data carefully to examine the potential loss of information that occurs when restricting analysis to common genes present across all datasets and to instead consider imputation strategies.

6. 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).

References

1. J. J. Hughey and A. J. Butte, Robust meta-analysis of gene expression using the elastic net, *Nucleic acids research* **43** (2015).
2. T. E. Sweeney, L. Braviak, C. M. Tato and P. Khatri, Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis, *The Lancet Respiratory Medicine* **4** (2016).
3. C. A. Bobak, A. J. Titus and J. E. Hill, Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets, *Applied Soft Computing* **74** (2019).
4. C. A. Bobak, A. J. Titus and J. E. Hill, Investigating random forest classification on publicly available tuberculosis data to uncover robust transcriptional biomarkers., in *HEALTHINF*, 2018.
5. W. Zhou, L. Han and R. B. Altman, Imputing gene expression to maximize platform compatibility, *Bioinformatics* **33** (2016).
6. T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko *et al.*, Ncbi geo: archive for functional genomics data sets?update, *Nucleic acids research* **41** (2012).
7. R. Aghdam, T. Baghfalaki, P. Khosravi and E. S. Ansari, The ability of different imputation methods to preserve the significant genes and pathways in cancer, *Genomics, proteomics & bioinformatics* **15** (2017).
8. WHO, World health organization global tuberculosis report 2018 (2018).
9. A. Suthar, R. Zachariah and A. Harries, Ending tuberculosis by 2030: can we do it?, *The international journal of tuberculosis and lung disease* **20** (2016).
10. M. P. Berry, C. M. Graham, F. W. McNab, Z. Xu, S. A. Bloch, T. Oni, K. A. Wilkinson, R. Banchereau, J. Skinner, R. J. Wilkinson *et al.*, An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis, *Nature* **466** (2010).
11. M. Kaforou, V. J. Wright, T. Oni, N. French, S. T. Anderson, N. Bangani, C. M. Banwell, A. J. Brent, A. C. Crampin, H. M. Dockrell *et al.*, Detection of tuberculosis in hiv-infected and-uninfected african adults using whole blood rna expression signatures: a case-control study, *PLoS medicine* **10** (2013).
12. S. T. Anderson, M. Kaforou, A. J. Brent, V. J. Wright, C. M. Banwell, G. Chagaluka, A. C. Crampin, H. M. Dockrell, N. French, M. S. Hamilton *et al.*, Diagnosis of childhood tuberculosis and host rna expression in africa, *New England Journal of Medicine* **370** (2014).
13. L. M. Verhagen, A. Zomer, M. Maes, J. A. Villalba, B. Del Nogal, M. Eleveld, S. A. van Hijum, J. H. de Waard and P. W. Hermans, A predictive signature gene set for discriminating active from latent tuberculosis in warao amerindian children, *BMC genomics* **14** (2013).
14. C. I. Bloom, C. M. Graham, M. P. Berry, F. Rozakeas, P. S. Redford, Y. Wang, Z. Xu, K. A. Wilkinson, R. J. Wilkinson, Y. Kendrick *et al.*, Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers, *PloS one* **8** (2013).

15. L. L. da Costa, M. Delcroix, E. R. Dalla Costa, I. V. Prestes, M. Milano, S. S. Francis, G. Unis, D. R. Silva, L. W. Riley and M. L. Rossetti, A real-time pcr signature to discriminate between tuberculosis and other pulmonary diseases, *Tuberculosis* **95** (2015).
16. M. Jacobsen, D. Repsilber, A. Gutschmidt, A. Neher, K. Feldmann, H. J. Mollenkopf, A. Ziegler and S. H. Kaufmann, Candidate biomarkers for discrimination between infection and disease caused by mycobacterium tuberculosis, *Journal of molecular medicine* **85** (2007).
17. S. Leong, Y. Zhao, N. M. Joseph, N. S. Hochberg, S. Sarkar, J. Pleskunas, D. Hom, S. Lakshminarayanan, C. R. Horsburgh Jr, G. Roy *et al.*, Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south india, *Tuberculosis* **109** (2018).
18. J. Maertzdorf, G. McEwen, J. Weiner, S. Tian, E. Lader, U. Schriek, H. Mayanja-Kizza, M. Ota, J. Kenneth and S. H. Kaufmann, Concise gene signature for point-of-care classification of tuberculosis, *EMBO molecular medicine* **8** (2016).
19. A. Sambarey, A. Devaprasad, A. Mohan, A. Ahmed, S. Nayak, S. Swaminathan, G. D'Souza, A. Jesuraj, C. Dhar, S. Babu *et al.*, Unbiased identification of blood-based biomarkers for pulmonary tuberculosis by modeling and mining molecular interaction networks, *EBioMedicine* **15** (2017).
20. H. Warsinske, R. Vashisht and P. Khatri, Host-response-based gene signatures for tuberculosis diagnosis: A systematic comparison of 16 signatures, *PLoS medicine* **16** (2019).
21. J. G. Burel, M. Babor, M. Pomaznoy, C. S. Lindestam Arlehamn, N. Khan, A. Sette and B. Peters, Host transcriptomics as a tool to identify diagnostic and mechanistic immune signatures of tuberculosis, *Frontiers in immunology* **10** (2019).
22. R. R. Chowdhury, F. Vallania, Q. Yang, C. J. L. Angel, F. Darboe, A. Penn-Nicholson, V. Rozot, E. Nemes, S. T. Malherbe, K. Ronacher *et al.*, A multi-cohort study of the immune factors associated with m. tuberculosis infection outcomes, *Nature* **560** (2018).
23. J. Maertzdorf, M. Ota, D. Repsilber, H. J. Mollenkopf, J. Weiner, P. C. Hill and S. H. Kaufmann, Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis, *PloS one* **6** (2011).
24. J. M. Cliff, J.-S. Lee, N. Constantinou, J.-E. Cho, T. G. Clark, K. Ronacher, E. C. King, P. T. Lukey, K. Duncan, P. D. Van Helden *et al.*, Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response, *The Journal of infectious diseases* **207** (2012).
25. J. Maertzdorf, J. Weiner, H.-J. Mollenkopf, T. Network, T. Bauer, A. Prasse, J. Müller-Quernheim and S. H. Kaufmann, Common patterns and disease-related signatures in tuberculosis and sarcoidosis, *Proceedings of the National Academy of Sciences* **109** (2012).
26. J. Cliff, J.-S. Lee, N. Constantinou, J.-E. Cho, T. G. Clark, K. Ronacher, E. C. King, P. T. Lukey, K. Duncan, P. D. Van Helden *et al.*, Tuberculosis patients blood gene expression through treatment (2012).
27. C. I. Bloom, C. M. Graham, M. P. Berry, K. A. Wilkinson, T. Oni, F. Rozakeas, Z. Xu, J. Rossello-Urgell, D. Chaussabel, J. Banchereau *et al.*, Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy, *PloS one* **7** (2012).
28. T. H. Ottenhoff, R. H. Dass, N. Yang, M. M. Zhang, H. E. Wong, E. Sahiratmadja, C. C. Khor, B. Alisjahbana, R. Van Crevel, S. Marzuki *et al.*, Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis, *PloS one* **7** (2012).
29. R. P. Lai, G. Meintjes, K. A. Wilkinson, C. M. Graham, S. Marais, H. Van der Plas, A. Deffur, C. Schutz, C. Bloom, I. Munagala *et al.*, Hiv–tuberculosis-associated immune reconstitution inflammatory syndrome is characterized by toll-like receptor and inflammasome signalling, *Nature communications* **6** (2015).
30. L. D. Tientcheu, J. Maertzdorf, J. Weiner, I. M. Adetifa, H.-J. Mollenkopf, J. S. Sutherland,

S. Donkor, B. Kampmann, S. H. Kaufmann, H. M. Dockrell *et al.*, Differential transcriptomic and metabolic profiles of m. africanum-and m. tuberculosis-infected patients after, but not before, drug treatment, *Genes and immunity* **16** (2015).

31. H. Esmail, R. P. Lai, M. Lesosky, K. A. Wilkinson, C. M. Graham, S. Horswell, A. K. Coussens, C. E. Barry, A. O?Garra and R. J. Wilkinson, Complement pathway gene activation and rising circulating immune complexes characterize early disease in hiv-associated tuberculosis, *Proceedings of the National Academy of Sciences* **115** (2018).
32. N. D. Walter, M. A. Miller, J. Vasquez, M. Weiner, A. Chapman, M. Engle, M. Higgins, A. M. Quinones, V. Rosselli, E. Canono *et al.*, Blood transcriptional biomarkers for active tuberculosis among patients in the united states: a case-control study with systematic cross-classifier evaluation, *Journal of clinical microbiology* **54** (2016).
33. A. Sambarey, A. Devaprasad, P. Baloni, M. Mishra, A. Mohan, P. Tyagi, A. Singh, J. Akshata, R. Sultana, S. Buggi *et al.*, Meta-analysis of host response networks identifies a common core in tuberculosis, *NPJ Systems Biology and Applications* **3** (2017).
34. S. Blankley, C. M. Graham, J. Turner, M. P. Berry, C. I. Bloom, Z. Xu, V. Pascual, J. Banchereau, D. Chaussabel, R. Breen *et al.*, The transcriptional signature of active tuberculosis reflects symptom status in extra-pulmonary and pulmonary tuberculosis, *PloS one* **11** (2016).
35. S. Marais, R. P. Lai, K. A. Wilkinson, G. Meintjes, A. O?Garra and R. J. Wilkinson, Inflammasome activation underlying central nervous system deterioration in hiv-associated tuberculosis, *The Journal of infectious diseases* **215** (2016).
36. J. Wu and R. I. with contributions from James MacDonald Jeff Gentry, *gcrma: Background Adjustment Using Sequence Information*, (2018). R package version 2.54.0.
37. W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke *et al.*, Orchestrating high-throughput genomic analysis with bioconductor, *Nature methods* **12** (2015).
38. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Research* **43** (2015).
39. J. M. Franks, G. Cai and M. L. Whitfield, Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data, *Bioinformatics* **34** (2018).
40. X. Chen and H. Ishwaran, Random forests for genomic data analysis, *Genomics* **99** (2012).
41. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman, Missing value estimation methods for dna microarrays, *Bioinformatics* **17** (2001).
42. H. Kim, G. H. Golub and H. Park, Missing value estimation for dna microarray gene expression data: local least squares imputation, *Bioinformatics* **21** (2004).
43. S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara and S. Ishii, A bayesian missing value estimation method for gene expression profile data, *Bioinformatics* **19** (2003).
44. M. C. De Souto, P. A. Jaskowiak and I. G. Costa, Impact of missing data imputation methods on gene expression clustering and classification, *BMC bioinformatics* **16** (2015).
45. A. Kamburov, C. Wierling, H. Lehrach and R. Herwig, Consensuspathdb?a database for integrating human functional interaction networks, *Nucleic acids research* **37** (2008).
46. A. M. Mandalakas and J. R. Starke, Current concepts of childhood tuberculosis, in *Seminars in pediatric infectious diseases*, (2)2005.
47. T. J. Scriba, A. Penn-Nicholson, S. Shankar, T. Hraha, E. G. Thompson, D. Sterling, E. Nemes, F. Darboe, S. Suliman, L. M. Amon *et al.*, Sequential inflammatory processes define human progression from m. tuberculosis infection to tuberculosis disease, *PLoS pathogens* **13** (2017).