

Published in final edited form as:

Eur Respir J. 2016 June ; 47(6): 1873–1876. doi:10.1183/13993003.02121-2015.

A 380 gene meta-signature of active TB compared to healthy controls

Simon Blankley¹, Christine M. Graham¹, Joe Levin¹, Jacob Turner², Matthew P.R. Berry^{1,3}, Chloe I Bloom¹, Zhaohui Xu⁴, Virginia Pascual⁴, Jacques Banchereau⁵, Damien Chaussabel⁶, Ronan Breen⁷, George Santis^{7,8}, Derek M. Blankenship², Marc Lipman^{9,10}, and Anne O'Garra^{1,11}

¹Laboratory of Immunoregulation and Infection, The Francis Crick Institute, Mill Hill Laboratory formerly The MRC National Institute for Medical Research), London, UK

²Baylor Research Institute, Baylor Scott & White, Dallas, Texas, United States of America

³Department of Respiratory Medicine, Imperial College Healthcare NHS Trust, St Mary's Hospital, London; UK

⁴Baylor Institute for Immunology Research/ANRS Center for Human Vaccines, INSERM, Dallas, Texas, USA

⁵The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032, Connecticut, USA

⁶Sidra Medical and Research Center, Doha, Qatar

⁷Department of Respiratory Medicine, Allergy and Lung Biology, King's College London, London, United Kingdom

⁸Division of Asthma, Allergy and Lung Biology, King's College London, London, United Kingdom

⁹Department of Respiratory Medicine, Royal Free London NHS Foundation Trust, London, UK

¹⁰Division of Medicine, University College London, London, UK

¹¹Department of Medicine, NHLI, Imperial College, London, UK

Introduction

Mycobacterium tuberculosis (*Mtb*) is estimated to have infected one third of the world's population and continues to be a significant cause of mortality and morbidity [1]. There is a need for new and improved diagnostics or treatment monitoring tools and blood based mRNA diagnostics are a potential solution [2]. Gene expression microarray analysis of human blood has been widely used to profile the host transcriptional response in active tuberculosis (TB) with the aim of identifying potential biomarkers as well as to better understand the host immune response [2]. So far, there has been a relative lack of concordance in the actual genes being identified from the published studies [2–4], although

there has been agreement in some of the pathways identified. Interferon signalling has been identified as a dominant signature in many of the individual studies [2], however when significant gene lists were combined from eight publicly available TB data-sets TREM1 signalling became the most significant pathway [5].

We therefore undertook this study to collectively reanalyse the publicly available datasets using differing methodologies to identify robustly differentially expressed genes which could distinguish active TB from controls. These identified genes are potential candidates for blood based mRNA biomarkers of active disease and additionally could provide valuable information regarding the immune and inflammatory response underlying TB pathogenesis.

Methods

In August 2014 a comprehensive search of Pubmed as well as microarray depositories was undertaken. Publicly available datasets that had active TB patients and healthy controls, latently infected or patients post treatment were identified and retained. The latter three cohorts are synonymous transcriptionally at the group level [3, 6–8]. HIV infected individuals were excluded from this analysis.

Where possible, data were imported in its raw format, Illumina and Agilent data were 75th centile normalised, and Affymetrix data RMA quantile normalised. If raw data were not available then the authors' normalisation was used. All datasets were then filtered for low expression transcripts (transcripts with expression 2 fold change (FC) from the median in at least 10% of all samples retained) followed by statistical filtering (independent t-test with Benjamini Hochberg multiple testing correction q -value <0.05). Probe/transcript IDs were matched to Entrez gene identifiers for each dataset. Multiple represented genes were filtered and the most significant (by q -value) retained. Venn Mapping [9] was used to check significance of the overlaps between any two datasets. Meta-profiling [10] of the significant gene lists was then undertaken to identify the number of overlaps required for inclusion in the meta-signature, additionally only those genes which were expressed in a consistent direction of regulation for at least the number of determined overlaps were retained as the meta-signature.

Modular analysis [11] was undertaken of compatible datasets, it was not possible to analyse GSE56153, GSE34608 and GSE28623 with this method as their technology platforms were not supported by the tool.

The canonical pathway, gene network analysis, gene function annotation and upstream analyses were generated through the use of IPA (Ingenuity[®] Systems, www.ingenuity.com).

Results

Modular and meta-profiling identify a common transcriptional response of patients with TB versus healthy controls, patients post-treatment and asymptomatic latently infected individuals

Sixteen datasets were identified for inclusion in the meta-analysis (GSE39939, GSE39940, GSE54992, GSE37250, GSE31348, GSE36238, GSE42825, GSE42826, GSE42830,

GSE40553, GSE56153, GSE34608, GSE28623, GSE19444, GSE19442 and GSE19439). Modular analysis of these datasets revealed similarities in the modular profiles with over-expression of modules annotated as cytotoxic, interferon, inflammation and DC/apoptosis, and under-expression of modules annotated as B-cells, T-cells, lymphocyte activation and mitochondrial stress (Figure 1A). Two datasets (GSE54992 and GSE36238) had different weaker modular patterns with less modules identified as significantly different from the control group; these two were also the smallest in terms of total number of individuals within the datasets.

Independently for each dataset differentially expressed genes between controls and TB groups were identified. There was significant overlap between the differentially expressed gene lists across the datasets with no difference in degree of overlap dependent on the choice of control group (Figure 1B). Meta-profiling identified 380 genes which were identified in nine or more datasets in a consistent direction of regulation (Figure 1C). Upregulated genes were more consistently identified across datasets than down-regulated genes (Figure 1D). Five genes were identified in all 16 datasets *AIM2*, *BATF2*, *FCGR1B*, *HP* and *TLR5*.

IFN- γ was the top predicted upstream regulator of these 380 meta-signature genes with 54 genes directly or indirectly linked to IFN- γ within the IPA database (Data not shown). The 380 meta-signature had enrichment for IPA canonical pathways involved in pattern recognition, interferon signalling, IL-6 signalling, TREM-1 signalling and complement (Data not shown). Based on these findings a curated cartoon summarising the major functional groups of genes and their relationships was created (Figure 1E).

Discussion

In this study we have identified a 380 gene meta-signature of active TB compared to healthy controls, patients post-treatment and asymptomatic latently infected individuals which showed enrichment for both innate and adaptive immune functions.

Two main methodologies were used to analyse the publicly available data: modular analysis and meta-profiling. Modular analysis depends on identifying differences in coordinately expressed groups of genes (modules) rather than individual genes [11]. We identified remarkable similarity between the datasets with over-expression of modules annotated with interferon, inflammation functions, monocyte and neutrophil functions, and under-expression of B and T-cell modules. These findings are in keeping with the individual studies which have included grouped modular analysis [7, 8]. Where grouped modular profiles were less consistent this may have resulted from the small cohort sizes used and emphasises the need for individual studies to be appropriately powered to detect all differentially expressed genes.

Using meta-profiling we found that 380 genes were consistently differentially expressed in nine or more datasets, with 5 genes identified as differentially expressed in all 16 datasets. These 5 genes included *AIM2*, *BATF2*, *FCGR1B* and *HP* which have been shown to potentially play a role following *Mtb* infection [12–15]. A role for *TLR5* has yet to be described and identification of such differentially regulated genes may therefore be part of a

programmatic response rather than specifically targeted for a tailored response to the pathogen.

Analysis of the 380 genes comprising the meta-signature identified IFN- γ as the most significant potential upstream regulated molecule, with a large network of IFN- γ regulated genes present within the 380 genes. IFN- γ is critical for control of mycobacterial disease in humans, with mutations either in the IFN- γ receptor or STAT1 resulting in increased susceptibility [16]. However, upregulation of gene expression molecules downstream of Type I IFN signalling was also observed, of relevance to TB exacerbation, since Type I IFN has been shown to antagonise signalling downstream of IFN- γ [17]. Thus, capturing the overall picture of significant enrichment may be more informative than identification of one individual pathway, as shown in the summary cartoon (Figure 1E). A number of immune pathways/functions are enriched for within the meta-signature including multiple pattern recognition receptors, cytokines, the inflammasome, complement and immunoglobulin. This supports diverse findings obtained from both mouse and human studies that the immune response following *Mtb* infection is complex and can be cross-regulatory [16].

This study confirms the reproducibility of blood based transcriptional analysis to identify the innate and adaptive host response in TB. It also identifies that upregulated mRNA transcripts are more reliably identified and highlights several mRNA candidates which could collectively be used as potential biomarkers of active disease. These findings have implications for the design and implementation of mRNA expression tools to support diagnostics and treatment monitoring of TB.

Funding

SB and AOG were supported by the Medical Research Council, United Kingdom (U117565642), now The Francis Crick (AOG Crick Budget 10126). SB was jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement. (MR/J010723/1). GS was supported in part by the Department of Health via the National Institutes of Health Research comprehensive Biomedical Research Centre award to Guy's and St Thomas' National Health Service Foundation Trust in partnership with King's College London. This study received financial support from SILA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. World Health Organization. World Health Organisation: Global tuberculosis report 2014. 2014.
2. Blankley S, Berry MP, Graham CM, Bloom CI, Lipman M, O'Garra A. The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2014; 369:20130427.
3. Bloom CI, Graham CM, Berry MP, et al. Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PloS one*. 2013; 8:e70630. [PubMed: 23940611]
4. Kaforou M, Wright VJ, Oni T, et al. Detection of Tuberculosis in HIV-Infected and -Uninfected African Adults Using Whole Blood RNA Expression Signatures: A Case-Control Study. *PLoS medicine*. 2013; 10:e1001538. [PubMed: 24167453]
5. Joosten SA, Fletcher HA, Ottenhoff TH. A helicopter perspective on TB biomarkers: pathway and process based analysis of gene expression data provides new insight into TB pathogenesis. *PloS one*. 2013; 8:e73230. [PubMed: 24066041]

6. Bloom CI, Graham CM, Berry MP, et al. Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy. *PloS one*. 2012; 7:e46191. [PubMed: 23056259]
7. Cliff JM, Lee JS, Constantinou N, et al. Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *The Journal of infectious diseases*. 2013; 207:18–29. [PubMed: 22872737]
8. Berry MP, Graham CM, McNab FW, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*. 2010; 466:973–7. [PubMed: 20725040]
9. Smid M, Dorssers LC, Jenster G. Venn Mapping: clustering of heterologous microarray data based on the number of co-occurring differentially expressed genes. *Bioinformatics*. 2003; 19:2065–71. [PubMed: 14594711]
10. Rhodes DR, Yu J, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101:9309–14. [PubMed: 15184677]
11. Chaussabel D, Quinn C, Shen J, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*. 2008; 29:150–64. [PubMed: 18631455]
12. Saiga H, Kitada S, Shimada Y, et al. Critical role of AIM2 in *Mycobacterium tuberculosis* infection. *International immunology*. 2012; 24:637–44. [PubMed: 22695634]
13. Tussiwand R, Lee WL, Murphy TL, et al. Compensatory dendritic cell development mediated by BATF-IRF interactions. *Nature*. 2012; 490:502–7. [PubMed: 22992524]
14. Jacobsen M, Repsilber D, Gutschmidt A, et al. Candidate biomarkers for discrimination between infection and disease caused by *Mycobacterium tuberculosis*. *Journal of molecular medicine*. 2007; 85:613–21. [PubMed: 17318616]
15. Kasvosve I, Gomo ZA, Mvundura E, et al. Haptoglobin polymorphism and mortality in patients with tuberculosis. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*. 2000; 4:771–5.
16. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MPR. The immune response in tuberculosis. *Annual review of immunology*. 2013; 31:475–527.
17. McNab F, Mayer-Barber K, Sher A, Wack A, O'Garra A. Type I interferons in infectious disease. *Nature reviews Immunology*. 2015; 15:87–103.

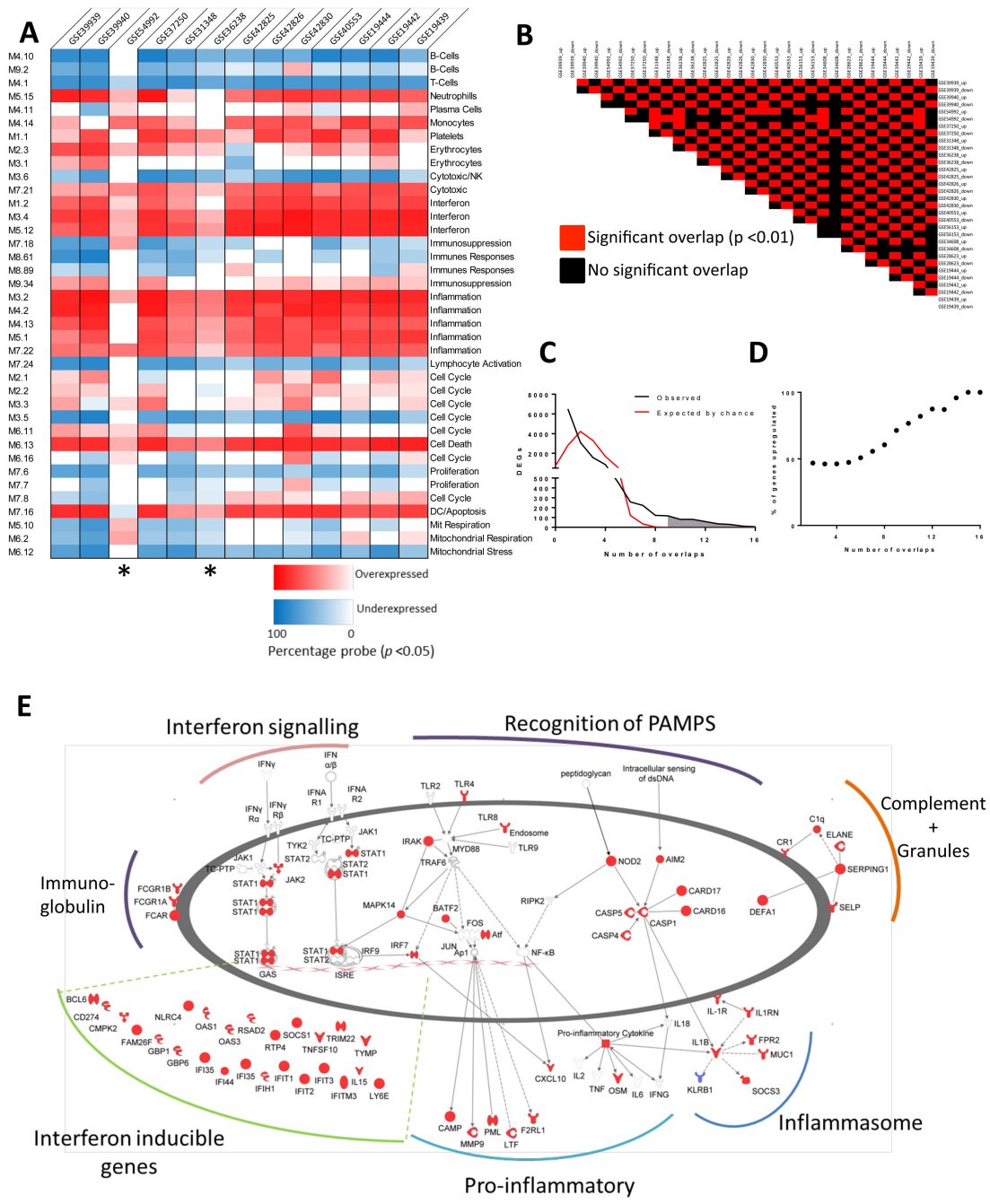


Figure 1: Modular and meta-profiling identifying common transcriptional response. (A) Modular analysis of 13 publicly available and compatible datasets of TB compared to control groups. 38 annotated modules are displayed as a heatmap with red indicating significant over-abundance of transcripts and a blue indicating significant under-abundance ($p < 0.05$). The colour intensity represents the percentage of genes in that module which are significantly differentially expressed. Two datasets with different modular profiles marked with asterisks. (B) Venn Mapper used to identify significance of overlap between any two differentially expressed gene lists. Significance of overlap between significant gene lists

calculated by Venn Mapper programme, with methodology adapted from *Smid et al* [9]. Significant genes identified following filtering of probes for low expression (probes eliminated if expression less than 2FC from median normalised value in 10% or more of samples), followed by statistical filtering; independent t-test with Benjamini Hochberg multiple testing correction between TB and control group. Probes were then matched to Entrez gene IDs which was used as the reference “array” for analysis. The fold change representation for each gene was the mean fold change of TB group compared to control group, where genes were multiply represented on an array the fold change associated with the most significant *q-value* was chosen. **(C)** Differentially expressed genes (DEGs) were identified for each of the sixteen datasets (probes filtered for low expression, followed by statistical filtering; independent t-test with Benjamini Hochberg multiple testing correction between TB and control group) and then using meta-profiling to simulate the data and identify the number of overlaps to define the meta-signature (shaded grey). **(D)** Genes were grouped by number of datasets in which they were significantly identified, and the percentage calculated which were upregulated (direction of regulation relative to control group; where there were inconsistencies in direction of regulation across datasets the direction most often observed was used to determine direction). **(E)** Curated cartoon of 380 meta-signature generated using IPA and IPA knowledge base.