A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection

Akul Singhania[1], Raman Verma[2], Christine M. Graham[1], Jo Lee[2], Tran Trang[3], Matthew Richarsdon[2], Patrick Lecine[3], Philippe Leissner[3], Matthew P.R. Berry[4], Robert J. Wilkinson[5,6,7], Karine Kaiser[8], Marc Rodrigue[8], Gerrit Woltmann[2], Pranabashis Haldar[2], Anne O'Garra[1,9]

[1] Laboratory of Immunoregulation and Infection, The Francis Crick Institute, London, United Kingdom.
[2] Respiratory Biomedical Research Centre, Institute for Lung Health, Department of Infection Immunity and Inflammation, University of Leicester, Leicester, United Kingdom.
[3] Bioaster Microbiology Technology Institute, Lyon, France.
[4] Department of Respiratory Medicine, Imperial College Healthcare NHS Trust, St Mary's Hospital, London, United Kingdom.
[5] Wellcome Centre for Infectious Diseases Research, Africa; Institute for Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory 7925, South Africa.
[6] Department of Medicine, Imperial College London, London, United Kingdom.
[7] Tuberculosis Laboratory, The Francis Crick Institute, London, United Kingdom.
[8] Medical Diagnostic Discovery Department, bioMérieux SA, Marcy l'Etoile, France.
[9] National Heart and Lung Institute, Imperial College London, London, United Kingdom.

Akul Singhania and Raman Verma contributed equally to this work.
Anne O'Garra and Pranabashis Haldar contributed equally to this work.

Correspondence and requests for materials should be addressed to A.O'G. (email: Anne.OGarra@crick.ac.uk)

Abstract

Whole blood transcriptional signatures distinguishing patients with active tuberculosis from asymptomatic latently infected individuals have been described but, no consensus exists for the composition of optimal reduced gene sets as diagnostic biomarkers that also achieve discrimination from other diseases. We have recapitulated a blood transcriptional signature of active tuberculosis using RNA-Seq, previously reported by microarray that discriminates active tuberculosis from latently infected and healthy individuals, also validated in an independent cohort. We show that an advanced modular approach, which preserves and presents a signature of the entire transcriptome, can better discriminate patients with active tuberculosis from both latently infected and acute viral and bacterial infections. We suggest a method of targeted gene selection across modules for constructing diagnostic biomarkers, more representative of the transcriptome that overcomes some limitations of existing techniques. Finally, we utilise the modular approach to demonstrate dynamic heterogeneity in a longitudinal study of recent tuberculosis contacts.

Tuberculosis (TB) is the leading cause of global mortality from an infectious disease. In 2016, there were 6.3 million new cases of TB disease and 1.67 million deaths and its diagnosis is problematic[1]. However, clinical disease represents one end of a spectrum of infection states. It is estimated that up to one third of all individuals worldwide have been infected with the causative pathogen, *Mycobacterium tuberculosis,* but the vast majority remain clinically asymptomatic with no radiological or microbiological evidence for active infection. This is termed latent TB infection (LTBI) and conceptually denotes a state in which *M. tuberculosis* persists within its host, while maintaining viability with the potential to replicate and cause symptomatic disease. Indeed, LTBI represents the primary reservoir for future incident TB, with 90% of all TB cases estimated to arise from reactivation of existing infection[1,2]. The risk of incident TB arising from existing LTBI is heterogeneous, poorly characterised and modifiable with anti-tuberculous treatment. Modelling studies indicate effective TB prevention to significantly reduce future TB incidence requires policies directed at the identification and treatment of LTBI[3]. However, implementation of mass screening programmes for this purpose are severely constrained by the size of the target population. Transformative advances in diagnostic tools that can effectively stratify TB risk in the LTBI population are therefore implicit to the realisation of systematic screening.

The basis for LTBI heterogeneity rests with the limited scope of the tools we have available to identify the state. LTBI is inferred solely through evidence that immune sensitization has occurred, by the tuberculin skin test (TST) or the *M. tuberculosis* antigen-specific interferon-gamma (IFN-γ) release assay (IGRA). Although these tests are both sensitive and specific for identifying exposure that has been associated with establishment of an adaptive immune response, neither distinguishes active from latent infection. Moreover, T-cell responses to mycobacterial antigens persist for several years after an infection has been treated, implying that these tests may not reliably inform the presence of viable organisms *in vivo*. For 'true' LTBI, in which the pathogen remains viable, it is envisaged that a dynamic equilibrium exists between the host immune response and the pathogen, with a shifting balance in favour of one or the other influencing the future risk of TB reactivation[4]. A recent study using highly sensitive radiological imaging with combined

Positron Emission Tomography and Computerised Tomography has reported evidence to support this dynamic state and demonstrated phenotypic imaging characteristics associated with the risk of developing TB among subjects with conventionally defined LTBI[5]. A proportion of these LTBI patients were identified with radiological features of subclinical active TB[5], with a subgroup failing to respond to prophylactic LTBI treatment regimens. These observations support the view that injudicious use of LTBI chemoprophylaxis using presently available diagnostic tools for mass screening, risks promoting drug resistance in unrecognised active infection.

We have previously characterised an interferon (IFN) inducible transcriptional signature of 393 gene transcripts in whole blood that discriminates patients with active pulmonary TB (from high and low-incidence TB burden countries) from healthy individuals, patients with other chronic respiratory and systemic conditions, and the majority of patients with LTBI[6,7]. This TB signature revealed an unexpected dominance of type I IFN-inducible genes[6] more frequently associated with viral infections[8]. We[9-11] and others[10-20] have since shown that elevated and sustained levels of type I IFN, result in an enhanced mycobacterial load and disease exacerbation in experimental models of TB. Similar findings of a blood signature in active TB patients have since been reported[21-27], and our meta-analysis of 16 datasets, including many of these studies, identified 380 genes differentially abundant in active TB across all datasets[28]. However, there is a relative lack of concordance across studies that have reported a reduced and optimised diagnostic gene signature, although agreement exists for some of the pathways they represent[21-23,29,30]. While some genes overlap between the different reduced signatures, the overall composition of each reduced signature is unique, both in size and transcript profile. In this respect, we note that a consistent statistical approach to optimising gene selection has not been used across studies and where the approach was consistent, a different optimal reduced signature was reported for discriminating active TB from either LTBI and controls or other diseases[22]. Additionally, recent reports have demonstrated the failure of these signatures in discriminating between TB and other diseases such as pneumonia, highlighting their inadequacy as stand-alone diagnostic tests, and a need for more accurate tests[26,27].

We have previously observed and reported that 10 – 20% of subjects with IGRA positive LTBI in our studies had a transcriptional signature that overlapped with active TB patients and clustered with this group[6]. By definition, the transcriptional signature in this LTBI outlier group shares important similarities with the signature of active TB that requires further characterisation. Importantly, the biological significance of this statistical observation remains unclear. However, these observations support utilisation of a transcriptional approach to explore LTBI heterogeneity. In keeping with this, Zak et al.[30] have recently reported evidence for a gene signature of TB several months in advance of clinical presentation with disease among a cohort of South African adolescents, suggesting that transcriptional signatures of TB in subjects with presumed LTBI may indicate either a high risk of progression to active disease or existing subclinical disease. The study was unable to determine a transition in the signature prior to developing TB and was limited by the confounding risk of new exposure in a high TB incidence setting. Utilisation of the transcriptome to interrogate immunological heterogeneity within the cohort was not undertaken and analysis was largely confined to the subgroup with IGRA defined LTBI, the assumption being that IGRA negative subjects do not have latent infection. However, a proportion of prospective TB cases identified in the study were IGRA negative at baseline[30], suggesting either that this cohort had new exposure during prospective observation in a high TB incidence setting and / or that the IGRA test did not reliably inform underlying

LTBI. In this context, studies evaluating the diagnostic performance of IGRAs in microbiologically confirmed active TB report an overall sensitivity of approximately 85%[31], implying that a proportion of *M. tuberculosis* infections may be missed using this test alone.

To address some of these questions, we undertook RNA Sequencing (RNA-Seq) of our earlier Berry et al.[6] cohorts and additionally set up a prospective cohort study at Leicester (UK) in subject groups of incident TB and recent TB contacts, respectively. In the Leicester cohort, we performed systematic longitudinal sampling and clinical characterisation first, to validate our TB signature using RNA-Seq in a new and independent cohort of individuals with active TB and LTBI, and secondly to provide longitudinal data in a low TB incidence setting. Using an advanced modular approach for characterising transcriptional signatures, we now identify important similarities and differences between active TB, LTBI and other diseases that informs limitations of existing signatures and provides a template for targeted gene selection using information from the entire transcriptome. We also demonstrate utilisation of the modular approach to characterise phenotypes of LTBI among recent close contacts of TB.

Results

RNA-Seq reproduces the gene-signature developed using microarray recapitulating the clustering of active TB and LTBI cases

We validated our microarray-derived blood 393-transcript signature[6] in patients with active TB using RNA-Seq in the Berry London and South Africa cohorts showing identical clustering of active TB and LTBI cases (**Supplementary Figure 1a** and **1b**). A 373-gene signature was then independently re-derived from the Berry London RNA-Seq data **(Supplementary Figure 1c; Supplementary Table 1; Figure 1a)** and validated in the Berry South Africa cohorts **(Figure 1a)** and a new Leicester cohort **(Supplementary Table 2; Figure 1b).** Consistent with our previous microarray signature, the RNA-Seq signature was absent in the majority of individuals with LTBI and healthy controls, and identified with perfect agreement the LTBI subjects that cluster with active TB, henceforth referred to as LTBI outliers, in both Berry cohorts **(Supplementary Figure 1b** and **1d).** A similar proportion of outliers were also observed in the Leicester cohort **(Figure 1b; Supplementary Figure 1e).** There was great similarity in the composition of the microarray and RNA-Seq based signatures, with over-abundance of IFN-inducible genes and under-abundance of B- and T-cell genes (data not shown) as previously reported[6]. This was supported by an *in silico* cellular deconvolution analysis of the RNA-Seq that showed diminished percentages of CD4, CD8 and B cells in the blood of active TB patients, and an increase in monocytes/macrophages and neutrophils (**Supplementary Figure 2**), in keeping with our previous findings using flow cytometry[6]**.**

Evaluation of published TB gene signatures identifies overlap with, and poor discrimination from gene expression in acute viral infections

Applying the published 27-gene and 44-gene signatures of Kaforou et al.[22] and the 16-gene signature of Zak et al.[30] to the Berry and Leicester TB cohorts, a high specificity and sensitivity for discriminating active TB and LTBI was identified with all three signatures across all three cohorts **(Figure 1c)**. This was supported by single sample Gene Set

Enrichment Analysis[32] (ssGSEA), demonstrating high enrichment of the Zak et al. signature[30] in active TB and a low enrichment in healthy controls and the majority of LTBI patients **(Figure 1d)**. We observed higher enrichment scores in the LTBI outlier groups **(Figure 1a** and **b)** of all three cohorts that overlapped with scores observed in active TB cohorts **(Figure 1d)**. Higher enrichment scores were also noted in a small proportion of $IGRA^{-ve}$ individuals recruited as healthy controls **(Figure 1d)**. There was comparable discrimination in enrichment scores between TB and LTBI using all three signatures **(Figure 2a),** although the Kaforou 44-gene signature demonstrated greater overlap of enrichment scores between groups, suggesting poorer discriminatory performance. In this context, it is notable that of the three signatures, only the Kaforou 44-gene signature was developed to discriminate between active TB and other diseases (including infectious meningitis, pneumonia, gastric diseases and malignancies)[22], rather than LTBI.

The composition of all three signatures[22,30] is dominated (≥50% of the signature) by IFN-inducible genes **(Supplementary Table 3)**, raising the possibility that they are not TB specific but may also be expressed in acute viral infections. We therefore evaluated enrichment of these signatures in two independent published datasets of influenza infection from Parnell et al.[33] and Zhai et al.[34] **(Supplementary Table 4; Figure 2b** and **c)**. Subjects with influenza at baseline showed a high enrichment score for the three TB signatures as compared with healthy controls, which diminished with time, in keeping with recovery **(Figure 2b** and **2c)**. In keeping with this, all three signatures, developed for distinguishing active TB and LTBI, also demonstrated excellent discrimination between influenza (day 0) and healthy controls **(Figure 2d),** comparable with their performance for TB **(Figure 1c)**. In contrast, enrichment scores for the three signatures, demonstrated heterogeneity in patients diagnosed with bacterial pneumonia from the Parnell study[33], with little change over 5 days and poor discrimination from controls, consistent with our previous findings for this group[6,7] **(Figure 2e).**

A modular approach to transcriptional data analysis identifies clear differences in the signature of active TB and other pulmonary infections

A limitation of the gene reduction methodologies[22,30] used to date has been the prioritisation of the most discriminant genes, with little consideration to the correlation between the selected genes in this iterative process. Although non-selective and lacking subjective bias, this approach favours selection of a highly correlated gene set with a narrow immunological focus. In this context, limited diversity risks loss of specificity, with an increased likelihood of overlap between multiple pathologies and responses to different infections for a specific immune pathway. We therefore hypothesised that methodologies which incorporate information from the entire transcriptome may better inform development of a unique biosignature for TB. Weighted gene co-expression network analysis[35] (WGCNA) is a well validated clustering technique for reducing high dimensional data into modules that preserve intrinsic relationships between variables within a network structure. When applied to the blood transcriptome, modules of co-ordinately expressed genes with a coherent functional relationship are generated. The complete transcriptome is thus expressed as a signature defined by the relative perturbation of individual modules.

We applied WGCNA analysis to the blood transcriptional data from our Berry and Leicester TB cohorts, those TB cohorts published by Zak and Kaforou[22,30], and to several others that included sample sets of other viral and bacterial infections[33,34,36,37], together with our

previous cohorts of sarcoidosis and lung cancer[7] as conditions that may mimic TB, all compared against their healthy controls **(Figure 3; Supplementary Table 4** (Information of published cohorts)**, 5** (genes in each module) and **6** (module annotation)**).** The modular signature for active TB was qualitatively consistent across all the TB cohorts and absent in LTBI. The IFN-modules (lightgreen and yellow) were over-abundant in TB **(Figure 3a)** as we have previously published[6,7] and also in acute influenza infection, but absent in bacterial infection[6,7] **(Figure 3b)**. However, we observed clear differences between TB and both influenza and other bacterial infections in the pattern of specific perturbation of other modules, including under-abundance of gene expression in the T-cell (blue and cyan) and B-cell (midnightblue) modules **(Figure 3)** for TB. On the other hand, we observed over-abundance of genes in the Cell Proliferation/Metabolism (darkturquoise) module and under-abundance of genes associated with Haematopoeisis (pink) in severe influenza but not in TB **(Figure 3)**. In this context, the classical approaches of gene signature reduction algorithms[38-40] used by Kaforou et al. to distinguish TB from LTBI or TB from other diseases[22], and Zak et al. to distinguish TB from LTBI, risk of progression[30] are notable for formulating gene signatures that we show here map predominantly to the yellow module (Interferon/complement/myeloid), with many of these genes also over-abundant in both influenza cohorts (**Supplementary Figure 3** and **4).**

Differential gene expression within modules informs gene selection to develop a powerful discriminatory transcriptional signature for active TB

Interrogating the whole gene-set of the yellow module in TB, influenza and bacterial infection, we observed a subset of genes expressed specifically in TB **(Figure 4, orange squares)**. Similarly, other genes were specifically expressed in influenza. Thus, although modular expression of the yellow module is comparable between TB and influenza, gene subsets within the module exhibit differential expression between the two conditions. This provides scope to select genes from this dominant module that can be used to develop a TB signature, while retaining discriminant value from viral infection. Using this rationale, we identified and extracted 303 unique gene candidates in the Berry London TB dataset that were selectively perturbed in TB, but not in any confounding viral infections, from all modules that contributed to and exhibited consistency across the TB datasets that we analysed (**Figure 3; Supplementary Figure 5a; Supplementary Figure 5b**). Using this gene set, we proceeded to develop a reduced gene signature to distinguish active TB from LTBI. We applied the Boruta algorithm[39] based on random forest to this set of genes, yielding 61 genes **(Supplementary Figure 5c)** that was further reduced by selecting the top 20 genes, ranked according to GINI score using Random Forest **(Supplementary Figure 5d)**. Our 20-gene signature **(Figure 5a)** included genes from six different modules **(Supplementary Figure 5d)**, representing both over-abundance and under-abundance in TB. Using a modified Disease Risk Score (See Methods), we identified powerful discrimination between active TB and LTBI/controls in Berry London & South Africa and Leicester cohorts (**Figure 5b;** Sensitivity/Specificity/Area under the curve for – Berry London 1/1/1, Berry South Africa 1/1/1, Leicester – 1/0.86/0.99). In contrast, the signature identified no difference between influenza and controls or between bacterial pneumonia and controls at any time-point across five days **(Figure 5c).**

LTBI outliers exhibit a distinct modular signature with features of Active TB

We have previously reported evidence for a small proportion of LTBI subjects that clustered with active TB using our 393-transcript signature[6] that we refer to as an LTBI outlier group. This group was reproduced using RNA-Seq in the Berry cohorts (10.9%) and a similar proportion were also identified in our new Leicester cohort (10%) (**Figure 1**). To compare and contrast the signature of this group with active TB and the majority of LTBI resembling healthy controls **(Supplementary Figure 1d** and **1e)**, we specifically examined the WGCNA modular signature in LTBI outliers using the combined Berry London and South Africa datasets and Leicester datasets respectively, compared with healthy controls **(Figure 6a).** The modular signature of LTBI outliers in both datasets showed over-abundance of the lightgreen (IFN/Pattern recognition receptors) and yellow (IFN/Complement/myeloid) modules as seen in active TB **(Figure 6a** and 6**b)**. This is entirely in keeping with our earlier finding that gene enrichment scores using the three published signatures of Kaforou and Zak **(Figures 1d** and **2a)**, all of which are comprised primarily of genes from the yellow module, were consistently higher in LTBI outliers. In addition to over-abundance of the IFN modules, the LTBI outlier group of the Leicester dataset showed changes in other modules also perturbed in active TB, suggesting a host response that is evolving towards the phenotype typically observed in active TB **(Figure 6a)**. Of particular interest was the observation of under-abundance in the tan module (Th1 and NK cells) that is associated with IFN-γ expression, a cytokine required for protection against TB[13,41-47]. Under-abundance of this module was a consistent finding across all the TB datasets that we analysed **(Figure 3, Figure 6)** and was a characteristic shared with bacterial pneumonia.

We performed differential gene expression analysis between active TB, LTBI outliers, and LTBI with outliers removed, and identified a set of 70 genes that was consistently upregulated in active TB and LTBI outliers compared to LTBI (without outliers) in both the Berry and Leicester datasets **(Figure 6d; Supplementary Table 7),** which were enriched for the IFN signalling pathway (data not shown).

The modular transcriptional signature is dynamic and exhibits heterogeneity in recent TB contacts

Longitudinal RNA-Seq was performed in a subset of our Leicester cohort **(Methods; Figure 7a)** that included 15 IGRA^{-ve} contacts, 16 IGRA^{+ve} contacts, both of whom remained healthy, and 9 subjects recruited as contacts that were subsequently diagnosed with microbiologically confirmed TB during prospective observation (**Figure 7a; Supplementary Table 8)**. Five contacts (4 IGRA^{+ve} and 1 IGRA^{-ve}) identified as outliers at baseline sequencing **(Figure 1b)** were included.

In contrast with other studies, the control population of our Leicester cohort comprised subjects that were IGRA^{-ve} contacts of TB. This is a group in which recent exposure to active TB is documented, placing them at higher risk of recently acquired infection. Our rationale for this approach was to evaluate whether transcriptional data may identify LTBI that is not detected using IGRA. The observations that: firstly, the Leicester control group had greater overlap in enrichment scores with the IGRA^{+ve} LTBI group using the Zak and Kaforou signatures, compared with the Berry London cohort **(Figure 1d** and **2a)**; and secondly, one subject from this group was identified as an outlier, together suggest that IGRA testing alone may miss some *M. tuberculosis* infection. We therefore elected to define our TB contacts henceforth as IGRA^{+ve} or IGRA^{-ve} with no deterministic reference to LTBI.

The modular signatures of both IGRA^{-ve} and IGRA^{+ve} contacts qualitatively demonstrated considerable between-subject heterogeneity and some within-subject variability; a comparison between the groups suggested more transcriptional activity, in the form of a higher frequency and greater breadth of modules exhibiting overabundance and under-abundance within the IGRA^{+ve} group **(Figure 7b** and **7c)**. For the cohort that developed TB after recruitment to the study **(Figure 7d),** we stratified subjects on the basis of their longitudinal clinical course as true progressors (no evidence of TB at baseline, with features developing during observation); subclinical TB (objective evidence of pathology, usually as radiological change, in the absence of reported symptoms); and active TB (symptoms at baseline with either radiological or microbiological evidence for TB subsequently identified) **(Supplementary Table 8)**. This stratification was performed to better understand the dynamic relationship between the modular signature and onset of TB.

To quantitatively evaluate the modular signatures in each group for their proximity to TB, we applied the modified disease risk score for our 20-gene signature **(Supplementary Figure 6)**. Higher risk scores were generally observed in the IGRA^{+ve}, compared with the IGRA^{-ve} cohort, although there was considerable variability and overlap. Longitudinal observations suggest relative stability of the risk score in the majority of both IGRA^{+ve} and IGRA^{-ve} subjects that were examined. In contrast, 6 of the 9 subjects that were diagnosed with TB demonstrated high baseline modified disease risk scores that tended to increase further, prior to diagnosis of active TB. In the other 3 contacts (Subjects 245, 348 and 278) the modified disease risk score remained low at all time-points, before and at the time of TB diagnosis **(Supplementary Figure 6c)**.

To improve specificity, we additionally devised a weighted TB agreement score as a quantitative measure of proximity to TB, based on a composite of categorical scoring of agreement between the test signature and a reference signature for active TB across all 23 modules (Methods). Baseline agreement scores demonstrated clustering near zero for IGRA^{-ve} subjects **(Supplementary Figure 7; Figure 8)**. In contrast, both the IGRA^{+ve} group, and the group that developed TB exhibited a wide range of positive scores **(Figure 8b** and **8c)**. Subjects identified as outliers **(Figure 1b)** had higher TB agreement scores than the majority of LTBI subjects that were not outliers **(Figure 8b)**. However, some discordance between clustering outcomes and the TB agreement score was observed, with two subjects that were not outliers having TB agreement scores within the outlier range (subject 185 and 040). Furthermore, the IGRA^{-ve} subject categorised as an outlier (Subject 209) had a very low TB agreement score **(Figure 8a)**, with resolution of the baseline module perturbations after 4 months **(Figure 7b)**. Overall, the longitudinal within-subject expression of the modular TB signature in both the IGRA^{+ve} and IGRA^{-ve} cohorts could be categorised into three groups: i. Subjects that did not express the signature at any time-point (9 of 15 IGRA^{-ve} subjects and 6 of 16 IGRA^{+ve} subjects); ii. Subjects that transiently expressed the signature in the first three to four months (4 of 15 IGRA^{-ve} subjects and 6 of 16 IGRA^{+ve} subjects; iii) Subjects that had or developed a persistent TB signature at and beyond 4 months (2 of 15 IGRA^{-ve} subjects and 4 of 16 IGRA^{+ve} subjects)**(Figure 8b)**. We did not observe subjects developing the signature *de novo* after 3 months.

In the cohort that developed TB, 5 of the 9 subjects demonstrated high baseline modified disease risk scores that were similar to values observed in the 6 out of 16 IGRA^{+ve} subjects **(Figure 8c)**. In 7 of the 9 subjects a moderate to high TB agreement score was observed at the visit prior to TB diagnosis. For the remaining 2 subjects (Subject 245 and Subject 348), a modular signature of TB was not expressed. For Subject 245 an explanation may be that

this patient received antibiotics for bacterial pneumonia which are known to have immunosuppressive effects and therefore will diminish an immune signature. For Subject 348 we have not identified potential confounding factors for this observation. Subjects categorised as true progressors exhibited a dynamic modular signature, with increasing TB agreement scores at all visit time-points within 2 months of diagnosis.

Discussion

We have recapitulated a blood transcriptional signature of active TB using RNA-Seq, previously reported by microarray[6,22,23,28,48] that discriminates active TB from LTBI and healthy individuals, and is largely characterised by an over-abundance of IFN-inducible genes and an under-abundance of B and T cell genes. We show that an advanced modular approach, rather than a traditionally derived reduced gene set, is more robust in discriminating active TB patients from individuals with LTBI and acute viral and bacterial infections. Using this modular approach, we also demonstrate heterogeneity of LTBI in a prospective study of contacts of patients with active TB.

RNA-Seq[30] has now replaced microarray[6,7,21-24,28,29,48-50] for transcriptional studies and the existing literature is limited by uncertainty regarding the equivalence of RNA-Seq and microarray. In this study, we repeated analysis of our previous Berry et al.[6] cohorts using RNA-Seq and provide reassurance that RNA-Seq recapitulates outcomes derived using microarray. The vast majority of genes in our RNA-Seq derived 373-gene signature also comprised our original 393-gene transcript signature. Furthermore, there was equivalence in allocation of subjects to clusters, including those with LTBI that clustered with active TB.

In transcriptomic studies of disease there has been focus on deriving reduced gene signatures to develop clinical diagnostics, with inconsistencies in both deriving and defining the optimal reduced gene signature. Studies defining signatures distinguishing active TB and LTBI[21-23,28,30] are illustrative of this issue. We evaluated the diagnostic performance of some of the published reduced gene signatures[22,30] on our independent TB cohorts and confirm excellent specificity and sensitivity for distinguishing active TB patients from those with LTBI. However, we identified dominance of IFN-inducible genes in these signatures and demonstrated enrichment of these signatures in published datasets of acute influenza infection, but not bacterial pneumonia. These observations highlight the lack of IFN-inducible genes in the immunological response to bacterial pneumonia, which contrasts both TB and viral infection. It follows that discordance exists such that signatures optimised for discriminating active TB patients and healthy individuals, with and without LTBI, may not provide robust discrimination of TB from other pathologies and/or infectious diseases that may exhibit a similar clinical presentation. It is clear that IFN-inducible genes are dominant discriminators of active TB from healthy LTBI, leading to preferential selection of this gene set to define an optimal signature. However, this dominance precludes consideration of most other gene sets that may better inform discrimination from other diseases. This view is supported by differences in the reported signatures of Kaforou et al.[22] that were independently derived to discriminate active TB from LTBI, or active TB from other diseases. The 44-gene signature, derived using the latter approach, included more genes and exhibited greater diversity but lost discriminatory power, when compared

with the 27-gene signature for discriminating LTBI from TB. These observations suggest that a trade-off exists between these two objectives and that a single signature may not be optimal for both. In a clinical context, the two objectives of a TB signature fulfil distinct requirements. A signature that discriminates active TB from LTBI is a useful screening tool for testing in healthy populations. Identification of an active TB signature when screening for LTBI can inform the need for further investigation. In contrast, a signature that discriminates active TB from other diseases would be applied for the investigation of unwell patients presenting with symptoms that suggest the possibility of TB.

To tackle the challenge of developing a single high performing TB signature, we explored WGCNA as a tool for systematic gene reduction into biologically meaningful modules that together represent the entire transcriptome. We determined consistency of the modular signature for active TB across our cohorts and other published datasets and demonstrated that, when taking into consideration all 23 modules, the signature in active TB was distinct from both viral and bacterial infections. Furthermore, we additionally identified gene clusters that were differentially expressed in TB but not influenza from within the modules of IFN signalling. This was an important observation as the opportunity to select specific genes from these dominant modules offered scope to improve the sensitivity of the signature and discrimination of TB from LTBI and other diseases. Based on these findings we developed and evaluated a two-step approach for targeted gene selection to derive a TB signature. Modules perturbed in TB were first interrogated to establish a gene set comprising genes that are differentially expressed in TB compared with other diseases. *A priori* gene selection in this way provided a gene set with high TB specificity against other diseases. In the second step, traditional gene reduction methodology was applied to separate TB from LTBI using this gene set. As a proof of principle, we developed a 20-gene signature using this approach that was diverse in its representation, incorporating genes from 6 modules. We demonstrate here that this 20-gene signature had excellent sensitivity and specificity for discriminating active TB from LTBI, but did not discriminate either viral or bacterial infection from health, implying effectiveness to discriminate TB from other infections.

Heterogeneity of LTBI was suggested in our previous study[6] with the identification of an outlier group after clustering. In the present study, we identified a similar proportion of LTBI outliers in the new Leicester cohort. We demonstrated enrichment scores using the published signatures of Zak et al.[30] and Kaforou et al.[22] that were higher in outliers compared with other LTBI in both the Berry and Leicester cohorts, and overlapped with scores obtained in active TB. These observations suggested LTBI outliers are characterised by an overabundance of IFN-inducible genes, a view that was corroborated in their modular signatures demonstrating overabundance of the corresponding modules, together with identification of seventy selectively upregulated genes, common to both the Berry and Leicester LTBI outliers, which mapped to IFN signalling pathways. The clinical significance of these observations remains unclear, however the recent study of Zak et al.[30] suggests expression of a TB-like signature, characterised by enrichment of IFN-inducible genes which we show from our analysis, may indicate either subclinical disease or increased risk of progression to TB within a few months.

We utilised the modular signature for deeper characterisation of heterogeneity in recent TB contacts. We developed a weighted TB agreement score that provided a composite quantitative measure for defining the modular signature of TB. We identified instances of discordance for similarity with TB, between clustering and the modular weighted TB

agreement score that appeared to be driven by differences in the pattern of perturbation in modules other than those representing IFN signalling pathways, such that cumulative similarity in these less dominant modules was a significant determinant of the composite weighted agreement score. Our observations of low agreement scores for the IGRA^{-ve} cohort is consistent with the absence of LTBI and likely to reflect a robust finding. In contrast, enrichment scores of the published signatures we tested indicated considerably more overlap of IGRA^{-ve} subjects with the IGRA^{+ve} group, again indicating impaired specificity of these signatures.

We observed evidence of dynamic change in the modular signature of some TB contacts that can be categorised into three patterns of longitudinal expression that may reflect early immunological events following TB exposure. We suggest the absence of a signature at any time point may indicate the absence of infection being acquired. This pattern was seen in 60% of our IGRA^{-ve} cohort and 37% of our IGRA^{+ve} cohort. A transient signature may indicate an infection that was acquired but has either been controlled or cleared. In this context the observation that 25% of our IGRA^{-ve} cohort demonstrated this pattern suggests that the blood transcriptional signature represents immune responses that precede priming and activation of IFN-γ producing CD4 T-cells. Finally, subjects with an evolving and persistent modular TB signature may represent subjects that have acquired an infection requiring active control to maintain latency. This pattern was seen in 25% of IGRA^{+ve} and 15% of IGRA^{-ve} subjects. These observations require validation in larger longitudinal cohorts, but do suggest that the blood transcriptome may offer a more sensitive approach to characterising the state of latent infection following TB exposure, with implications for better stratification of prospective TB risk.

For our cohort of 9 subjects identified with TB during prospective observation, a high or rising TB agreement score was observed in the majority. This was most apparent in the subjects defined as true progressors. Our study was limited by small numbers and the identification of TB within a short period of prospective observation. We are therefore presently unable to comment on the dynamic properties of this response or determine the interval between the signature becoming detectable and manifestation of active TB. It is notable also that two subjects did not express a signature at any time point and yet went on to be diagnosed with TB. Interrogating the modules for these subjects indicates a weak transcriptional response that may suggest pathogen induced host immunomodulation, which is well recognised in active TB[13,41]. In keeping with this, we observed the TB agreement score dropping at the time of diagnosis in 4 subjects with evidence of prior signature expression.

In conclusion, a modular approach to characterising the blood transcriptional signature in active TB is robust and confers specificity for evaluating important clinical outcomes and heterogeneity in LTBI, and its use improves the development of reduced gene signatures for discriminating active TB from LTBI and other infections.

Methods

Study cohorts for analysis

Cohorts analysed in Berry et al. 2010[6] using microarrays were subjected to RNA-Seq and analysed as part of this study. Test and validation sets, termed Berry London and Berry South Africa sets, respectively, based on the geographical location of patient recruitment, were retained for RNA-seq analysis in this study (**Supplementary Figure 1a**).

An independent cohort was recruited (between 09/2015 and 09/2016) at the Glenfield Hospital, University Hospitals of Leicester NHS Trust, Leicester, UK. The cohort consisted of active TB patients (n=53) and recent close contacts (n=108). Patients who were pregnant, immunosuppressed, had previous TB or previous treatment for LTBI were excluded from this study. All participants had routine HIV testing and patients with a positive result were excluded. Patients with active TB were confirmed by laboratory isolation of *M. tuberculosis* on culture of a respiratory specimen (sputum or bronchoalveloar lavage) with sensitivity testing performed by the Public Health Laboratory Birmingham, Heart of England NHS Foundation Trust, Birmingham, UK. All recent close contacts were IGRA tested using the QuantiFERON Gold In-Tube Assay (Qiagen) and were subsequently categorised as either IGRA negative (n=50) or IGRA positive (n=49). All participants were prospectively enrolled and sampled before the initiation of any anti-mycobacterial treatment. A subset of subjects recruited initially as close contacts were identified with active TB during longitudinal assessment (n=9), based on microbiological confirmation of *M. tuberculosis* by culture or positive Xpert MTB/RIF (Cepheid). (**Supplementary Tables 2** and **8**; **Figure 7a**). The Research Ethics Committee (REC) for East Midlands – Nottingham 1, Nottingham, UK (REC 15/EM/0109) approved the study. All participants were older than 16 years and gave written informed consent.

Additional TB datasets were retrieved from Gene Expression Omnibus (GEO) that included datasets from Kaforou et al. 2013[22] (GEO accession: GSE37250) and Zak et al. 2016[30](GEO accession: GSE79362, BioProject PRJNA315611, SRA SRP071965) (**Supplementary Table 4**). Other datasets with additional diseases downloaded from GEO included Parnell et al. 2011[33] (GEO accession: GSE20346), Zhai et al. 2015[34] (GEO accession GSE68310), Herberg et al. 2013[36] (GEO accession: GSE42026), Suarez et al. 2015[37] (GEO accession: GSE60244) and Bloom et al. 2015[7] (GEO accession: GSE42834) (**Supplementary Table 4**).

RNA extraction, globin reduction, cDNA library preparation and RNA-Seq

3 ml whole blood were collected by venepuncture into Tempus™ blood RNA tubes (Fisher Scientific UK Ltd), tubes were mixed vigorously immediately after collection, and then stored in a −80°C freezer prior to use. Total RNA was isolated from 1 ml whole blood using the MagMAX™ for Stabilized Blood Tubes RNA Isolation Kit (Applied Biosystems/Thermo Fisher Scientific) according to the manufacturer's instructions. Globin RNA was depleted from total RNA (1.5–2 µg) using the human GLOBINclear kit (Thermo Fisher Scientific) according to manufacturer's instructions. RNA yield of total and globin-reduced RNA was assessed using a NanoDrop™ 8000 spectrophotometer (Thermo Fisher Scientific). Quality and integrity of total and globin-reduced RNA were assessed with the HT RNA Assay reagent kit (Perkin Elmer) using a LabChip GX bioanalyser (Caliper Life Sciences/Perkin Elmer) and

assigned an RNA Quality Score (RQS). Samples (200 ng) with an RQS > 6 were used to prepare a cDNA library using the TruSeq Stranded mRNA HT Library Preparation Kit (Illumina). The tagged libraries were sized and quantitated in duplicate (Agilent TapeStation system), using D1000 ScreenTape and reagents (Agilent), normalised, pooled and then clustered using the HiSeq® 3000/4000 PE Cluster Kit (Illumina). The libraries were imaged and sequenced on an Illumina HiSeq 4000 sequencer using the HiSeq® 3000/4000 SBS kit (Illumina) at a minimum of 25 million paired end reads (75 bp) per sample.

RNA-seq data analysis

Raw paired-end RNA-seq data obtained for Berry London, Berry South Africa and Leicester cohorts was processed separately and subjected to quality control using FastQC (Babraham Bioinformatics) and MultiQC[52]. Trimmomatic[53] v0.36 was used to remove adapters and filter raw reads below the 36 bases long and leading and trailing bases below quality 25. Filtered reads were aligned to the *Homo sapiens* genome Ensembl GRCh38 (release 86) using HISAT2[54] v2.0.4 with default settings and RF rna-strandedness including unpaired read reads resulting from Trimmomatic. Mapped and aligned reads were quantified to obtain gene-level counts using HtSeq[55] v0.6.1 with default settings and reverse strandedness. Raw counts were processed using the *bioconductor* package *edgeR*[56] v3.14.0 in R. Genes expressed with counts per million (CPM) >2 in at least 5 samples were considered and normalised using trimmed mean of M- values (TMM) to remove library-specific artefacts. Only protein coding genes were considered for subsequent analyses. Differentially abundant genes were calculated using likelihood ratio tests in edgeR by fitting generalized linear models to the non-normally distributed RNA-seq data. Genes with $\log_2$ fold change >1 or <-1 and false discovery rate (FDR) p-value < 0.05 corrected for multiple testing using the Benjamini-Hochberg (BH) method[57] were considered significant. For subsequent analysis, voom transformation was applied to RNA-seq count data to obtain normalized expression values on the $\log_2$ scale. For Berry Combined dataset, raw counts from Berry London and South Africa cohorts were combined as one dataset and processed in edgeR as described above and batch effects were removed from $\log_2$ expression values using surrogate variable analysis (sva) using the *bioconductor* package *sva*[58] in R. RNA-seq data obtained from Zak et al. 2016[30] in the SRA format were converted to fastq files using the SRA toolkit and processed as above.

Microarray data analysis

External microarray datasets retrieved from GEO as non-normalized matrices were processed in GeneSpring GX v14.8 (Agilent Technologies). Flags were used to filter out probe sets that did not result in a 'present' call in at least 10% of the samples, with the 'present' lower cut-off of 0.8. Signal values were then set to a threshold level of 10, $\log_2$ transformed, and per-chip normalised using 75th percentile shift algorithm. Next per-gene normalisation was applied by dividing each messenger RNA transcript by the median intensity of all the samples. The training, test and validation sets in Bloom et al. 2013[7] were combined and batch effects were removed using *sva*[58]. In Kaforou et al. 2013[22], HIV+/- groups were combined and analysed as one dataset. In all datasets, multiple probes mapping to the same gene were removed and the probe with the highest inter-quartile range across all samples was retained, to match with the RNA-seq data. Differentially expressed genes were identified using the *bioconductor* package *limma*[59] in R and only

genes with FDR p-value < 0.05 corrected for multiple testing using the BH method[57] were considered significant.

Gene signature enrichment analysis

Enrichment of TB gene signatures was carried out on a per sample basis using single sample Gene Set Enrichment Analysis (ssGSEA)[32] using the *bioconductor* package *gsva*[60] in R. Enrichment scores were obtained similar to those from Gene Set Enrichment Analysis (GSEA) but based on absolute expression rather than differential expression[32], to quantify the degree to which a gene set is over-represented in a particular sample.

Weighted gene co-expression network analysis (WGNCA)

Modular analysis was performed using the WGCNA package in R. Modules were constructed using the Berry Combined dataset (combined Berry London and South Africa sets) using 5,000 genes with highest covariance across all samples using $\log_2$ RNA-seq expression values. A signed weighted correlation matrix containing pairwise Pearson correlations between all genes across all samples was computed using a soft threshold of ß = 14 to reach a scale-free topology. Using this adjacency matrix, the Topological Overlap Measure (TOM) was calculated, which measures the network interconnectedness and used as input to group highly correlated genes together using average linkage hierarchical clustering. The WGCNA dynamic hybrid tree-cut algorithm[61] was used to detect network modules of co-expressed genes, with a minimum module size of 20. All modules were assigned a colour arbitrarily and annotated using Ingenuity Pathway Analysis (IPA) (QIAGEN Bioinformatics) and Literature Lab (Acumenta Biotech, Massachusetts, USA). For each module, module eigengene (ME) values were calculated, which represent the first principal component of a given module and summarize the gene abundance profile in that module. For each module, top 50 hub genes with high intramodular connectivity and a minimum correlation of 0.75 were calculated and exported into Cytoscape v3.4.0 to create interaction networks.

WGCNA module enrichment analysis

Fold enrichment for the WGCNA modules was calculated using quantitative set analysis for gene expression (QuSAGE)[62] using the *bioconductor* package *qusage* in R, to identify the modules of genes over- or under-expressed in a dataset compared to a control group. Linear mixed models were incorporated in the analysis using QGen algorithm in QuSAGE, and patients in datasets with repeated measures were modelled as random effects. Only modules with FDR p-value < 0.05 were considered significant. To test the modules in microarray datasets, only those modules with a >70% match in genes symbols was present in the microarray dataset. To obtain a modular profile of a disease group, single sample enrichment scores were calculated using ssGSEA and the average enrichment score of the control group was subtracted from the average enrichment score of the disease group. To obtain a modular profile on a single sample basis, average enrichment score of the control group was subtracted from the enrichment score of the sample.

Class prediction

For classifying patients as active TB or latent TB, the random forest algorithm was used in *caret*[63] package in R, using leave one out cross validation (LOOCV) over 1,000 iterations. For the Zak et al. 2016[30] and Kaforou et al. 2013[22] gene signatures, the Berry London set was used to train the model that was tested in Berry South Africa and Leicester cohorts to calculate the accuracy and the sensitivity and specificity of the gene signatures in classifying patient as active TB or latent TB. For the Zhai et al. 2015[34], the Influenza A group at Day 0 was randomly split into training (70%) and test (30%) sets to classify patients as infected with Influenza A or healthy controls.

In order to develop a TB-specific gene signature, only genes significantly differentially expressed in Berry London set and not in other flu cohorts were considered, from only those modules that were perturbed in TB (a module was considered perturbed in TB if it followed a similar profile (up or down compared to control) in at least 4 of the 5 TB datasets (Berry London, Berry South Africa, Leicester cohort, Kaforou et al. 2013[22] and Zak et al. 2016[30]), and given that for the 5th dataset the module did not reach significance when compared to control). These genes were then reduced using the *Boruta*[39] package in R. Boruta is a feature selection wrapper algorithm based on random forest and is particularly useful in biomedical applications as it captures features by incorporating the outcome variable. Next, the features identified as predictive using Boruta were ranked using the GINI score in random forest and the top 20 genes were selected. For classifying patients as active TB or latent TB, the random forest algorithm was used in *caret*[63] package in R, using LOOCV over 1,000 iterations. Each of the TB datasets was randomly split into training (70%) and test (30%) sets to classify patients as active TB or latent TB. For the Zhai et al. 2015[34], the Influenza A group at Day 0 was randomly split into training (70%) and test (30%) sets to classify patients as infected with Influenza A or healthy controls.

Modified Disease Risk Score

To test the TB-specific 20-gene signature, a modified version of the Disease Risk Score (DRS) established by Kaforou et al. 2013[22] was used. Briefly, the DRS is obtained from normalized data in a non-log space, by adding the total intensity of up-regulated transcripts and subtracting the total intensity of down-regulated transcripts from a gene signature. In this study, normalized CPM values were used for the RNA-seq data and non-log normalized expression values were used for microarray data. As part of the modification of the DRS, the absolute values of the total intensity of up-regulated transcripts and total intensity of down-regulated transcripts were added to obtain a composite score.

Weighted TB Agreement Score

This was a score developed to quantify the level of similarity of modular signatures between a test sample and a TB reference signature. The reference signature was defined according to the pattern of perturbation within individual modules that was consistent across all the TB dataset we examined. Individual modules that expressed perturbation in a single dataset not seen in the other datasets were assigned a signal of no perturbation. The pattern in the test set was compared with this reference for each module. Agreement with the reference was assigned a score of 1 for the given module and disagreement, defined as module perturbation in the opposing direction to the reference, was scored as

-1. For modules that failed to exhibit significant perturbation from control, a score of zero was assigned if the test sample did not agree with the reference. As the modules representing IFN signalling (yellow and light green) are dominant, agreement in perturbation with one or both of these modules was mandated for scores of the other modules to be valid. The weighted score was therefore calculated as the product of the sum of agreement scores for the 21 modules (excluding the yellow and lightgreen modules), with the score for the sum of agreement for the yellow and lightgreen modules. The sum of scores for the yellow and lightgreen modules was assigned as zero if negative. The final score was divided by 42 to normalise the scale between -1 and 1.

Deconvolution analysis

Deconvolution analysis for quantification of relative levels of distinct cell types on a per sample basis was carried out using CIBERSORT[64]. CIBERSORT estimates relative subsets of RNA transcripts using linear support vector regression. Cell signatures for 22 cell types were obtained using the LM22 database from CIBERSORT and grouped into 11 representative cell types. Fractions of cell types were compared across different groups using One-way ANOVA, and p-value < 0.05 was considered significant.

Data availability

Sequence data that support the findings of this study is being deposited in NCBI SRA *SRPXXXX*, under the BioProject code *PRJNAXXXXXX* (GEO accession: *GSEXXXX*). All other data that support the findings of this study are available upon request.

Acknowledgements

We acknowledge the Francis Crick Advanced Sequencing Facility, and Bioinformatics and Biostatistics Science Technology Platforms for their contribution to our sequencing processing. We acknowledge the NIHR Leicester Biomedical Research Centre for their support of the study at Leicester. The views expressed are those of the author(s) and not necessarily those of the NHS the NIHR or the Department of Health. We thank the patients for their participation. We thank Asmaà Fritah-Lafont for help in co-ordinating the meetings regarding the study. We thank Dr. Lúcia Moreira-Teixeira for reviewing the manuscript and for valuable discussion. AOG, CMG and AS were funded by The Francis Crick Institute, (Crick 10126; Crick 10468), which receives its core funding from Cancer Research UK, the U.K. Medical Research Council, and the Wellcome Trust; and the sequencing project by the Bioaster Microbiology Technology Institute, Lyon, France; Medical Diagnostic Discovery Department, bioMérieux SA, Marcy l'Etoile, France; and funded in part by Illumina Inc., San Diego, CA, USA. RV and JL, University of Leicester, were funded by Bioaster Microbiology Technology Institute, Lyon, France. RJW was supported by The Francis Crick Institute, (Crick 10128), which receives its core funding from Cancer Research UK, the U.K. Medical Research Council, and Wellcome; by Wellcome (104803; 203135); MRC South Africa under strategic health innovation partnerships; and NIH 019 AI 111276.

Author contributions

AOG and PH co-led the whole study; AOG, MPRB, PH, RV, GW, MR designed the study; RV and JL recruited TB, LTBI and contacts to the study for the Leicester cohort; CMG led and performed the RNA-Seq sample and raw data generation. RV and CMG helped to co-ordinate logistics of the study; RJW, PL, PL gave feedback and concrete discussion during the study; TT and MR contributed towards bioinformatics analysis; AS led and performed all the bioinformatics analysis; PH and GW contributed the weighted TB agreement score analysis. AOG, AS, RV and PH wrote the manuscript; all co-authors have read, reviewed and approved the paper.

ORCID IDs: 0000-0001-9845-6134 (AOG); 0000-0002-6941-3618 (AS)

References

1 World Health Organisation. Global Tuberculosis Report. (2017).

2 Vynnycky, E. & Fine, P. E. Lifetime risks, incubation period, and serial interval of tuberculosis. *Am J Epidemiol* **152**, 247-263 (2000).

3 Abu-Raddad, L. J. *et al*. Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. *Proc Natl Acad Sci U S A* **106**, 13980-13985, doi:10.1073/pnas.0901720106 (2009).

4 Barry, C. E., 3rd *et al*. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat Rev Microbiol* **7**, 845-855, doi:10.1038/nrmicro2236 (2009).

5 Esmail, H. *et al*. Characterization of progressive HIV-associated tuberculosis using 2-deoxy-2-[18F]fluoro-D-glucose positron emission and computed tomography. *Nat Med* **22**, 1090-1093, doi:10.1038/nm.4161 (2016).

6 Berry, M. P. *et al*. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466**, 973-977, doi:10.1038/nature09247 (2010).

7 Bloom, C. I. *et al*. Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PLOS One* **8**, e70630, doi:10.1371/journal.pone.0070630 (2013).

8 Yan, N. & Chen, Z. J. Intrinsic antiviral immunity. *Nat Immunol* **13**, 214-222, doi:10.1038/ni.2229 (2012).

9 McNab, F. *et al*. Type I IFN induces IL-10 production in an IL-27-independent manner and blocks responsiveness to IFN-γ for production of IL-12 and bacterial killing in Mycobacterium tuberculosis-infected macrophages. *J Immunol*, doi:10.4049/jimmunol.1401088 (2014).

10 McNab, F. W. *et al*. TPL-2-ERK1/2 signaling promotes host resistance against intracellular bacterial infection by negative regulation of type I IFN production. *J Immunol* **191**, 1732-1743, doi:10.4049/jimmunol.1300146 (2013).

11 Redford, P. S. *et al*. Influenza A virus impairs control of Mycobacterium tuberculosis coinfection through a type I interferon receptor-dependent pathway. *J Infect Dis* **209**, 270-274, doi:10.1093/infdis/jit424 (2014).

12 McNab, F., Mayer-Barber, K., Sher, A., Wack, A. & O'Garra, A. Type I interferons in infectious disease. *Nat Rev Immunol* **15**, 87-103, doi:10.1038/nri3787 (2015).

13 O'Garra, A. *et al*. The immune response in tuberculosis. *Annu Rev Immunol* **31**, 475-527, doi:10.1146/annurev-immunol-032712-095939 (2013).

14 Antonelli, L. R. *et al*. Intranasal Poly-IC treatment exacerbates tuberculosis in mice through the pulmonary recruitment of a pathogen-permissive monocyte/macrophage population. *J Clin Invest* **120**, 1674-1682, doi:10.1172/JCI40817 (2010).

15 Dorhoi, A. *et al*. Type I IFN signaling triggers immunopathology in tuberculosis-susceptible mice by modulating lung phagocyte dynamics. *Eur J Immunol* **44**, 2380-2393, doi:10.1002/eji.201344219 (2014).

16 Manca, C. *et al*. Virulence of a *Mycobacterium tuberculosis* clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-a/b. *Proc Natl Acad Sci U S A* **98**, 5752-5757, doi:10.1073/pnas.091096998 (2001).

17 Manca, C. *et al*. Hypervirulent M. tuberculosis W/Beijing strains upregulate type I IFNs and increase expression of negative regulators of the Jak-Stat pathway. *J Interferon Cytokine Res* **25**, 694-701, doi:10.1089/jir.2005.25.694 (2005).

18 Mayer-Barber, K. D. *et al*. Host-directed therapy of tuberculosis based on interleukin-1 and type I interferon crosstalk. *Nature* **511**, 99-103, doi:10.1038/nature13489 (2014).

19 McNab, F. W. *et al*. Type I IFN induces IL-10 production in an IL-27-independent manner and blocks responsiveness to IFN-gamma for production of IL-12 and bacterial killing in Mycobacterium tuberculosis-infected macrophages. *J Immunol* **193**, 3600-3612, doi:10.4049/jimmunol.1401088 (2014).

20 Ordway, D. *et al*. The hypervirulent Mycobacterium tuberculosis strain HN878 induces a potent TH1 response followed by rapid down-regulation. *J Immunol* **179**, 522-531 (2007).

21 Joosten, S. A., Fletcher, H. A. & Ottenhoff, T. H. A helicopter perspective on TB biomarkers: pathway and process based analysis of gene expression data provides new insight into TB pathogenesis. *PLOS One* **8**, e73230, doi:10.1371/journal.pone.0073230 (2013).

22 Kaforou, M. *et al*. Detection of tuberculosis in HIV-infected and-uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLOS Med* **10**, e1001538 (2013).

23 Maertzdorf, J. *et al*. Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes and immunity* **12**, 15-22, doi:10.1038/gene.2010.51 (2011).

24 Ottenhoff, T. H. *et al*. Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis. *PLOS One* **7**, e45839, doi:10.1371/journal.pone.0045839 (2012).

25 Roe, J. K. *et al*. Blood transcriptomic diagnosis of pulmonary and extrapulmonary tuberculosis. *JCI Insight* **1**, e87238, doi:10.1172/jci.insight.87238 (2016).

26 Walter, N. D. *et al*. Blood Transcriptional Biomarkers for Active Tuberculosis among Patients in the United States: a Case-Control Study with Systematic Cross-Classifier Evaluation. *J Clin Microbiol* **54**, 274-282, doi:10.1128/JCM.01990-15 (2016).

27 Walter, N. D., Reves, R. & Davis, J. L. Blood transcriptional signatures for tuberculosis diagnosis: a glass half-empty perspective. *Lancet Respir Med* **4**, e28, doi:10.1016/S2213-2600(16)30038-8 (2016).

28 Blankley, S. *et al*. A 380-gene meta-signature of active tuberculosis compared with healthy controls. *Eur Respir J* **47**, 1873-1876, doi:10.1183/13993003.02121-2015 (2016).

29 Blankley, S. *et al*. The Transcriptional Signature of Active Tuberculosis Reflects Symptom Status in Extra-Pulmonary and Pulmonary Tuberculosis. *PLOS One* **11**, e0162220, doi:10.1371/journal.pone.0162220 (2016).

30 Zak, D. E. *et al*. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *The Lancet* **387**, 2312-2322 (2016).

31 Diel, R., Loddenkemper, R. & Nienhaus, A. Evidence-based comparison of commercial interferon-gamma release assays for detecting active TB: a metaanalysis. *Chest* **137**, 952-968, doi:10.1378/chest.09-2350 (2010).

32 Barbie, D. A. *et al*. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108 (2009).

33 Parnell, G. *et al*. Aberrant cell cycle and apoptotic changes characterise severe influenza A infection–a meta-analysis of genomic signatures in circulating leukocytes. *PLOS One* **6**, e17186 (2011).

34 Zhai, Y. *et al*. Host transcriptional response to influenza and other acute respiratory viral infections–a prospective cohort study. *PLOS Pathog* **11**, e1004869 (2015).

35 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).

36 Herberg, J. A. *et al*. Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes. *J Infect Dis* **208**, 1664-1668 (2013).

37 Suarez, N. M. *et al*. Superiority of transcriptional profiling over procalcitonin for distinguishing bacterial from viral lower respiratory tract infections in hospitalized adults. *J Infect Dis* **212**, 213-222 (2015).

38 Friedman, J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38**, 367-378 (2002).

39 Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J Stat Softw* **36**, 1-13 (2010).

40 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301-320 (2005).

41 Cooper, A. M. Cell-mediated immune responses in tuberculosis. *Annu Rev Immunol* **27**, 393-422, doi:10.1146/annurev.immunol.021908.132703 (2009).

42 Altare, F. *et al*. Impairment of mycobacterial immunity in human interleukin-12 receptor deficiency. *Science (New York, N.Y.)* **280**, 1432-1435 (1998).

43 Casanova, J. L. & Abel, L. Genetic dissection of immunity to mycobacteria: the human model. *Annu Rev Immunol* **20**, 581-620, doi:10.1146/annurev.immunol.20.081501.125851 (2002).

44 de Jong, R. *et al*. Severe mycobacterial and Salmonella infections in interleukin-12 receptor-deficient patients. *Science (New York, N.Y.)* **280**, 1435-1438 (1998).

45 Fortin, A., Abel, L., Casanova, J. L. & Gros, P. Host genetics of mycobacterial diseases in mice and men: forward genetic studies of BCG-osis and tuberculosis. *Annu Rev Genomics Hum Genet* **8**, 163-192, doi:10.1146/annurev.genom.8.080706.092315 (2007).

46 Jouanguy, E. *et al*. A human IFNGR1 small deletion hotspot associated with dominant susceptibility to mycobacterial infection. *Nature genetics* **21**, 370-378, doi:10.1038/7701 (1999).

47 Newport, M. J. *et al*. A mutation in the interferon-gamma-receptor gene and susceptibility to mycobacterial infection. *The New England journal of medicine* **335**, 1941-1949, doi:10.1056/nejm199612263352602 (1996).

48 Cliff, J. M., Kaufmann, S. H., McShane, H., van Helden, P. & O'Garra, A. The human immune response to tuberculosis and its treatment: a view from the blood. *Immunol Rev* **264**, 88-102, doi:10.1111/imr.12269 (2015).

49 Bloom, C. I. *et al*. Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy. *PLOS One* **7**, e46191, doi:10.1371/journal.pone.0046191 (2012).

50 Cliff, J. M. *et al*. Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *J Infect Dis* **207**, 18-29, doi:10.1093/infdis/jis499 (2013).

51 Lee, S. W. *et al*. Time interval to conversion of interferon-gamma release assay after exposure to tuberculosis. *Eur Respir J* **37**, 1447-1452, doi:10.1183/09031936.00089510 (2011).

52 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).

53 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).
54 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).
55 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).
56 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).
57 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300 (1995).
58 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. SVA: surrogate variable analysis. *R package version* **3** (2013).
59 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).
60 Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7, doi:10.1186/1471-2105-14-7 (2013).
61 Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719-720, doi:10.1093/bioinformatics/btm563 (2008).
62 Yaari, G., Bolen, C. R., Thakar, J. & Kleinstein, S. H. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Res* **41**, e170-e170 (2013).
63 Kuhn, M. Caret: classification and regression training. *Astrophysics Source Code Library* (2015).
64 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457, doi:10.1038/nmeth.3337 (2015).

Figures

Figure 1. Whole-blood transcriptional gene signatures in TB. a Heatmaps depicting unsupervised hierarchical clustering of active TB (red), LTBI (black) and control samples (purple) using a 373-gene signature derived using Berry London cohort and tested in Berry South Africa cohort, and **(b)** validated in an independent Leicester cohort. Gene expression values were averaged and scaled across the row to indicate the number of standard deviations above (red) or below (blue) the mean, denoted as row Z-score. **c** Receiver operating characteristic curves depicting the predictive potential of the 27-gene (TB vs. LTBI) and 44-gene (TB vs. other diseases (OD)) signatures from Kaforou et al.[22] and the 16-gene signature from Zak et al.[30], in classifying a sample as TB or LTBI. **d** Bar graphs depicting enrichment scores derived from ssGSEA for active TB, LTBI and control samples from Berry London, Berry South Africa and Leicester cohorts using the 16-gene signature from Zak et al[30]. Purple, black and red bars represent control, LTBI and active TB samples, respectively, and * (control outliers), # (LTBI outliers) and § (active TB outliers) represent outlier samples.

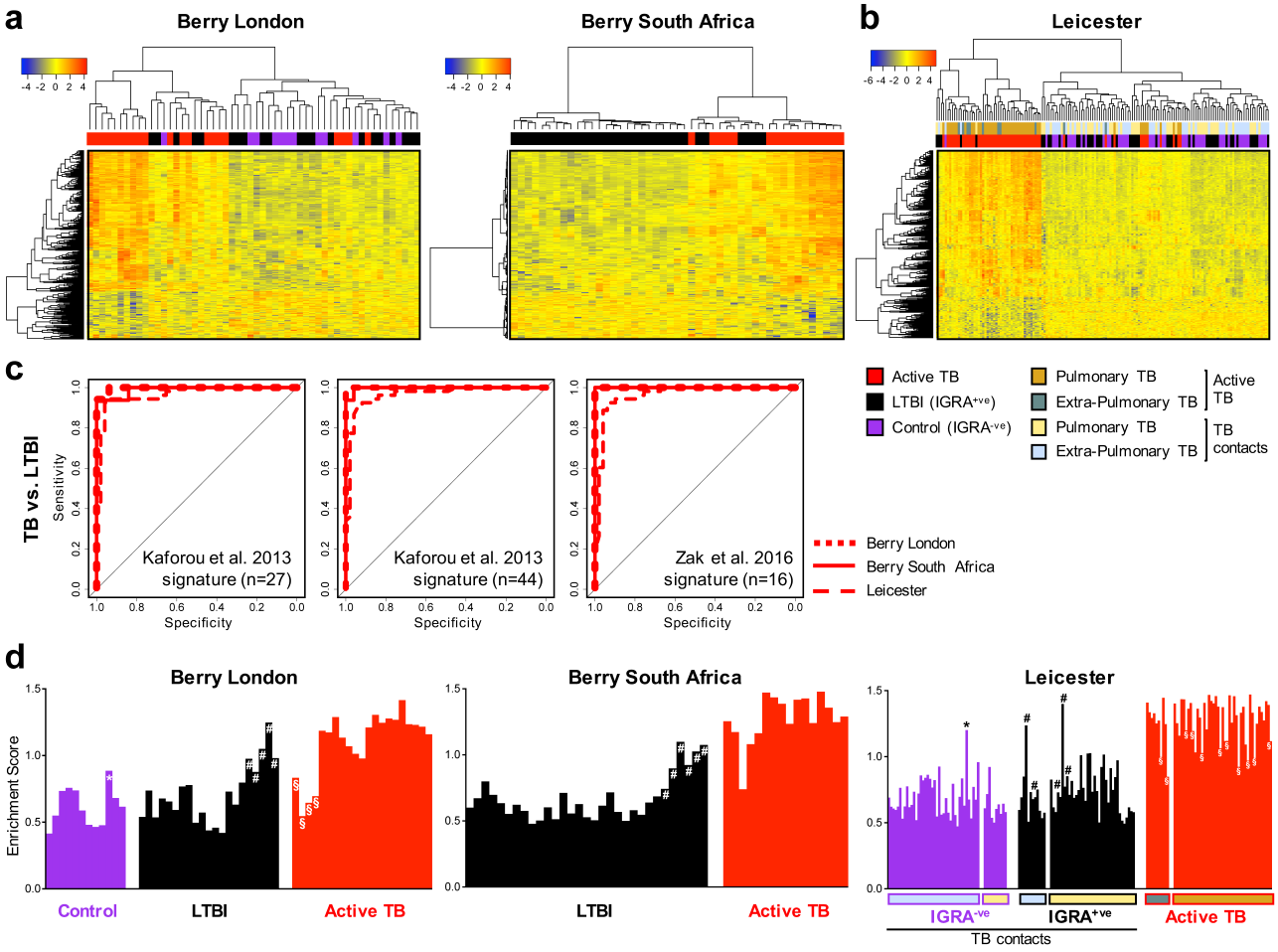

Figure 2. Enrichment of transcriptional gene signatures in TB and other viral and bacterial infections. a Box plots depicting enrichment scores derived from ssGSEA using the 16-gene signature from Zak et al.[30], and the 27-gene (TB vs. LTBI) and 44-gene (TB vs. other diseases (OD)) signatures from Kaforou et al.[22] in tuberculosis datasets (Berry London, Berry South Africa and Leicester), and in datasets with additional diseases – **(b)** severe influenza cohort from Parnell et al.[33], **(c)** Influenza A cohort from Zhai et al.[34] with **(d)** receiver operating characteristic curves depicting the predictive potential of these signatures in classifying a sample as influenza A or control from Zhai et al.[34], and **(e)** box plots for bacterial pneumonia cohort from Parnell et al.[33].

Figure 2.

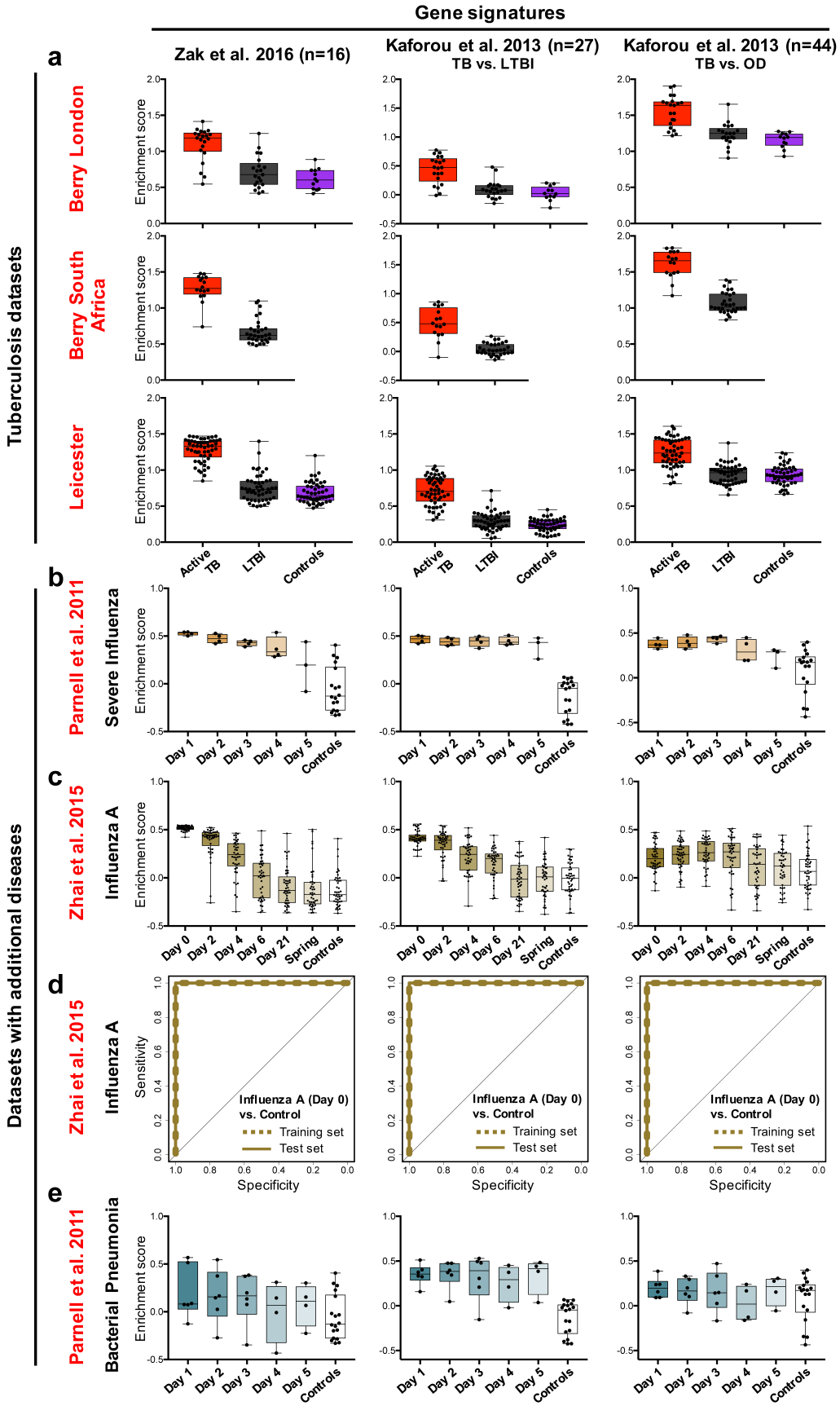

Figure 3. Modular transcriptional profiles of TB and other diseases. a Twenty-three modules of co-expressed genes derived using WGCNA from Combined Berry dataset (London & South Africa) were tested in TB datasets and **(b)** datasets with additional diseases. QuSAGE fold enrichment scores are depicted with red and blue indicating modules over- or under-expressed compared to controls. Colour intensity and size represent the degree of enrichment. Only modules with FDR p-value < 0.05 were considered significant and are depicted here. The enrichment scores for the lightgreen module for severe influenza cohort from Parnell et al.[33] and influenza A cohort from Zhai et al.[34] were greater than the maximum score depicted on the scale (i.e. > 1.3), and the actual scores are listed on the module.

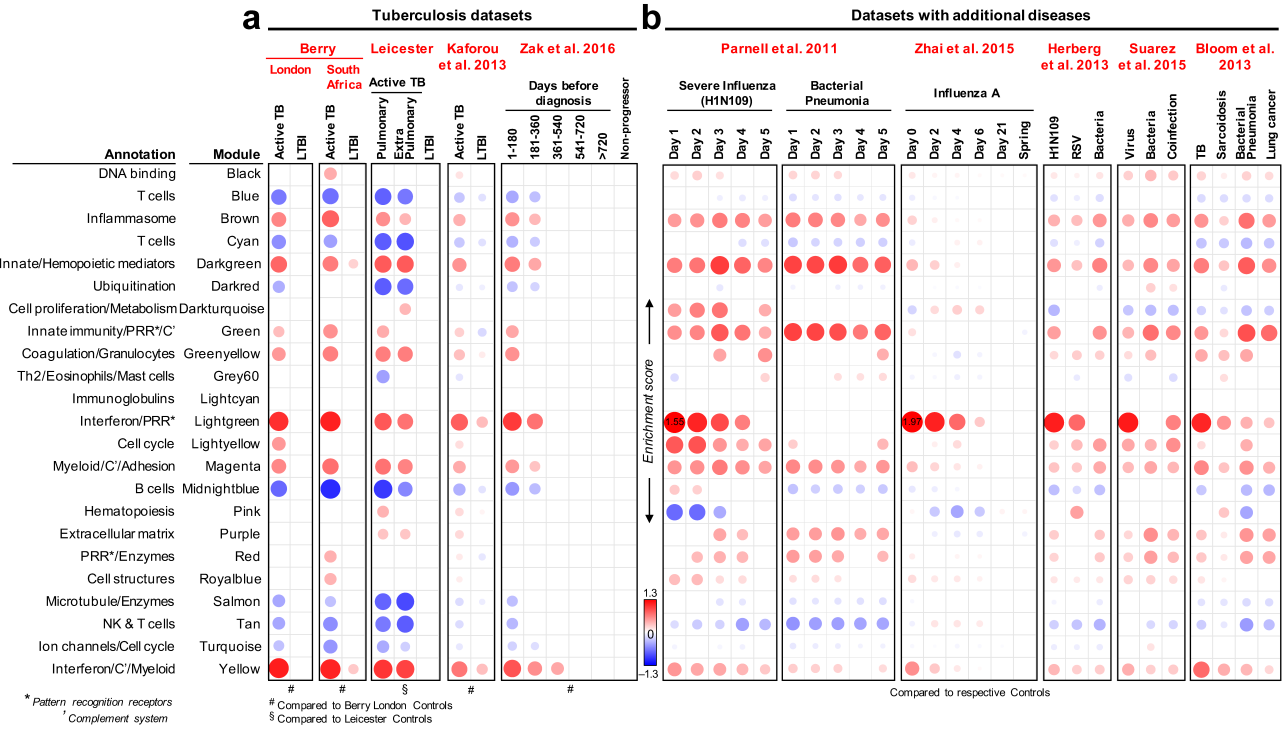

Figure 4. Gene expression in TB compared to other viral and bacterial infections. a Log_2 fold changes for genes in the yellow module from Berry London cohort derived from active TB vs. controls (*y-axis*) compared to log_2 fold changes derived from case vs. controls from other datasets (*x-axis*) in TB, and **(b)** datasets with other diseases (Herberg et al.[36], Suarez et al.[37], and time-course data from influenza A from Zhai et al.[34], and severe influenza and bacterial pneumonia from Parnell et al.[33]). Shapes and colours represent significance associated with the fold changes (FDR p-value < 0.05) in either Berry London only (orange squares), respective dataset only (cyan diamonds), both (yellow circles) or neither (black triangles).

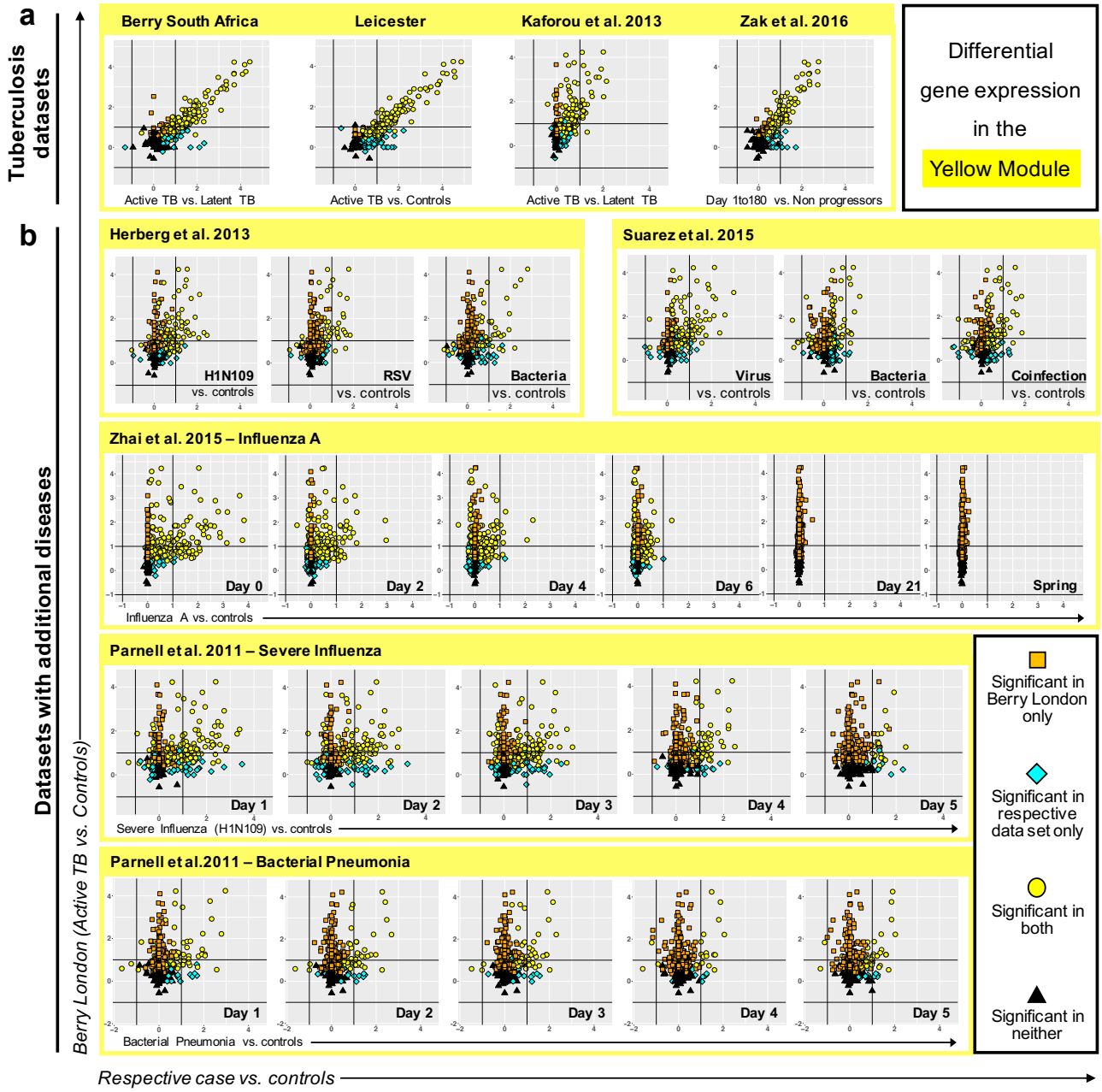

Figure 5. Whole-blood TB-specific 20-gene signature. a A 20-gene signature derived from genes significantly differentially expressed in Berry London cohort only, and not in other flu datasets **(Supplementary 5b)**. **b** Box plots depicting the modified Disease Risk Scores in tuberculosis datasets and a receiver operating characteristic curve depicting the predictive potential of this 20-gene signature, in classifying a sample as TB or LTBI. **c** Box plots depicting the modified Disease Risk Scores in datasets with additional diseases and a receiver operating characteristic curve depicting the predictive potential of this 20-gene signature, in classifying a sample as influenza A or control.

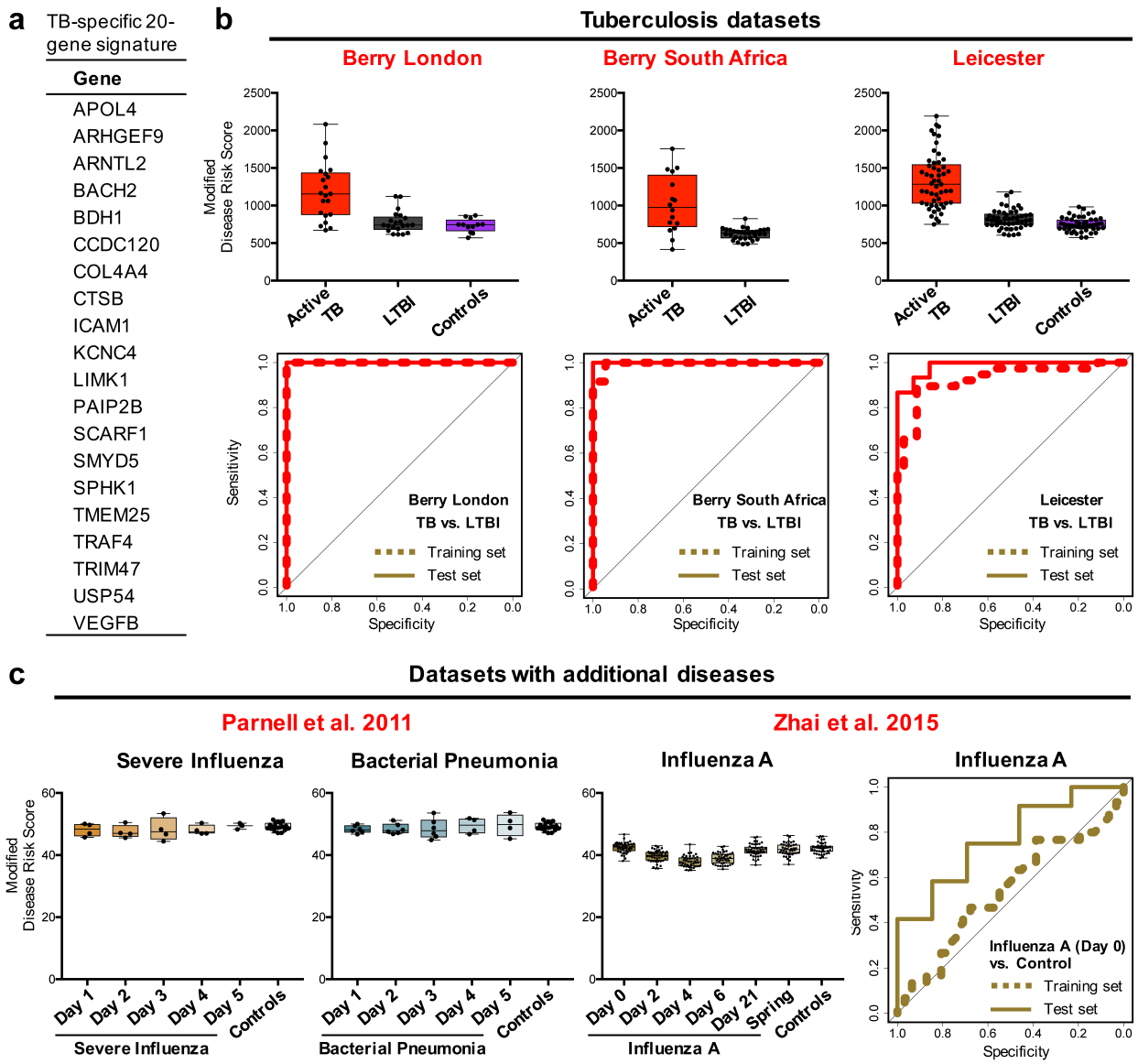

Figure 6. Transcriptional profiles of LTBI outliers. a Twenty-three modules of co-expressed genes derived using WGCNA from Combined Berry dataset (London & South Africa) and tested in Combined Berry dataset and Leicester cohort, with LTBI outliers as a separate group. QuSAGE fold enrichment scores are depicted with red and blue indicating modules over- or under-expressed compared to controls. Colour intensity and size represent the degree of enrichment. Only modules with FDR p-value < 0.05 were considered significant and are depicted here. **b** Interaction networks depicting the top 50 hub genes with high intramodular connectivity for the yellow, lightgreen and **(c)** tan modules. Each gene is represented as a square node with edges representing correlation between the gene expression profiles of the two genes (minimum Pearson correlation of 0.75). A key describing the four different partitions within each square node is shown, with each partition representing $\log_2$ fold changes for active TB (without outliers) and LTBI–Outliers from Berry Combined and Leicester cohorts, compared to respective controls (without outliers). Red and blue represent up- and down-regulated genes, respectively. In the tan module, the expression for IFNG is also shown, although it was not one of the top 50 hub genes. Boxplots depicting module eigengene expression, i.e. the first principal component, are also shown for the yellow, lightgreen and tan modules for samples from Berry Combined and Leicester datasets. **d** Volcano plots depicting differentially expressed genes for active TB (without outliers) and LTBI–Outliers compared to LTBI (without outliers) in the Berry Combined and Leicester datasets. Significantly differentially expressed genes ($\log_2$ fold change >1 or <-1, and FDR p-value < 0.05) are represented as red (upregulated) or blue (downregulated) dots, along with a Venn diagram and table summarising overlaps between these different comparisons.

Figure 6.

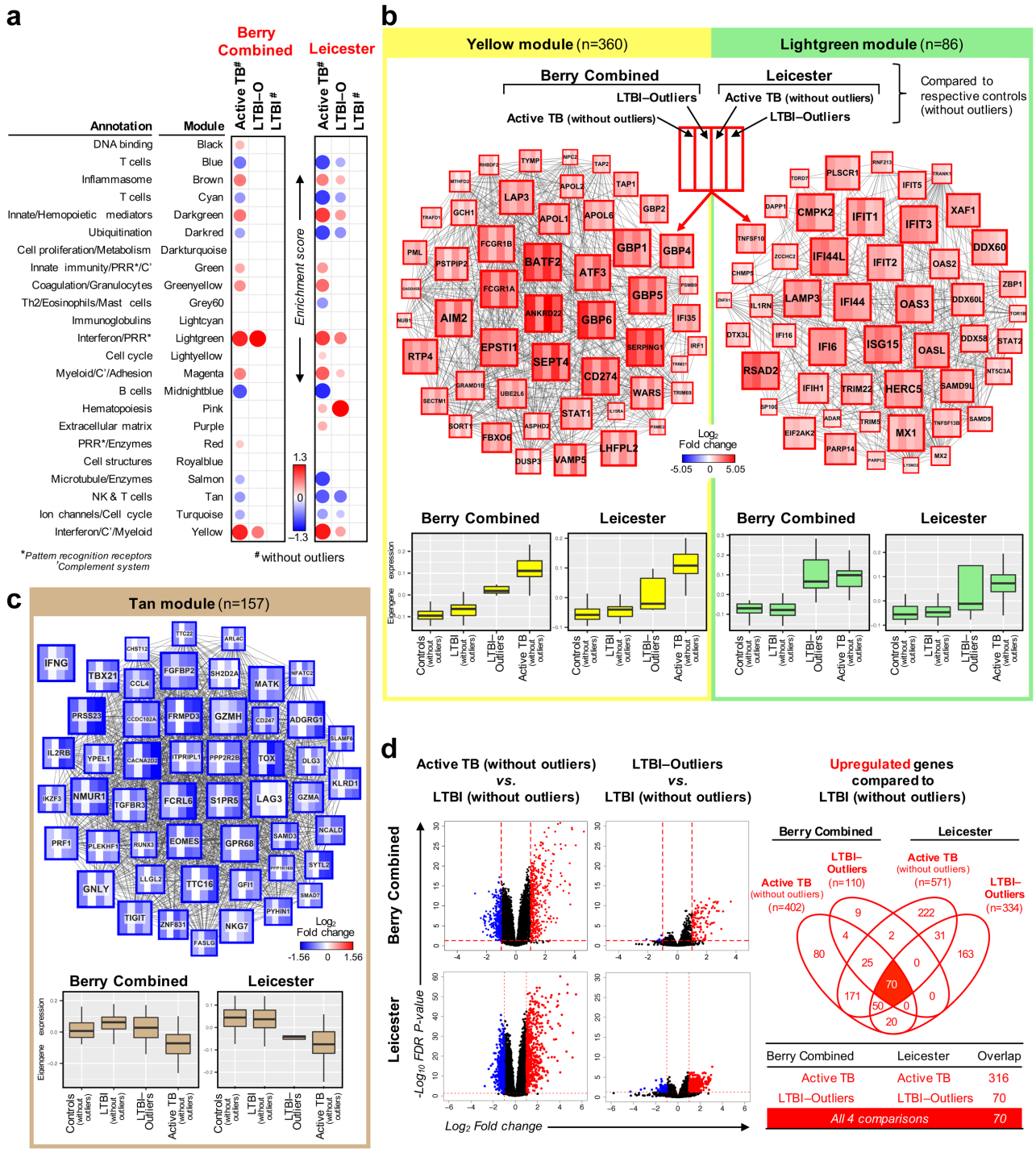

Figure 7. Modular transcriptional profiles of TB contacts followed over time. a Schematic depicting active TB patients and their contacts followed over time in the Leicester cohort. Purple, black and red represent $IGRA^{-ve}$ (controls), $IGRA^{+ve}$ (LTBI) and active TB patients, respectively. **b** Longitudinal modular profiles of TB contacts who remained $IGRA^{-ve}$ without developing TB (n=15), **(c)** TB contacts who remained $IGRA^{+ve}$ without developing TB (n=16), and **(d)** TB contacts who developed TB during the study (n=9). Enrichment scores derived using ssGSEA compared to the average enrichment scores of $IGRA^{-ve}$ controls are depicted, with red and blue indicating modules over- or under-expressed compared to controls. Colour intensity and size represent the degree of enrichment. For each patient time course data is depicted with a sample at baseline followed by months from baseline. For the contacts who developed TB during the study, the time point when the contact was diagnosed with active TB in the clinic is represented in red letters. Representative modular TB profiles depicting modules that were perturbed in TB, are shown for visual comparison on either side of each modular figure. The x-axis depicts the time in months of recruitment for the study from Baseline.

Figure 7.

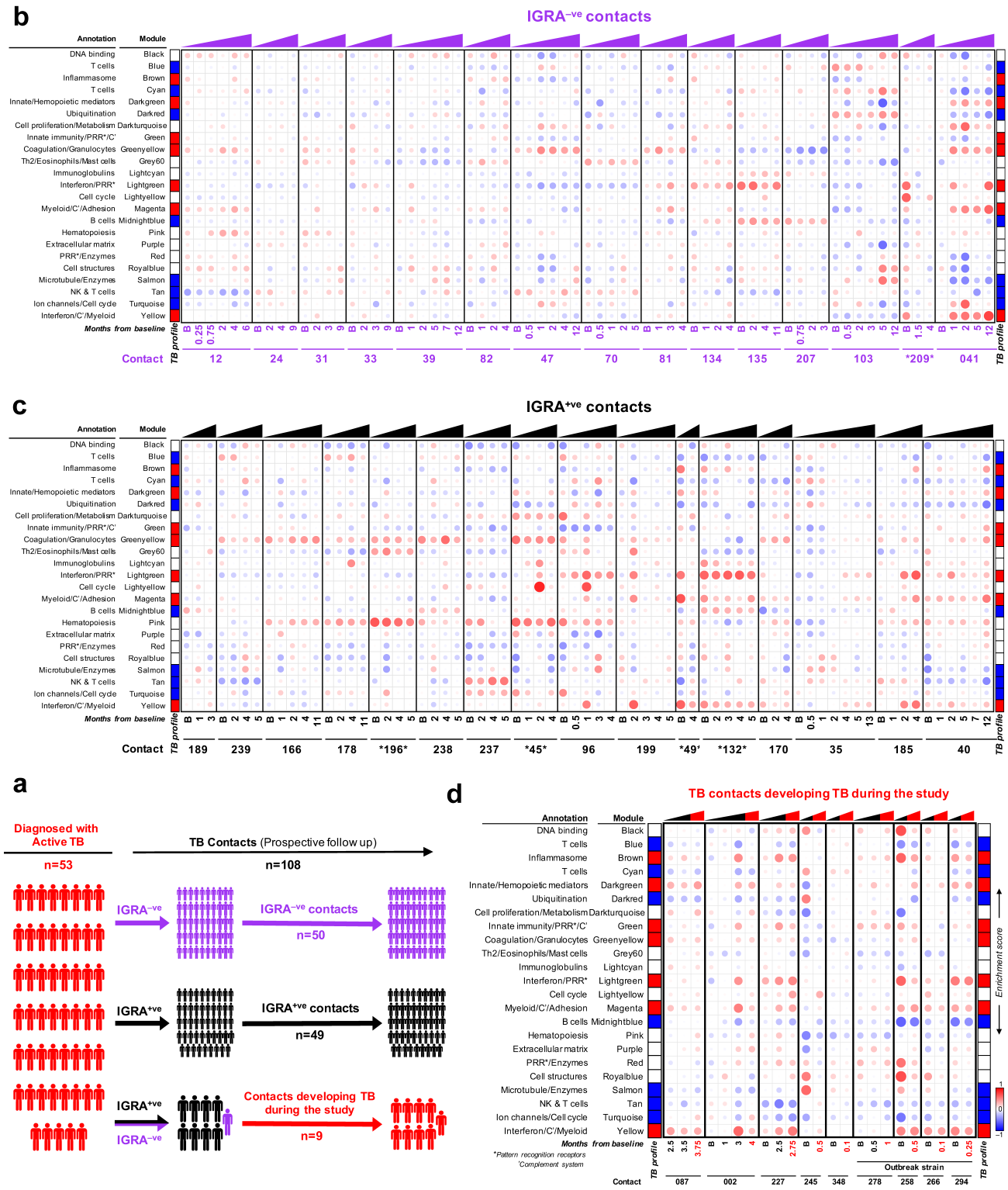

Figure 8. Whole-blood weighted-TB agreement score in TB contacts followed over time. **a** Bar plots depicting the weighted-TB agreement scores in TB contacts who remained IGRA^{-ve} (n=15), **(b)** TB contacts who remained IGRA^{+ve} (n=16), and **(c)** TB contacts who developed TB during the study (n=9). For TB contacts who developed TB during the study, the time point when the contact was diagnosed with active TB in the clinic is represented by a red bar.