Unravelling the transcriptome of the human tuberculosis lesion and its clinical implications

Albert Despuig[1,2,&], Kaori L. Fonseca[1,2,&], Dominic Habgood-Coote[3], Álvaro Del Río-Álvarez[4], Juan Carrillo-Reixach[4,5], Lilibeth Arias[1,2], Nino Gogichadze[1,6], Aaron Goff[7], Leticia Muraro Wildner[7], Shota Gogishvili[8], Keti Nikolaishvili[8], Natalia Shubladze[8], Zaza Avaliani[8], Pere-Joan Cardona[1,2,9], Federico Martinón-Torres[2,10,11], Antonio Salas[2,11,12], Alberto Gómez-Carballa[2,10,11], Simon J Waddell[7], Carolina Armengol[4,5], Sergo Vashakidze[8,13], Myrsini Kaforou[3,14], Cristina Vilaplana[1,2,15]*

1. Experimental Tuberculosis Unit (UTE). Fundació Institut Germans Trias i Pujol (IGTP). Universitat Autònoma de Barcelona (UAB). Edifici Mar. Can Ruti Campus. Crtra. de Can Ruti, Camí de les Escoles, s/n. 08916, Badalona; Catalonia; Spain.
2. Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES). Av. Monforte de Lemos, 3-5. Pabellón 11. Planta 0. 28029, Madrid, Spain.
3. Department of Infectious Disease, Imperial College London, London, United Kingdom.
4. Childhood Liver Oncology Group, Germans Trias i Pujol Research Institute (IGTP), Program of Predictive and Personalized Medicine of Cancer (PMPCC), Edifici Muntanya. Can Ruti Campus. Ctra. de Can Ruti, camí de les Escoles, s/n, 08916 Badalona, Catalonia, Spain.
5. Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREhD). Av. Monforte de Lemos, 3-5. Pabellón 11. Planta 0. 28029, Madrid, Spain.
6. UCBL, UnivLyon, Université Claude Bernard Lyon 1 (UCBL1), 69100 Villeurbanne, France
7. Global Health and Infection, Brighton and Sussex Medical School, University of Sussex, Falmer, Brighton, BN1 9PX, UK.
8. National Center for Tuberculosis and Lung Diseases (NCTLD). 50, Maruashvili Str. 0101 Tbilisi, Georgia.
9. Hospital Universitari Germans Trias i Pujol (HUGTIP). Microbiological dept. Crtra. Del Canyet, s/n. 08916. Badalona, Catalonia; Spain.
10. Genetics, Vaccines and Pediatric Infectious Diseases Research Group (GENVIP), Instituto de Investigación Sanitaria de Santiago (IDIS) and Universidad de Santiago de Compostela (USC), 15706, Galicia, Spain.
11. Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela (SERGAS), 15706, Galicia, Spain.
12. Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), 15706, Galicia, Spain.
13. The University of Georgia, Tbilisi, Georgia.
14. Centre for Paediatrics and Child Health, Imperial College London, United Kingdom
15. Direcció Clínica Territorial de Malalties Infeccioses i Salut Internacional. Gerència territorial Metropolitana Nord, Institut Català de la Salut. Can Ruti Campus. Crtra. del Canyet, s/n. 08916, Badalona; Catalonia; Spain

Corresponding Author*:
Cristina Vilaplana, MD, PhD
ORCID: https://orcid.org/0000-0002-2808-7270
Principal Investigator
Tuberculosis Experimental Unit (UTE)
Fundació Institut Germans Trias i Pujol (IGTP)
Campus Can Ruti. Crtra. De Can Ruti, Camí de les Escoles, s/n.
Badalona, Barcelona, 08916
Catalonia (Spain)
Telephone: +34 93 033 0527
cvilaplana@igtp.cat

[&]These authors contributed equally: Albert Despuig and Kaori L. Fonseca

ABSTRACT

Tuberculosis (TB) disease causes up to 1.5 million deaths every year and represents an important problem of public health at worldwide level. Here, we quantified the gene expression signatures of granuloma biopsies across human TB pulmonary lesions, and validated the best gene candidates using NanoString technology, profiling 157 samples from 40 TB patients who underwent surgery.

We characterised the transcriptional profile of the TB granuloma in comparison to healthy tissue, described an 11-gene signature and measured 7 proteins in plasma associated with it. We demonstrated a gradient of immune-related transcript abundance across the granuloma substructure and evidenced metabolically-active *Mycobacterium tuberculosis* in the lesions. Patients who converted to sputum negative after two months of starting treatment, showed enriched inflammatory pathways in the lesion several months after, supporting use of sputum culture conversion (SCC) as a prognostic biomarker during clinical management and as a factor to prioritise patients when considering lung surgery.

INTRODUCTION

Tuberculosis (TB) is an infectious disease caused by a single agent – *Mycobacterium tuberculosis* (*Mtb*). TB is a major cause of ill-health and mortality in many countries, being responsible of up to 10.5 million deaths each year worldwide[1]. Recent COVID-19 pandemic has severely impacted TB diagnosis and treatment, halting the progress made to control TB[2].

Globally, TB chemotherapy is successful in 85% of known drug sensitive cases[3]. Nevertheless, there is a fraction of patients who will fail treatment, who are prone to disease relapse and death, which results in 1.5 million deaths each year[1]. Multi-drug and extensively-drug resistant (MDR/XDR) TB cases have a higher rate of negative outcomes compared to drug-sensitive TB. In some countries with a high prevalence of MDR/XDR-TB, adjunctive surgical resection is a common therapeutic tool. Despite the major invasive procedure, the removal of cavities by therapeutic surgery reduces the transmission burden of MDR/XDR-TB and results in favourable outcomes for the patients[4]. The therapeutic resection of human pulmonary lesions has revealed insights into TB lesion architecture, the site of TB disease and immunopathology, that may contribute to the emergence of drug-resistant *Mycobacterium tuberculosis* (*Mtb)* populations[5,6]. Moreover, mycobacterial culture from resected granuloma tissue demonstrated that a subset of patients still harboured live *Mtb* bacilli despite showing preoperative microbiological clearance in sputa in both MDR/XDR-TB and drug-sensitive TB (DS-TB) patients[5,7].

Clinically relevant biomarkers, which can be used to predict TB clinical course and treatment outcome or to act as surrogate markers of sputum culture conversion (SCC), are urgently needed to improve its current clinical management. Time to SCC is currently used as a biomarker of sterilizing activity, although it does not correlate with the

persistence of the bacilli within the lesion and does not reflect the health status of the patient[8]. Host blood RNA-derived biomarker signatures, especially those derived from RNA-Sequencing (RNA-Seq), have emerged as promising tools to overcome these limitations[9]. RNA signatures from patients with pulmonary TB, have shown potential in disease diagnosis and prognosis to monitor treatment efficacy and/or predict failure; they also show some correlation to the extent of lung pathology[9,10]. However, literature regarding the RNA-Seq derived transcriptome of the human TB lesion biopsies is scarce[11] and host factors that drive cavitary lesion formation or trigger poor clinical outcomes remain unknown.

Based on the hypothesis that characterising human TB lesions at the transcriptomic level will drive novel biomarker and therapeutics research, we explored, quantified and validated the gene expression signatures of granuloma biopsies across a large collection of human TB pulmonary lesions and confirmed the best gene candidates using NanoString technology, profiling a total of 157 samples from 40 TB patients who underwent surgery.

RESULTS

Identification of a gene signature of the human TB lesion

We analysed total RNA from three different sections of the human TB granuloma from surgical biopsies (Central Lesion (CL), Internal Wall (IW), External Wall (EW)) and Pulmonary Nodules (PN). Additionally, healthy parenchyma (HT) from the same involved lung was collected as tissue control (Fig. 1a). Patients were divided into a Discovery Set (14 patients) and a Validation Set (26 patients) (Supplementary Figure 1). The Discovery Set comprised 61 samples from the 14 patients and was employed to characterise genome-wide gene expression across TB lesions. (Supplementary Table 1). The patients were matched according to sex and drug sensitivity of the *Mtb* isolate (Supplementary Table 2). The associations of the demographic, clinical and TB-related data of each participant in the Discovery Set at the time of surgery, and sex were assessed (Supplementary Table 3). Clinical and demographic features of the participants and removed TB lesions between Discovery and Validation Sets were also evaluated (Supplementary Table 4).

To identify genes differentially expressed between the TB granuloma and surrounding healthy lung tissue, we performed differential expression analysis between the granuloma (CL, IW, EW) and HT. We found a total of 6,056 significantly differentially expressed genes (DEGs). Of these, 3,405 genes were over-expressed in lung lesion tissues, whereas 2,651 were under-expressed compared to healthy lung tissue (Supplementary Data Set 1). The top 40 DEGs showed a clear separation between the granuloma samples and the HT samples (Fig.1b). Only 10 granuloma-derived samples including 7 of the EW clustered with HT samples. This may reflect contamination with HT tissue due to the fact the HT is just adjacent to the EW, or could be justified by the well-accepted variability of granulomas within the same individual[12,13].

Next, we performed Gene-Set-Enrichment Analysis (GSEA) for the 6,056 DEGs, which revealed 22 Hallmark gene sets significantly enriched in the TB granuloma lesion (Supplementary Data Set 2). The most significant gene sets found to be overexpressed in the TB granuloma were mainly related to an immune inflammatory response (i.e. allograft rejection, interferon gamma response, inflammatory response, IL6/JAK/STAT3, TNFalpha/NFKb, IL2/STAT5 signalling hallmarks) and cellular proliferation (i.e. E2F targets and G2M checkpoint hallmarks) (Fig. 1c and d; Supplementary Data Set 2).

A total of 11 genes specific to the TB lesion (*MMP-3, LTB, LTF, FCMR, CCL-19, MMP-9, IL2RG, CD14, MAP4K1, CXCL-10* and *CD5L*) were identified for validation in an independent set of 96 tissue samples from 26 TB patients (Validation set) through nCounter technology (NanoString) (Fig.2a, Supplementary Figure 2, Supplementary Table 5). The 11 genes characteristic of the TB lesion were significantly differentially expressed in the lesion *vs* healthy lung tissue comparison ($p < 0.0001$). Moreover, $\log_2$FC values followed the same gene expression directionality pattern in both Discovery and Validation sets and showed a positive significant correlation ($r = 0.94$; $p < 0.0001$), confirming the reproducibility of the results using a different technology (Fig.2b, Supplementary Tables 6 and 7). Altogether, our results show a TB lesion 11-gene signature with *MMP-3, LTB, LTF, FCMR, CCL-19, MMP-9, IL2RG, CD14, MAP4K1* and *CXCL-10* being over-expressed and *CD5L* under-expressed in the granuloma compared with the healthy tissue.

Seven proteins of the TB lesion 11-gene signature were found in plasma

To further evaluate if proteins encoded by the TB lesion 11-gene signature could mirror the ongoing tissue pathology in circulation, we studied further 7 secreted proteins corresponding to the genes. We used the pre-surgery plasma of whole cohort of 40 with

a poor prognosis requiring surgery, and compared the results with those from an independent cohort of patients who demonstrated better prognosis (completed anti-TB treatment (ATT)) and therefore, did not require surgery). Plasma levels of 5 out of 7 proteins expected to be induced (soluble CD14 (sCD14), MMP-9, CXCL-10, CCL19 and lactoferrin (LTF)) were significantly higher in patients requiring surgery, while CD5L levels were significantly lower as, as expected (Fig. 2c). No statistical differences were found for MMP3 plasma levels. Our findings suggest that this 6-plasma protein TB signature may be clinically useful as marker of continued lung pathology after completion of antibiotic therapy, and should be further validated as circulating biomarkers of disease worsening and treatment failure.

Gene expression reveals spatially distinct profiles across the lesion

Next, we compared the expression profiles derived from each section (CL, IW, and EW) against HT to characterise each granuloma tissue type in the TB lesion architecture. A total of 753 DEGs were common in all pairwise comparisons between the three lesion compartments *vs* HT, with CL *vs* HT having the highest number of unique DEGs (2,974) (Fig.3a). The list of DEGs for the CL, IW and EW *vs* HT comparisons was constituted by 6,769, 4,712 and 867 genes respectively (Supplementary Data Sets 3, 4 and 5). The expression of genes associated with the different lesion compartments revealed an enrichment of over-expressed genes in the central lesion (Fig 3b). Notably, the magnitude of differential expression decreased gradually towards the edge of the granuloma structures (Supplementary Data Set 6). As a Pulmonary Nodulus (PN) is often found near the primary lesion and were resected too, we also explored its molecular profile (Fig 1a). The differential expression analysis between granuloma *vs* HT and PN

vs HT, revealed that all the DEGs in PN *vs* HT were also significantly differentially expressed in granuloma *vs* HT (Supplementary Figure 3, Supplementary Data Set 7).

Next, we performed GSEA for the three sets of DEGs (CL *vs* HT, IW *vs* HT and EW *vs* HT), which revealed 34 Hallmark gene sets significantly enriched in the centre of TB lesion compared to the EW (Fig. 3c, Supplementary Data Set 8). The most significant gene sets found to be enriched in CL *vs* EW were mainly related to an immune inflammatory response and complement (Fig. 3c and d), supporting the observation that changes found in granuloma *vs* HT were driven mainly by CL transcriptomic profile.

The expression of seven genes (*P2RX7, LHFPL2, C1QB, SLC6A9, GNPTAB, FAM71A, HPSE*) derived from fitting a spatial granuloma model was measured in the Validation set (Supplementary Figure 1, Supplementary Table 5). All seven genes followed a statistically significant descendant trend through the TB granuloma compartments towards healthy tissue (Fig.4).

Metabolically active *Mtb* were detected in patient lung tissue

Next, we assessed if the presence of *Mtb* inside the lesions[14] could explain the inflammatory response found in the granuloma. RNA is produced by live bacilli, and therefore acts as a marker of metabolically active *Mtb*. We found evidence of *Mtb* ribosomal RNA transcripts in the RNA-Seq data of the lung lesion samples; therefore, we modified the Molecular Bacterial Load (MBL) assay measuring *Mtb* 16S rRNA transcripts[15] to detect and quantify *Mtb* in matched patient lung lesion. Eighteen of 27 patient lung lesions (66%) were positive for *Mtb*, with 16S rRNA abundance above the detection limit equivalent to $1.6x10^2$ CFU (Fig. 5, Supplementary Figure 4). *Mtb* load ranged from $1.7x10^2$ CFU to $2.0x10^4$ CFU per lung sample. Furthermore, we stratified the patients considering *Mtb* drug sensitivity but no statistically significant differences were found. The data demonstrate live metabolically-active *Mtb* in two-thirds of the

patients despite prior anti tuberculosis drug therapy, which might explain the pro-inflammatory profile of the lesions.

A lung lesion gene expression signature differentiates fast and slow sputum culture conversion (SCC) responders

In order to get further insights into the gene expression of the TB lesion according to patient prognosis, we performed a supervised analysis comparing fast and slow SCC responders. The supervised analysis showed 950 significant DEGs (Fig. 6a; Supplementary Data Set 9). According to the GSEA analysis, two hallmark gene sets were significantly enriched in fast sputum converters, related to epithelial-mesenchymal transition and angiogenesis. The most significant gene sets found enriched in the TB lesion of slow sputum converters instead were related to interferon-gamma response and TNFalpha/NFKb signalling hallmarks (Fig. 6b and c, Supplementary Data Set 10). While fast sputum converters show evidence of tissue remodelling, slow sputum converters exhibit a pro-inflammatory profile. Our data therefore suggest that an initial slow clearance of *Mtb* (> 2 months after the start of ATT) may be associated with greater lung inflammatory responses detected many months later (at 6 months after the start of ATT).

DISCUSSION

Through RNA-Seq and NanoString nCounter technology, we have characterised a unique and large collection of 157 samples of TB lesions from 40 patients undergoing therapeutic surgery. One of the main strengths of the project is that fresh human TB lesion specimens (without formalin fixation or paraffin embedding) were transcriptionally characterised using two cutting-edge approaches in both Discovery and Validation participant cohorts. Although some variability was found between samples, we identified a clear pattern across all the patients, with a predominant pro-inflammatory transcriptional signature in TB lung lesion compared to the healthy tissue. As these patients are advanced TB cases with cavitary TB, who underwent lung resection surgery, the results may not be generalised to the whole spectrum of TB disease, which constitutes a limitation to the study. Nevertheless, our study included a larger number of samples in comparison to previous approaches and there are few papers published working on TB lung lesion, in total using 79 samples from 24 patients[13,16,17]. To our knowledge, this is the first study that has built and validated a transcriptional signature of a human TB granuloma from fresh TB lesion biopsy samples, using different granuloma compartments and including, not also MDR/XDR-TB patients as Dheda et al[11], but also, DS-TB patients.

Here we show the important involvement of pro-inflammatory pathways in the TB human lesion, which is supported by Dheda's and Subbian's studies[11,13]. Dheda et al described the pathways involved in different parts of cavitary lesions from 14 failed MDR-TB subjects that underwent surgery, pointing the cavity wall as the main source of pro-inflammatory activity[11]. In line with our findings and through ingenuity pathway analysis (IPA), they showed that pro-inflammatory pathways were especially over-represented in the cavity wall, including nitric oxide production, reactive oxygen species,

IL-1, IL-6, IFN-γ and NF-κβ transcriptional signatures. Furthermore, Subbian et al. demonstrated through Gene Ontology analysis and using 4 granuloma samples, the involvement of immune cell signalling and activation, interferon response and tissue remodelling processes in the complex TB granuloma microenvironment and the gene expression pattern of vitamin D receptors[13]. Using a different approach (GSEA), our results point in the same direction while providing novel and additional information. Even considering different sample preservation methods, RNA extraction, and analysis technologies, some of the differentially expressed genes in the above-mentioned studies overlapped with our validated genes of interest. In contrast to Dheda's work, in our hands the centre of the TB lesion showed a more perturbed transcript abundance compared to the other lesion compartments. Additionally, we found a characteristic highly pro-inflammatory gene expression profile at the centre of the lesion, while the edges of the TB lesion cavities (EW) were more similar to healthy lung tissue. These differences in terms of spatial expression could be due to differences in the sample preparation or analyses, or to differences between the two patient cohorts.

The TB lesion 11-gene signature described here, predominantly encodes proteins related to the immune response, inflammatory processes, and cell signalling, such as LTB and MAP4K1[18,19], to orchestrate the antimycobacterial response. Some are already targets for host-directed therapies such as LTF, FCMR and CD5L[20–23]. In addition, *CCL19* gene expression has been found to be increased in lungs post-infection in mice acting to induce lymphoid structures[24]. MMP-3 and MMP-9, as metalloproteinases, are known to participate in the orchestration of the TB granuloma[25,26]. It is also worth noting that *MMP9*, which has been suggested to be responsible for the persistence of cavities, was ranking among the top 10 DEG genes in the Subbian study[13]. In addition, *IL2RG* and *CD14* were expressed in a similar pattern, unlike *LTB*. Similarly, *LHFPL2*, *C1QB*, and

HPSE genes from our spatial model of the TB granuloma also matched[13]. We found *C1Q1* gene to be over-expressed in central parts of the granuloma, contributing for the pro-inflammatory profile. The expression of this gene has been highlighted as a disease severity biomarker[27] and linked to poor treatment response[28]. Finally, interleukin-2, CXCL-10, and soluble CD14 have been proposed as biomarkers of disease severity[29–31].

Six proteins described as protein TB signature (sCD14, CD5L, CXCL-10, MMP-9, LTF and CCL19) were found statistically different in the plasma of patients ending up in surgery, in spite of completing treatment, compared to successfully treated TB patients. Previously published blood microarray-derived signatures found up-regulated levels of *CXCL10*[28,32,33], *MMP9*[28,32,33], *LTB* [28], *LTF* [28], *CD14* [28], *LHFPL2* [32,33], *C1QB*[32,33] and *HPSE*[32,33], when comparing baseline to end of treatment samples. *CXCL-10*, *CD14*, *MMP-9* and *LTF* were found up-regulated in blood of patients at baseline when compared to 26 weeks after treatment[28]. Recently, these genes were also found differentially expressed in TB compared to control samples in a study published by Reichmann et al. using mediastinal and neck lymph nodes samples of non-treated TB patients, a different set of samples from the used in our work[34]. We found them over-expressed in the lesion of patients undergoing surgery for unresolved TB, and also detected the proteins encoded by these genes increased in circulation, suggesting that it might be a biomarker for treatment response, poor prognosis or both.

Our results showing an important component of inflammation in granulomas in patients after completing drug therapy are supported by the findings of Malherbe et al. These authors showed through 18-F fluorodeoxyglucose positron emission tomography-computed tomography (18-FDG PET-CT) lung scans that some patients still have an increased FDG uptake in the lesions when compared to surrounding healthy tissue after

6 months of treatment[35]. More recently they have also related both a larger burden of disease and a slower rate of reduction in scan metrics to delayed sputum conversion[36].

The main novelty of our work is that our data showed that slow sputum converters have different transcriptional profiles in TB granulomas when compared to fast sputum converters, and this has clinical management implications. TB surgery is usually pursued in cases with persistent pulmonary cavities after long-term standard ATT despite successful sputum sterilisation. Achieving sputum conversion after two months of treatment has been associated with TB cavity persistence[7]. In our cohort, DS-TB/MDR-TB cases in the fast SCC converters was higher (5/8) than in slow converters (4/6), as previously described[37]. Although the SCC has been considered a surrogate of prognosis[38,39], it is still not used for clinical decision making. Importantly, our study suggests that patients who converted the sputum after two months still show inflammatory pathways enriched in the lesion after several months, while fast converters have enriched pathways involved in tissue repair. These data support using SCC at month 2 after starting treatment as a prognostic biomarker, and suggest it might be used as a factor to prioritise patients when considering lung surgery.

The TB lesion 11- gene signature of ongoing lung pathology, including the 6-protein TB signature, should be further studied as potential biomarkers for performance evaluation, using larger cohorts including patients at different stages of disease. Our study is the first one showing evidences of metabolically-active bacilli in the lesions of two thirds of patients despite ATT, which may be the reason why these patients needed surgical intervention. We provide evidence that some circulating biomarkers mirror what is happening *in situ* in the lung, which may help for the development of new diagnostic and prognostic tools, and treatment strategies. Moreover, this is the first study showing different pulmonary transcriptomic patterns associated with sputum culture conversion

status, enabling the stratification of patients at an early timepoint in the clinical management of TB disease (month 2 after starting treatment) as part of a patient cantered approach to TB treatment.

ONLINE METHODS

Ethics

This paper is part of the SH-TBL project (ClinicalTrials.gov Identifier: NCT02715271). The protocol and research methodology, as well as associated documents (informed consent sheet, informed consent form) were reviewed and approved by both ethics' committees at the NCTLD (IRB00007705 NCTLD Georgia #1, IORG0006411) and the Germans Trias i Pujol University Hospital (EC: PI-16-171). For the analysis of biomarkers in plasma, we used a pre-existing collection of samples from patients with TB from the STAGE-TB cohort (ClinicalTrials.gov Identifier: NCT03691883) approved by the ethics committees of the Germans Trias i Pujol University Hospital (CEI PI-17-064). Written informed consent was obtained from all study participants before they were enrolled.

Study design and patient cohort

We conducted a cross-sectional study, in which we recruited 40 adult patients from the National Center of Tuberculosis and Lung Diseases (NCTLD) in Tbilisi (Georgia), from May 2016 to May 2018, after receiving therapeutic surgery indication for their pulmonary TB. Study participants constituted the SH-TBL Cohort. All volunteers received standard anti-TB treatment (ATT) regimen for their pulmonary TB according to Georgia national guidelines[40,41] and achieved sputum culture conversion. They required surgery because of persistent radiological signs of cavitary lesions on chest X-Ray (CXR) and computed tomography scan, according to the Georgian National Guideline "Surgical Treatment of Patients with Pulmonary Tuberculosis" based on official guidelines of surgery recommendations for pulmonary TB[42]. Thoracic surgery decisions were made by the NCTLD Resistant Tuberculosis Treatment Committee, composed of two surgeons and 18 pulmonary TB specialists.

Data and sample collection

Anonymised data regarding the socio-epidemiological factors, clinical aspects, and information referring to the current TB episode for the SH-TBL Cohort were collected in an electronic case report form platform.

Patients were classified according to their sputum conversion date, being considered as fast converters or slow converters depending on if sputum conversion happened before or after 2 months of anti TB treatment, respectively. Drug sensitivity of the *Mtb* strain responsible for the TB episode was also recorded, considering DS-TB (Drug Sensitive) or MDR (MultiDrug Sensitive)/XDR-TB (ExtraDrug Resistant).

All data were transferred to a data frame (SH-TBL Cohort metadata) and are freely accessible, including continuous and categorical variables (doi: 10.17632/knhvdbjv3r.1). During surgery, granuloma TB lesions 3.2 cm in diameter (median; 2.2-5.5 cm) were removed after performing segmentectomy, lobectomy, or pneumonectomy, according to surgical criteria. Biopsies from three different sections of this TB lesion were collected, namely Central Lesion (CL), Internal Wall (IW), External Wall (EW). Together with the resection of the human TB granuloma, a solitary Pulmonary Nodulus (PN) sample was removed. In addition, healthy lung parenchyma (Healthy Tissue, HT), unaffected by eye and by palpation, was collected from the same patient's lung (Fig. 1a). Biopsy fragments (~0.5 cm^3) of tissue samples were collected for RNA-Seq. Fresh tissue samples intended for RNA-Seq were incubated in RNA*later* solution (Qiagen) at 4ºC overnight, before storage at -80ºC. In total, 199 SH-TBL biological samples were collected for the study, composed of 39 CL, 40 IW, 40 EW, 40 PN, and 40 HT tissues (Supplementary Table 1). The samples were shipped from the NCTLD to the Comparative Medicine and Bioimage Centre of Catalonia (CMCiB) BioSafety Level 3 (BSL-3) laboratory in Badalona (Spain), where total-RNA extraction was performed.

Plasma samples from the same 40 patients were collected prior to surgery (SH-TBL Cohort) and kept frozen at -80ºC at the IGTP until further analysis. Samples used as controls in the biomarker analysis in plasma came from 13 patients from the ongoing STAGE-TB project (registered at the ClinicalTrials.gov database under the code NCT03691883) who had completed 6 months of ATT, and that were stored frozen at -80ºC in the IGTP facilities.

Total-RNA extraction procedure

For an optimal RNA recovery, TB lesion biopsy samples were divided into 0.16-0.21g single pieces and placed into new tubes. Samples were reduced to powder by mechanical cryofracturing using a BioPulverizer device (Biospec Products) after being cooled by liquid nitrogen. The powdered tissue was then transferred to 2mL Lysing Matrix D tubes together with a lysis solution for homogenization by FastPrep® instrument (MP Biomedicals). RNA was purified using the mirVana miRNA Isolation Kit (Thermo Fisher Scientific), followed by genomic DNA digestion using the DNA-*free* DNA Removal Kit (Thermo Fisher Scientific) according to manufacturer's instructions. Quantitative and qualitative RNA integrity number equivalent (RINe) values were obtained by Agilent Bioanalyzer 2100 (Agilent Technologies). In general, a standard RINe score for good quality RNA is set at seven. However, considering the source and status of the human TB lesion samples, we used a minimum RINe cutoff of four, accepting that low RNA integrity is an important bias factor in RNA-Seq experiments.

Discovery Set RNA-Seq library preparation, sequencing, and gene alignment

The Discovery Set for the RNA-Seq experiments comprised 61 samples from 14 patients and was constituted considering participant sex and *Mtb* drug-sensitivity classification as potential confounding factors. Seven males and seven females were included, 3 DS-TB, 3 MDR-TB and 1 XDR-TB patient per gender group. Nine samples were discarded as the

RINe threshold was not achieved (Supplementary Table 1). Discovery Set participants were included consecutively as per recruitment order until reaching the pre-set criteria.

Purified RNA from Discovery Set participants was diluted to 25 ng/µl per aliquot and then shipped on dry ice to Macrogen (Seoul, South Korea), where the RNA-sequencing was performed. Libraries were constructed using the TruSeq Stranded Total RNA LT Sample Prep Kit (Human Mouse Rat) (Illumina) following the TruSeq Stranded Total RNA Sample Prep guide (Part #15031048 Rev. E), including prior removal of ribosomal RNA using the RNA Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) (Illumina).

RNA-Seq was performed on an Illumina platform HiSeq 4000 (Illumina), at 50 million reads per sample, 100bp stranded paired-end reads. Preprocessing of raw data included quality control through FastQC (v.011.7) and MultiQC (v.1.9)[43]. Before further steps in read preprocessing, Illumina adapters were trimmed off with Trimmomatic (v.0.39)[44]. The human genome sequence GRCh38.89 and human gene annotations were downloaded from the ENSEMBL web repository. Files from each Discovery Set sample were aligned to the human reference genome through the Spliced Transcripts Alignment to a Reference (STAR) package (v.2.7.5b)[45], with the built-in gene counts quantification mode for stranded RNA-Seq data. BAM files were generated, and the SAMtools package applied to calculate the percentage of successful read alignment against the reference human genome (v.1.10)[45].

Discovery Set statistical analysis

The overall pipeline for the data handling, plotting and statistical analysis was conducted in R (v.3.6.2)[46]. After STAR mapping, a gene count data table was obtained including CL, IW, EW, PN, and HT samples belonging to the Discovery Set. Genes with mean raw counts that were below or equal to five were discarded as they were too lowly expressed

to be reliably quantified and could potentially interfere with the differential gene expression analysis.

The differential expression analysis from tissue count tables was conducted using the DESeq2 Bioconductor package (v.1.28.1)[47]. Genes were considered as significant DEGs when the Benjamini-Hochberg adjusted *p*-value was equal to or less than 0.05 ($P \leq 0.05$). The differential expression examination for two-factor levels under an experimental design was employed through the contrast argument. The contrast argument is the linear combination of the estimated $\log_2$FC, testing if differences between the groups were equal to zero. Patient's sex and drug-sensitivity type were accounted for in DESeq2 modelling. The spatial transcriptional model analysis within the human TB granuloma was evaluated in a linear additive manner, obtaining a gradient of gene expression throughout the TB lesion architecture. Gene intersections were explored with the online tool Venny[48]. The R package pheatmap (v.1.0.12) was used to generate heatmaps and dendrograms for the genes and samples by hierarchical clustering after DESeq2 depth normalization. Heatmaps are showing the Euclidean distances between samples.

Pathophysiological analysis of generated molecular signatures

After conducting the differential expression analysis, resultant transcriptional signatures were analyzed by GSEA (The Broad Institute) (v.4.0.3)[49]. Preranked GSEA was performed using the human TB granuloma and the time to SCC signatures according to the $\log_2$FC values. The signature derived from the spatial analysis of the human TB granuloma was not suitable for this analysis given its linear additive design.

The file format was set following GSEA instructions for RNA-Seq projects. The gene sets database was established with the [Hallmarks] option and the chip platform as "Human ENSEMBL Gene ID MSigDB.v7.0.chip" as a reference database for GSEA analysis. A Statistical significance for every enrichment score per gene set was given by

the nominal *p*-value, considering values <0.05 as significant. False Discovery Rate (FDR) was set at 25%, reducing the probability that a gene set with a given NES becomes a false-positive result. The rank metric score indicates the position of the gene within the gene set after a calculated score. The Running ES is the ES in the corresponding gene set at this point in the ranked list of genes.

Genes of interest selection

We used a statistical and a pathophysiological rationale to select the most meaningful genes from each signature from the Discovery Set to build a custom gene expression panel for the validation analysis using the NanoString platform (Supplementary Figure 1, Supplementary Table 5). The pathophysiological rationale consisted of selecting the significant genes from each signature after running pre-ranked GSEA. Molecular pathways related to inflammatory processes, immune response, tissue remodeling, and destruction mechanisms as well as hypoxia and the complement system were prioritized for their role in the generation and maintenance of granulomatous lesions. We selected those genes with a FC >2.00/<-2.00 in the 6,056-gene human TB signature considering only the top DEG genes according to adjusted *p*-value.

To complete the genes of interest sub-set, we used a statistical rationale consisted of selecting those DEG genes that had a $\log_2$FC equal or superior to |1| as well as a per gene mean normalized count greater than 50 in all samples, considering only the top 20 significantly differentially expressed (DEG) ranked genes. From the 6,056-gene human TB granuloma signature, we selected *MAP4K1*, *FCMR*, and *CCL19*.

From the granuloma spatial analysis, we selected those genes whose $\log_2$ FC represents the entire gradient of expression, covering the genes mainly expressed in CL towards HT to be further validated. This was traduced into selecting the seven first DEG ranked genes

according to the Benjamini-Hochberg adjusted *p-value*: *P2RX7, LHFPL2, C1QB, SLC6A9, GNPTAB, FAM71A, HPSE*.

Finally, we were also interested in validating the expression of genes encoding circulating biomarkers related to prognostic features according to the literature and that were found DEG in the 6,056-human TB granuloma signature, namely *IL2RG, CD14, CXCL10, CD5L*, and *MMP9*, as they could be potentially related to the worsening phenotype of the SH-TBL Cohort.

Validation Set using NanoString technology

In order to validate the RNA-Seq findings, we used a second patient set and an alternative measure of transcript abundance. The Validation Set of patients consisted of 95 tissue samples (70 human TB granuloma biopsies and 25 HT) from 26 SH-TBL Cohort subjects (Supplementary Table 1). Six CL and one HT were discarded as the RINe values were below four. The Genes of Interest panel consisted of 18 genes (*MMP3, LTB, LTF, MAP4K1, FCMR, CCL19, P2RX7, LHFPL2, C1QB, SLC6A9, GNPTAB, FAM71A, HPSE, IL2RG, CD14, CXCL10, CD5L*, and *MMP9*), and five housekeeping (HK) genes (*ACTB, ACTG1, ADCY3, PCBP1*, and *PLXNB2*) (Supplementary Table 5), assayed using the NanoString platform.

A 27-probe set custom panel was designed and the Validation Set samples assayed using the NanoString nCounter XT assay, following manufacturer's instructions. We used the standard gene expression protocol for 12x RNA hybridization with 100 ng total-RNA as input, a hybridization time of 18h for all samples and mixing controls and TB samples in order to avoid any technical bias. Samples were analysed in a SPRINT nCounter system (NanoString). Raw counts were obtained and analysed through NanoString's nSolver (v.3.0) digital analyser software. Technical normalization was carried out using the geometric mean of the positive controls, whereas the background was established using

a threshold calculated by the mean of negative probes plus two standard deviations. We discarded those genes below the defined threshold. Differences in sample input were corrected by selecting the optimal number and best reference candidates from the set of HK genes included in the panel. GeNorm algorithm[50] implemented in the R package CrtlGene[51] was used to test not only the most stable reference genes but also the optimal number to carry out the normalization step. We disregarded HK candidate genes with less than 50 raw counts in any of the sample comparisons.

Validation Set statistical analysis

After obtaining the NanoString nCounter results, endogenous genes from the Validation Set were first analysed to determine statistical differences. We applied the Mann-Whitney U test to identify differentially expressed genes between groups, namely the pooled counts of granuloma and HT as well as between the sub-set of fast SCC and slow SCC. For visualization purposes, data were Log_2 transformed. Correlation coefficients between Discovery Set and Validation Set Log_2FC values were determined using the Pearson correlation coefficient. Validation Set Log_2FC values were calculated with the formula log2(mean(group1)/mean(group2)) whereas Discovery Set Log_2FC values were extracted from the DESeq2 results. Linear regression was used to corroborate the linear additive model results and exhibit the gradient of expression through the human TB granuloma substructure. Results were considered as significant when p-value ≤ 0.05.

Assessment of biomarkers in plasma samples

Levels of circulating proteins in plasma (encoded by transcripts validated to be differentially expressed by NanoString) were assessed through ELISA and Luminex assays. Plasma levels of soluble CD14 (sCD14) and CXCL-10 were determined by sandwich ELISA (cat#DC140 and DIP100, Quantikine ELISA and Kits, R&D Systems, MN, USA). Plasma samples were 200-fold diluted for sCD14 and used neat for CXCL-

10 detection. The assays were performed according to the manufacturer's instructions. CD5L plasma levels were measured using the CircuLex Human AIM/CD5L/Spα ELISA kit, according to manufacturer's instructions (cat#CY-8080, Medical & Biological Laboratories Co., Ltd, Aichi, Japan) and samples diluted 100-fold. ELISA plates were read in a Thermo Scientific Varioskan® plate reader (Thermo Fisher) on 450nm wavelength.

Luminex multiplex assay was used to detect MMP-9 and Lactoferrin (LTF) (Human Premixed Multi-Analyte Kit, cat#LXSAHM-02, R&D Systems, MN, USA). Plasma samples were diluted 50-fold for MMP-9 and 100- or 200-fold for LTF detection. Assays were performed according to the manufacturer's instructions and plates read using a Luminex® 200 System.

Biomarkers and *Mtb* RNA quantification data were analysed using GraphPad Prism (version 9.0). For circulating biomarker data, the Mann-Whitney test was used to compare two groups. Significant differences are as follows: $*p \leq 0.05$, $**p \leq 0.01$; $***p \leq 0.001$ and $****p \leq 0.0001$.

Extraction and quantification of *Mtb* RNA from tissue

Mtb RNA was extracted from lung tissues using a modified GTC/TRIzol differential lysis extraction method[15]. Briefly, lesion tissue samples from 27 patients stored in RNALater (described above) were first disrupted in PBS using the gentleMACS Octo Dissociator H12.01 (Miltenyi Biotec) before adding 5M guanidine thiocyanate solution. The samples were centrifuged, passed through a 0.8 mm needle, and re-suspended in 1 mL TRIzol (Thermo Scientific). After bead beating with 0.1 mm silica beads (MP Biomedicals), RNA was DNase-treated and purified using the RNeasy Mini Kit (Qiagen) as previously described[52]. RNA yield and integrity were assessed using NanoDrop One (Thermo Scientific) and 2100 Bioanalyzer (Agilent Technologies).

Quantification of *Mtb* was determined from 16S rRNA abundance using the molecular bacterial load (MBL) assay with some adjustments[53]. Expression of *Mtb* 16S rRNA was measured on a QuantStudio 7 RT-PCR machine using QuantiNova SYBR Green (Qiagen) reverse transcriptase (RT)-quantitative PCR (RT-qPCR) reagents following the manufacturer's instructions. Each sample was quantified in three independent RT-qPCR runs. An RT minus control was also included for each patient sample. *Mtb* 16S RNA yields were extrapolated from a standard curve of *Mtb* H37Rv RNA, ranging from 200 ng to 0.0001 ng, run in triplicate on each plate (Supplementary Figure 4). To estimate *Mtb* molecular bacterial load in tissue, 16S RNA abundance was compared to 16S rRNA yields from a dilution series of *Mtb* H37Rv in liquid culture measured in triplicate (colony forming units (CFU) from $3x10^6$ CFU to $3x10^1$ CFU). After adjusting for sample volume, samples with an estimated bacterial load above the detection limit of $1.6x10^2$ CFU were considered positive for *Mtb*.

ACKNOWLEDGEMENTS

The authors would like to thank the patients who agreed to participate in the study and the staff from the NCTLD who helped with the SH-TBL project. The authors also thank the patients, Arantxa Romero and the medical staff (Zoran Stojanovic, Adrián Antuori, Sílvia Roure, Antoni Noguera-Julian, Antoni Soriano-Arandes, Andrea Martín, Maria Espiau, Maria Teresa Tortola, Malú De Souza, Mª Ángeles Jiménez, Israel Molina, Xavier Casas, Núria López, Marisol Domínguez-Álvarez, Adrián Sánchez-Montalvá, Joan Pau Millet) from the hospitals involved in STAGE-TB Project, who participated in the samples collection that was used as controls for the biomarkers assessment in plasma. We would also like to thank Eric Garcia for their technical support in acquiring the data.

FUNDING

This work was supported by: 1) the Plan Nacional I + D + I co-financed by ISCIII-Subdirección General de Evaluación and Fondo-EU de Desarrollo Regional (FEDER) through PI16/01511, PI20/01424, CP13/00174, CPII18/00031 and CB06/06/0031. 2) The European Union's Horizon 2020 research and innovation program under grant agreement No 847762. 3) The Catalan Agency for Management of University and Research Grants (AGAUR) through 2017SGR500 and 2017 FI_B_00797. 4) The "Spanish Society of Pneumology and Thoracic Surgery" (SEPAR) (16/023). 5) The Wellcome Trust, the Medical Research Foundation grants (206508/Z/17/Z and MRF-160-0008-ELP-KAFO-C0801) and the NIHR Imperial College BRC to M.K. 6). N.G., registered in the EMJMD LIVE (Erasmus+ Mundus Joint Master Degree Leading International Vaccinology Education), co-funded by the EACEA (Education, Audiovisual and Culture Executive Agency, award 2018-1484) of the European commission, receives a scholarship from the EACEA. 7). Wellcome Trust (204538/Z/16/Z) and National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) (NC/R001669/1) grants to S.J.W.

AUTHOR CONTRIBUTIONS

C.V. and S.V. conceived the study. C.V., M.K., PJ.C. and C.A. made substantial contributions to the design of the study pipeline. S.V., S.G., K.N., N.S. and Z.A. (NCTLD) participated in the human sample and data collection used for this study. A.D. and L.A. performed the total-RNA extraction of the human samples. A.D., D.H.C., A.d R., J.C., K.L.F., C.A. and M.K. performed the bioinformatic analysis. A.G.C., F.M-T. and A.S executed the nCounter analysis. K.L.F and N.G. performed the assessment of biomarkers in patients' plasma. L.M.W. extracted the bacterial samples, and A.G. performed the analysis of *Mtb* RNA from patient tissues. A.D., K.L.F., M.K., S.J.W., C.A. and C.V. interpreted the study results. A.D. and K.L.F. wrote the manuscript. C.V., K.L.F., M.K., S.V., S.J.W. and C.A. substantively revised the work. All authors approved the submitted version (and any substantially modified version) and agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

COMPETING INTERESTS

The authors declare the following competing interests: F.M-T has received honoraria from GSK group of companies, Pfizer Inc, Sanofi Pasteur, MSD, Seqirus, Biofabri and Janssen for taking part in advisory boards and expert meetings and for acting as a speaker in congresses outside the scope of the submitted work. F.M-T has also acted as principal investigator in randomized controlled trials of the above-mentioned companies as well as Ablynx, Gilead, Regeneron, Roche, Abbott, Novavax, and MedImmune, with honoraria paid to his institution. C.V salary is covered by ISCIII (CPII18/00031) and Fundacio Institut d'Investigació en Ciencies de la Salut Germans Trias I Pujol (IGTP). IGTP is the

applicant of the patent EP3880702A1, which inventors are: MR Sarrias Fornés, G Aran Canals, E Téllez Caballero, C Vilaplana Massaguer and FE Borràs Serres. The patent relates to method and kit for the diagnosis and/or prognosis of inflammatory conditions based on analysing the levels of CD5L in a sample and it has entered into National Phase. In the present manuscript, we found CD5L under-expressed in the granuloma compared with the healthy tissue, as well as lower CD5L plasma levels in patients requiring surgery compared to those who didn't. Consequently, CD5L was included both in the TB lesion 11-gene signature and in the 6-plasma protein TB signature described in the present manuscript. C.V is an unpaid board member of the following non-profit organizations: the FUITB foundation and the ACTMON foundation. Neither the ISCIII, the IGTP, the FUITB nor ACTMON have had any role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

DATA AVAILABILITY

All metadata and sequencing data from this manuscript have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) Database and will be released after manuscript acceptance in a peer-reviewed scientific journal.

REFERENCES

1. World Health Organization. *Global Tuberculosis Report*. *Global tuberculosis report* https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2021 (2021).
2. Petersen, E. *et al.* World TB Day 2022: Revamping and Reshaping Global TB Control Programs by Advancing Lessons learnt from the COVID-19 pandemic. *Int. J. Infect. Dis.* (2022) doi:10.1016/J.IJID.2022.02.057.
3. World Health Organization, (WHO). *WHO report on TB 2020*. *WHO* vol. 1 (2020).
4. Kempker, R. R., Vashakidze, S., Solomonia, N., Dzidzikashvili, N. & Blumberg, H. M. Surgical treatment of drug-resistant tuberculosis. *Lancet. Infect. Dis.* **12**, 157–166 (2012).
5. Kempker, R. R. *et al.* Additional drug resistance in Mycobacterium tuberculosis isolates from resected cavities among patients with multidrug-resistant or extensively drug-resistant pulmonary tuberculosis. *Clin. Infect. Dis.* **54**, 51–54 (2012).
6. Dartois, V. & E. Barry, C. Clinical pharmacology and lesion penetrating properties of second- and third-line antituberculous agents used in the management of multidrug-resistant (MDR) and extensively-drug resistant (XDR) tuberculosis. *Curr. Clin. Pharmacol.* **5**, 96–114 (2010).
7. Vashakidze, S. *et al.* Retrospective study of clinical and lesion characteristics of patients undergoing surgical treatment for Pulmonary Tuberculosis in Georgia. *Int. J. Infect. Dis.* **56**, 200–207 (2017).
8. Perrin, F. M., Lipman, M. C., McHugh, T. D. & Gillespie, S. H. Biomarkers of treatment response in clinical trials of novel antituberculosis agents. *Lancet. Infect. Dis.* **7**, 481–490 (2007).

9. Warsinske, H. C. *et al.* Assessment of Validity of a Blood-Based 3-Gene Signature Score for Progression and Diagnosis of Tuberculosis, Disease Severity, and Treatment Response. *JAMA Netw. open* **1**, e183779 (2018).

10. Thompson, E. G. *et al.* Host blood RNA signatures predict the outcome of tuberculosis treatment. *Tuberculosis* **107**, 48–58 (2017).

11. Dheda, K. *et al.* Spatial network mapping of pulmonary multidrug-resistant tuberculosis cavities using RNA sequencing. *Am. J. Respir. Crit. Care Med.* **200**, 370–380 (2019).

12. Lin, P. L. & Flynn, J. L. The End of the Binary Era: Revisiting the Spectrum of Tuberculosis. *J. Immunol.* **201**, 2541–2548 (2018).

13. Subbian, S. *et al.* Lesion-specific immune response in granulomas of patients with pulmonary tuberculosis: A pilot study. *PLoS One* **10**, 1–21 (2015).

14. Moreno-Molina, M. *et al.* Genomic analyses of Mycobacterium tuberculosis from human lung resections reveal a high frequency of polyclonal infections. *Nat. Commun. 2021 121* **12**, 1–11 (2021).

15. Honeyborne, I. *et al.* Profiling persistent tubercule bacilli from patient sputa during therapy predicts early drug efficacy. *BMC Med.* **14**, (2016).

16. Kim, M. J. *et al.* Caseation of human tuberculosis granulomas correlates with elevated host lipid metabolism. *EMBO Mol. Med.* **2**, 258–274 (2010).

17. McCaffrey, E. F. *et al.* The immunoregulatory landscape of human tuberculosis granulomas. *Nat. Immunol. 2022 232* **23**, 318–329 (2022).

18. Agyekum, S. *et al.* Expression of lymphotoxin-beta (LT-β) in chronic inflammatory conditions. *J. Pathol.* **199**, 115–121 (2003).

19. Sawasdikosol, S. & Burakoff, S. A perspective on hpk1 as a novel immuno-oncology drug target. *Elife* **9**, 1–15 (2020).

20. Ting, L. M., Kim, A. C., Cattamanchi, A. & Ernst, J. D. Mycobacterium tuberculosis inhibits IFN-gamma transcriptional responses without inhibiting activation of STAT1. *J. Immunol.* **163**, 3898–906 (1999).

21. Wang, H., Coligan, J. E. & Morse, H. C. Emerging Functions of Natural IgM and Its Fc Receptor FCMR in Immune Homeostasis. *Front. Immunol.* **7**, (2016).

22. Ingersoll, M. A., Kline, K. A., Nielsen, H. V. & Hultgren, S. J. G-CSF induction early in uropathogenic Escherichia coli infection of the urinary tract modulates host immunity. *Cell. Microbiol.* **10**, 2568–2578 (2008).

23. Sanjurjo, L., Aran, G., Roher, N., Valledor, A. F. & Sarrias, M.-R. AIM/CD5L: a key protein in the control of immune homeostasis and inflammatory disease. *J. Leukoc. Biol.* **98**, 173–184 (2015).

24. Kahnert, A. *et al.* Mycobacterium tuberculosis triggers formation of lymphoid structure in murine lungs. *J. Infect. Dis.* **195**, 46–54 (2007).

25. Parasa, V. R. *et al.* Inhibition of tissue matrix metalloproteinases interferes with Mycobacterium tuberculosis-induced granuloma formation and reduces bacterial load in a human lung tissue model. *Front. Microbiol.* **8**, 1–14 (2017).

26. Greenlee, K. J., Werb, Z. & Kheradmand, F. Matrix metalloproteinases in lung: Multiple, multifarious, and multifaceted. *Physiol. Rev.* **87**, 69–98 (2007).

27. Cai, Y. *et al.* Increased complement C1q level marks active disease in human tuberculosis. *PLoS One* **9**, (2014).

28. 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 (2013).

29. Sigal, G. B. *et al.* Biomarkers of Tuberculosis Severity and Treatment Effect: A Directed Screen of 70 Host Markers in a Randomized Clinical Trial.

EBioMedicine **25**, 112–121 (2017).

30. Petrone, L. *et al.* Blood and urine inducible protein 10 as potential markers of disease activity. *Int. J. Tuberc. Lung Dis.* **20**, 1554–1561 (2016).

31. Memar, M. Y. & Baghi, H. B. Presepsin: A promising biomarker for the detection of bacterial infections. *Biomedicine and Pharmacotherapy* (2019) doi:10.1016/j.biopha.2018.12.124.

32. Bloom, C. I. *et al.* Detectable Changes in The Blood Transcriptome Are Present after Two Weeks of Antituberculosis Therapy. *PLoS One* **7**, (2012).

33. Ottenhoff, T. H. M. *et al.* Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis. *PLoS One* **7**, (2012).

34. Reichmann, M. T. *et al.* Integrated transcriptomic analysis of human tuberculosis granulomas and a biomimetic model identifies therapeutic targets. *J. Clin. Invest.* **131**, (2021).

35. Malherbe, S. T. *et al.* Persisting PET-CT lesion activity and M. tuberculosis mRNA after pulmonary tuberculosis cure. *Nat. Med.* **22**, 1094 (2016).

36. Malherbe, S. T. *et al.* Quantitative 18F-FDG PET-CT scan characteristics correlate with tuberculosis treatment response. doi:10.1186/s13550-020-0591-9.

37. Kurbatova, E. V. *et al.* Sputum culture conversion as a prognostic marker for end-of-treatment outcome in patients with multidrug-resistant tuberculosis: a secondary analysis of data from two observational cohort studies. *Lancet. Respir. Med.* **3**, 201–209 (2015).

38. Assemie, M. A., Alene, M., Petrucka, P., Leshargie, C. T. & Ketema, D. B. Time to sputum culture conversion and its associated factors among multidrug-resistant tuberculosis patients in Eastern Africa: A systematic review and meta-analysis. *Int. J. Infect. Dis.* **98**, 230–236 (2020).

39. Wallis, R. S. *et al.* Biomarkers for tuberculosis disease activity, cure, and relapse. *Lancet Infect. Dis.* **9**, 162–172 (2009).

40. World Health Organization (WHO). *Guidelines for treatment of drug-susceptible tuberculosis and patient care*. *WHO* (2017).

41. WHO consolidated guidelines on drug-resistant tuberculosis treatment. (2019).

42. World Health Organization. The role of surgery in the treatment of pulmonary TB and multidrug- and extensively drug-resistant TB. (2014).

43. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

44. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009).

46. R Core Team. R: A language and environment for statistical computing. (2019).

47. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

48. Oliveros, J. C. Venny. An interactive tool for comparing lists with Venn's diagrams.

49. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

50. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, (2002).

51. Zhong, S. ctrlGene: Assess the Stability of Candidate Housekeeping Genes. (2019).

52. Wildner, L. M., Gould, K. A. & Waddell, S. J. Transcriptional profiling Mycobacterium tuberculosis from patient sputa. doi:10.1007/978-1-4939-7638-6_11.

53. Honeyborne, I. *et al.* Molecular bacterial load assay, a culture-free biomarker for rapid and accurate quantification of sputum Mycobacterium tuberculosis bacillary load during treatment. *J. Clin. Microbiol.* **49**, 3905–3911 (2011).

FIGURE LEGENDS

Figure 1. Human TB granuloma transcriptomic profile is different from the healthy tissue and enriched for proinflammatory genes. A set of 6,056 genes was identified as significantly differentially expressed in the Discovery Set after comparing the human TB granuloma counts (Central Lesion – CL, Internal Wall – IW, and External Wall – EW pooled counts) to HT expression. **a.** TB lesions samples, ranging from 2.2 to 5.5 cm, were kept and collected from each of the three granuloma compartments (CL, IW and EW). Healthy pulmonary tissue (HT) from the same involved lung was collected to perform as tissue control, plus pulmonary nodulus (PN) biopsies. **b.** Heatmap showing the top 40 DEGs comparing the human TB granuloma (CI + IW + EW pooled counts) to HT expression profile (Benjamini-Hochberg adjusted *p*-value ≤ 0.05). Scale bar red colour refers to gene's over-expression; blue colour refers to under-expression. Heatmap dendrogram shows hierarchical clustering of exhibited DEGs. **c, d**. GSEA dot plots showing the enrichment of 22 hallmark gene sets, based on the differentially expressed genes in the granuloma tissue compared to HT. The top regulators with positive activation z scores and highest inverse Fisher's exact enrichment *p* values are shown. GSEA plots were built considering 1 sample/patient, using the mean of CL+IW+EW counts.

Figure 2. Human TB lesion 11-gene set performance in Validation Set and detection in plasma samples. a. Boxplot of genes of interest expression in granuloma pooled counts (brown) versus HT (green) in the Validation Set. Gene expression is expressed as count's $\log_2$. Mann-Whitney U test was applied to study gene expression differences between both tissue sources (***$P \leq 0.001$). **b.** Log_2 FC Pearson's correlation plot of the 11-gene signature between RNA-Seq's Discovery Set and NanoString's Validation Set. Correlation coefficient (R) and *p*-value (significant as *p*-value ≤ 0.05). **c.** Proteins encoded

by the TB granuloma set of genes were identified in plasma samples of TB patients whose underwent surgery (Surgery) and their levels compared to Non-surgery TB patients who had completed ATT (Non-surgery). Statistical analysis was performed with unpaired two-tailed Mann-Whitney test. Biomarkers levels are represented on a log_{10}-scale * refer to statistical differences between Non-surgery and Surgery patients' groups. *$p < 0.05$; **$p < 0.01$; ****$p < 0.0001$.

Figure 3. Granuloma transcriptomic profile considering the three compartments contributions. a. Venn diagram showing gene intersection of CI (yellow), IW (orange), and EW (brown) generated signatures of DEGs (Benjamini-Hochberg adjusted *p*-value ≤ 0.05). **b**. Heatmap showing a gene expression gradient is shown throughout CL (yellow), IW (orange), EW (brown) and HT (green), in the Discovery Set. (Benjamini-Hochberg adjusted *p*-value ≤ 0.05. Scale bar red colour refers to over-expression; blue colour refers to under-expression. **c,d.** GSEA dot plots showing the enrichment of 34 hallmark gene sets, based on the differentially expressed genes in the granuloma's central lesion (CL) compared to external wall (EW), in the Discovery Set. The top regulators with positive activation z scores and highest inverse Fisher's exact enrichment *p* values are shown. GSEA plots were built using the CL *vs* EW counts.

Figure 4. Spatial-transcriptional expression of the compartmentalized human TB granuloma. Boxplot of genes of interest expression in CL (yellow), IW (orange), EW (brown) and HT (green) in the Validation Set. Gene expression is represented as normalized count's log2. The gradient of expression throughout the three defined compartments of the human TB granuloma lesion was compared between compartments, and to HT. Statistical analysis was performed with a one-way ANOVA using Kruskal-

Wallis test for multiple comparisons showing: * refer to statistical differences between the CL, IW or EW compartments when compared to HT; # refer to statistical differences between CL or IW compartments when compared to the EW. $*^{,\#}$ p <0.05; $**^{\#\#}$ p <0.01; *** p <0.001; **** p <0.0001.

Figure 5. Estimated bacterial load in human lung tissue samples, determined using a modified MBL assay measuring *Mtb* 16S rRNA abundance. The number of *Mtb* bacilli was extrapolated from *Mtb* 16S ribosomal RNA RT-qPCR measurements (adjusted for RT-minus controls) in comparison to RNA isolated from a *Mtb* dilution series. *Mtb* -positive patient samples above the detection limit are marked with an asterisk. Error bars represent +1 SD from three independent technical replicates.

Figure 6. Transcriptomic differences between fast and slow sputum converters subset of patients. a. Heatmap showing top 24 DEGs comparing the human TB granuloma (CI + IW + EW pooled counts) to HT expression profile (Benjamini-Hochberg adjusted p-value <0.001). Gene expression differences is relative to Log2FC. Scale bar red colour refers to gene's over-expression; blue colour refers to under-expression. Heatmap dendrogram shows unsupervised hierarchical clustering of exhibited DEGs. **b, c.** GSEA dot plots showing the enrichment of hallmark gene sets comparing fast and slow SCC, in the Discovery Set. The top regulators with positive activation z scores and highest inverse Fisher's exact enrichment p values are shown. GSEA plots were built considering 1 sample/patient, using the mean of CL+IW+EW counts.

Figure 1

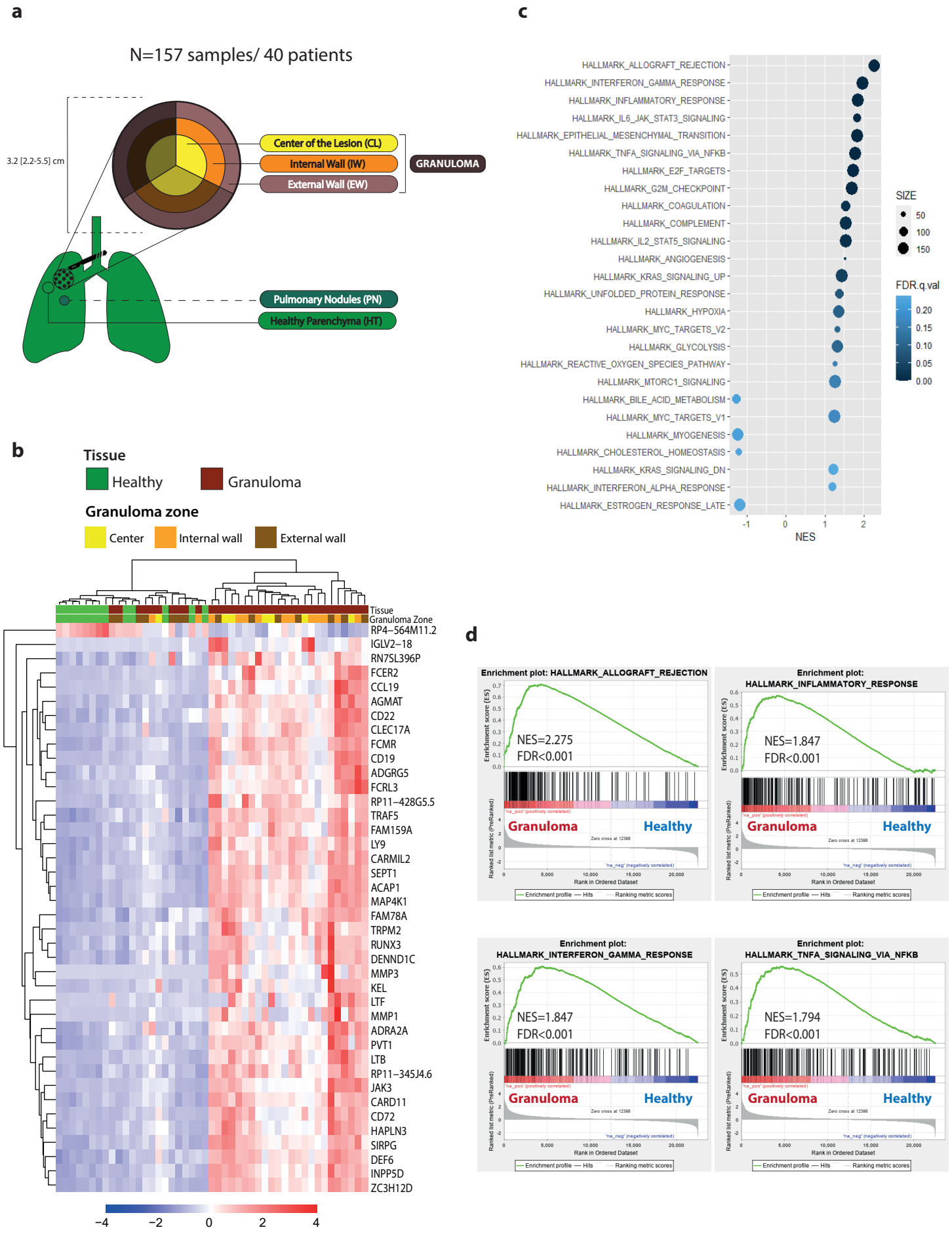

a
N=157 samples/ 40 patients
3.2 [2.2-5.5] cm
Center of the Lesion (CL)
Internal Wall (IW)
External Wall (EW)
GRANULOMA
Pulmonary Nodules (PN)
Healthy Parenchyma (HT)
b
Tissue
Healthy
Granuloma
Granuloma zone
Center
Internal wall
External wall
Tissue
Granuloma Zone
RP4−564M11.2
IGLV2−18
RN7SL396P
FCER2
CCL19
AGMAT
CD22
CLEC17A
FCMR
CD19
ADGRG5
FCRL3
RP11−428G5.5
TRAF5
FAM159A
LY9
CARMIL2
SEPT1
ACAP1
MAP4K1
FAM78A
TRPM2
RUNX3
DENND1C
MMP3
KEL
LTF
MMP1
ADRA2A
PVT1
LTB
RP11−345J4.6
JAK3
CARD11
CD72
HAPLN3
SIRPG
DEF6
INPP5D
ZC3H12D
−4
−2
0
2
4
c
HALLMARK_ALLOGRAFT_REJECTION
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_E2F_TARGETS
HALLMARK_G2M_CHECKPOINT
HALLMARK_COAGULATION
HALLMARK_COMPLEMENT
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_ANGIOGENESIS
HALLMARK_KRAS_SIGNALING_UP
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_HYPOXIA
HALLMARK_MYC_TARGETS_V2
HALLMARK_GLYCOLYSIS
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY
HALLMARK_MTORC1_SIGNALING
HALLMARK_BILE_ACID_METABOLISM
HALLMARK_MYC_TARGETS_V1
HALLMARK_MYOGENESIS
HALLMARK_CHOLESTEROL_HOMEOSTASIS
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_ESTROGEN_RESPONSE_LATE
-1
0
1
2
NES
SIZE
50
100
150
FDR.q.val
0.20
0.15
0.10
0.05
0.00
d
Enrichment plot: HALLMARK_ALLOGRAFT_REJECTION
NES=2.275
FDR<0.001
Enrichment plot: HALLMARK_INFLAMMATORY_RESPONSE
NES=1.847
FDR<0.001
Enrichment plot: HALLMARK_INTERFERON_GAMMA_RESPONSE
NES=1.847
FDR<0.001
Enrichment plot: HALLMARK_TNFA_SIGNALING_VIA_NFKB
NES=1.794
FDR<0.001
Granuloma
Healthy
Enrichment score (ES)
Ranked list metric (PreRanked)
Rank in Ordered Dataset
Enrichment profile
Hits
Ranking metric scores

Figure 2

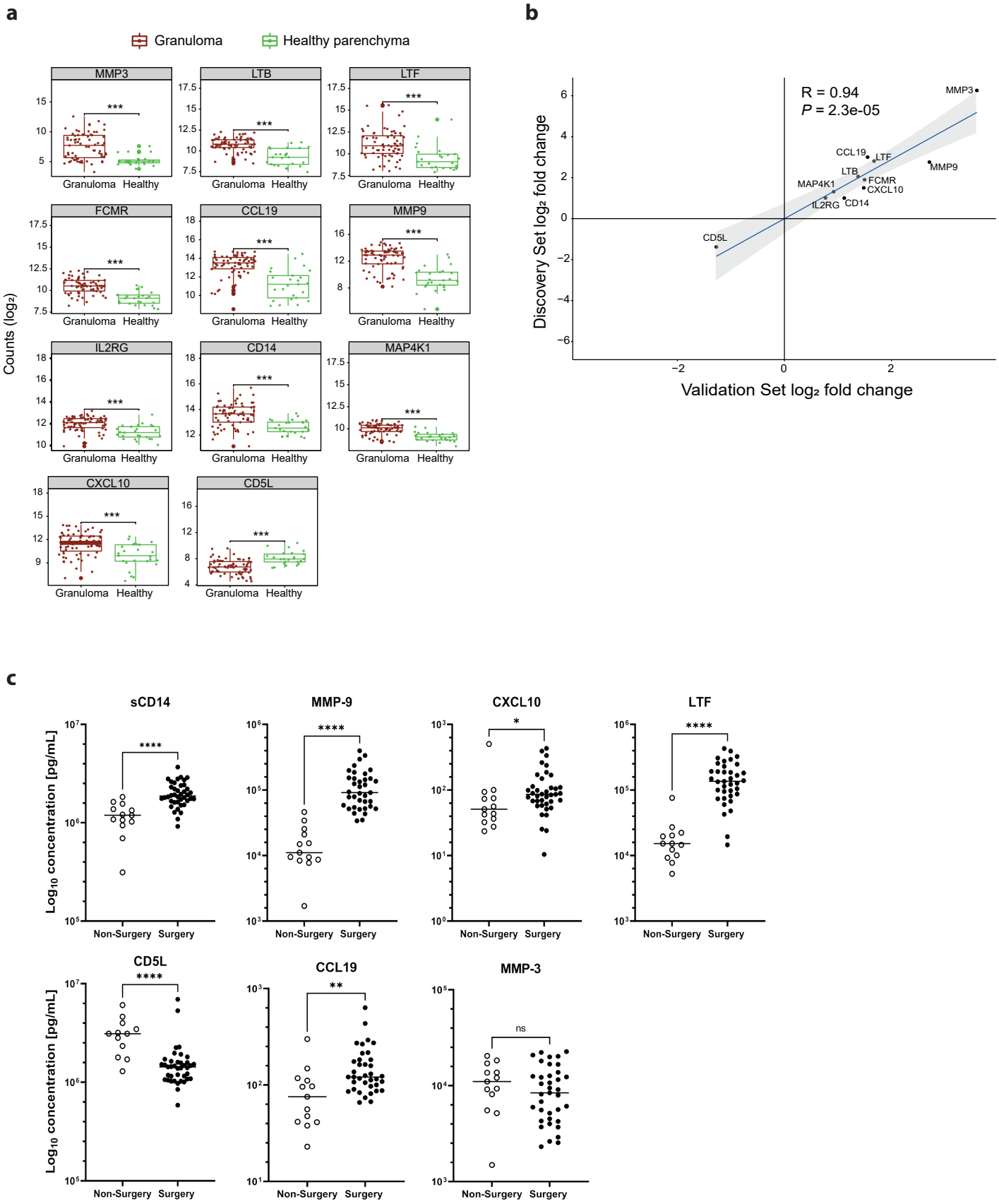

a
Granuloma
Healthy parenchyma
MMP3
LTB
LTF
FCMR
CCL19
MMP9
IL2RG
CD14
MAP4K1
CXCL10
CD5L
Counts (log2)
Granuloma
Healthy
***
b
R = 0.94
P = 2.3e-05
Discovery Set log2 fold change
Validation Set log2 fold change
MMP3
CCL19
LTF
MMP9
LTB
FCMR
MAP4K1
CXCL10
IL2RG
CD14
CD5L
c
sCD14
MMP-9
CXCL10
LTF
CD5L
CCL19
MMP-3
Log10 concentration [pg/mL]
Non-Surgery
Surgery
****
*
**
ns

Figure 3

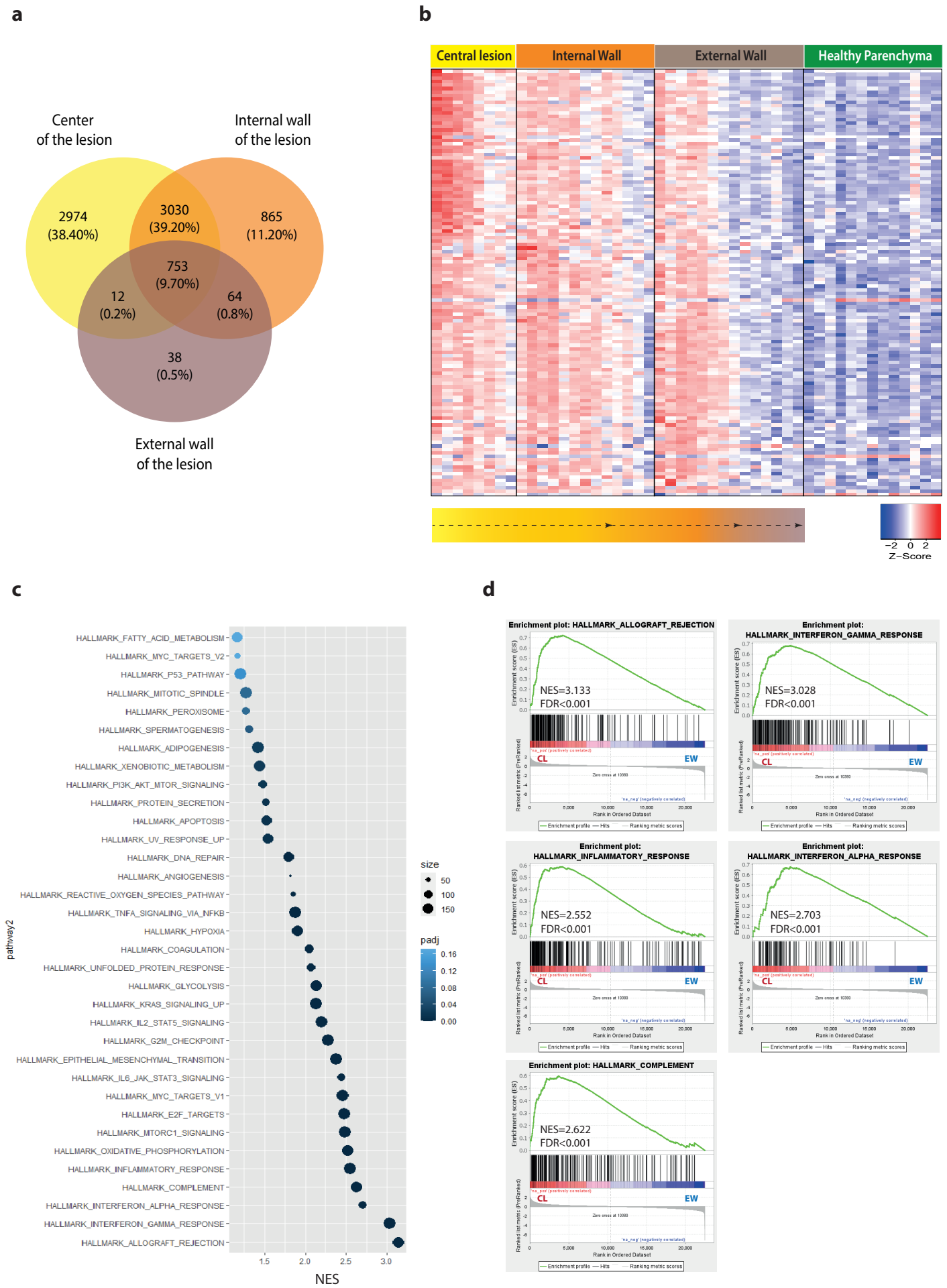

a
Center of the lesion
Internal wall of the lesion
External wall of the lesion
2974 (38.40%)
3030 (39.20%)
865 (11.20%)
753 (9.70%)
12 (0.2%)
64 (0.8%)
38 (0.5%)
b
Central lesion
Internal Wall
External Wall
Healthy Parenchyma
−2 0 2
Z−Score
c
HALLMARK_FATTY_ACID_METABOLISM
HALLMARK_MYC_TARGETS_V2
HALLMARK_P53_PATHWAY
HALLMARK_MITOTIC_SPINDLE
HALLMARK_PEROXISOME
HALLMARK_SPERMATOGENESIS
HALLMARK_ADIPOGENESIS
HALLMARK_XENOBIOTIC_METABOLISM
HALLMARK_PI3K_AKT_MTOR_SIGNALING
HALLMARK_PROTEIN_SECRETION
HALLMARK_APOPTOSIS
HALLMARK_UV_RESPONSE_UP
HALLMARK_DNA_REPAIR
HALLMARK_ANGIOGENESIS
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_HYPOXIA
HALLMARK_COAGULATION
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_GLYCOLYSIS
HALLMARK_KRAS_SIGNALING_UP
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_G2M_CHECKPOINT
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_MYC_TARGETS_V1
HALLMARK_E2F_TARGETS
HALLMARK_MTORC1_SIGNALING
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_COMPLEMENT
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_ALLOGRAFT_REJECTION
pathway2
1.5
2.0
2.5
3.0
NES
size
50
100
150
padj
0.16
0.12
0.08
0.04
0.00
d
Enrichment plot: HALLMARK_ALLOGRAFT_REJECTION
NES=3.133
FDR<0.001
Enrichment plot: HALLMARK_INTERFERON_GAMMA_RESPONSE
NES=3.028
FDR<0.001
Enrichment plot: HALLMARK_INFLAMMATORY_RESPONSE
NES=2.552
FDR<0.001
Enrichment plot: HALLMARK_INTERFERON_ALPHA_RESPONSE
NES=2.703
FDR<0.001
Enrichment plot: HALLMARK_COMPLEMENT
NES=2.622
FDR<0.001
CL
EW
Enrichment score (ES)
Rank in Ordered Dataset
Ranked list metric (PreRanked)
Enrichment profile
Hits
Ranking metric scores

Figure 4

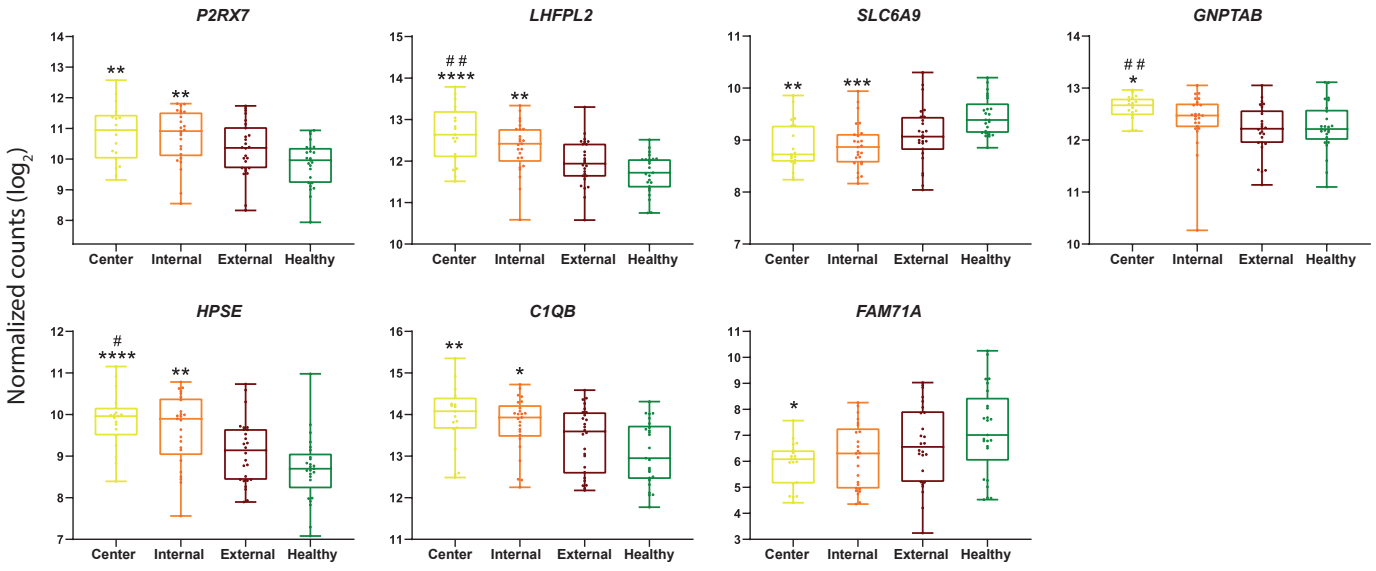

Normalized counts (log2)
P2RX7
LHFPL2
SLC6A9
GNPTAB
HPSE
C1QB
FAM71A
Center
Internal
External
Healthy

Figure 5

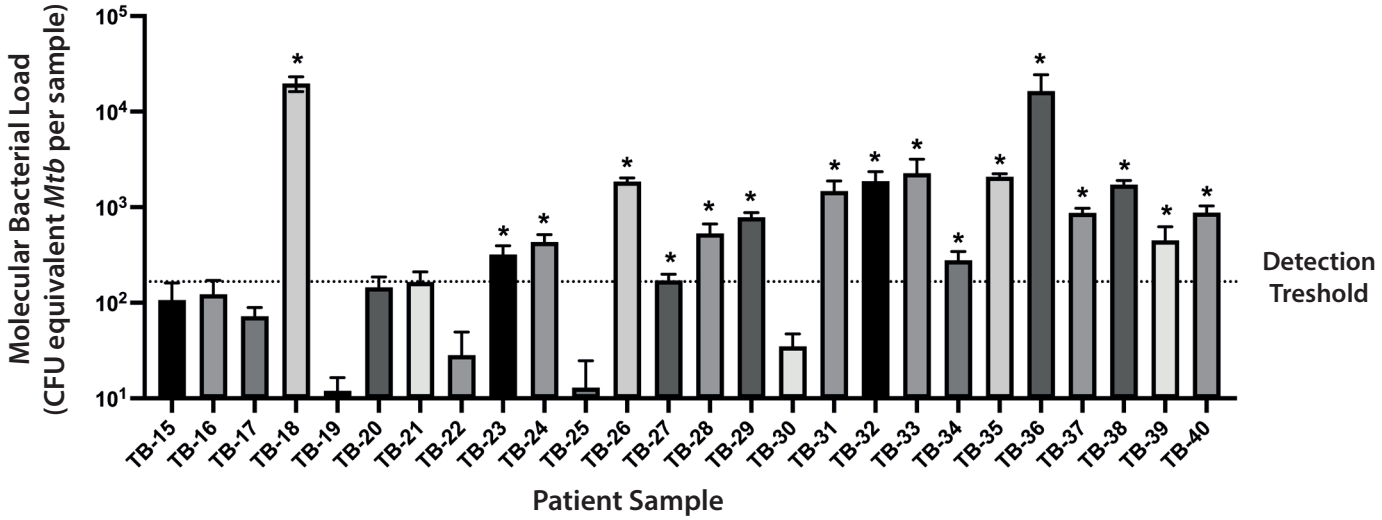
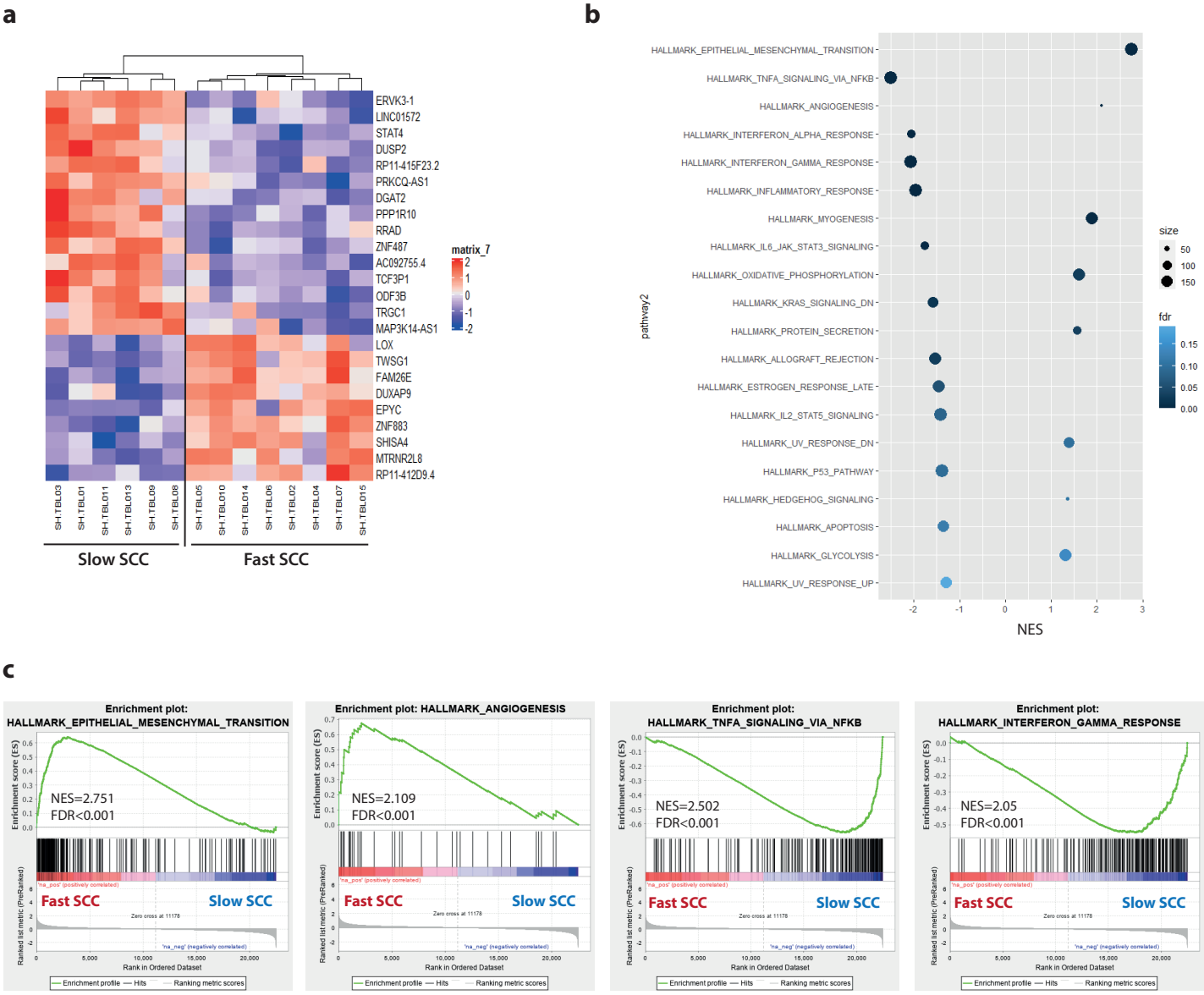

Molecular Bacterial Load
(CFU equivalent *Mtb* per sample)
10^5
10^4
10^3
10^2
10^1
Detection
Treshold
TB-15
TB-16
TB-17
TB-18
TB-19
TB-20
TB-21
TB-22
TB-23
TB-24
TB-25
TB-26
TB-27
TB-28
TB-29
TB-30
TB-31
TB-32
TB-33
TB-34
TB-35
TB-36
TB-37
TB-38
TB-39
TB-40
Patient Sample

Figure 6


a
ERVK3-1
LINC01572
STAT4
DUSP2
RP11-415F23.2
PRKCQ-AS1
DGAT2
PPP1R10
RRAD
ZNF487
AC092755.4
TCF3P1
ODF3B
TRGC1
MAP3K14-AS1
LOX
TWSG1
FAM26E
DUXAP9
EPYC
ZNF883
SHISA4
MTRNR2L8
RP11-412D9.4
matrix_7
2
1
0
-1
-2
SH.TBL03
SH.TBL01
SH.TBL011
SH.TBL013
SH.TBL09
SH.TBL08
SH.TBL05
SH.TBL010
SH.TBL014
SH.TBL06
SH.TBL02
SH.TBL04
SH.TBL07
SH.TBL015
Slow SCC
Fast SCC
b
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_ANGIOGENESIS
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_MYOGENESIS
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_PROTEIN_SECRETION
HALLMARK_ALLOGRAFT_REJECTION
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_UV_RESPONSE_DN
HALLMARK_P53_PATHWAY
HALLMARK_HEDGEHOG_SIGNALING
HALLMARK_APOPTOSIS
HALLMARK_GLYCOLYSIS
HALLMARK_UV_RESPONSE_UP
pathway2
NES
size
50
100
150
fdr
0.15
0.10
0.05
0.00
c
Enrichment plot:
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
NES=2.751
FDR<0.001
Enrichment plot: HALLMARK_ANGIOGENESIS
NES=2.109
FDR<0.001
Enrichment plot:
HALLMARK_TNFA_SIGNALING_VIA_NFKB
NES=2.502
FDR<0.001
Enrichment plot:
HALLMARK_INTERFERON_GAMMA_RESPONSE
NES=2.05
FDR<0.001
Fast SCC
Slow SCC
Enrichment score (ES)
Ranked list metric (PreRanked)
Rank in Ordered Dataset
Enrichment profile
Hits
Ranking metric scores

Supplementary Information

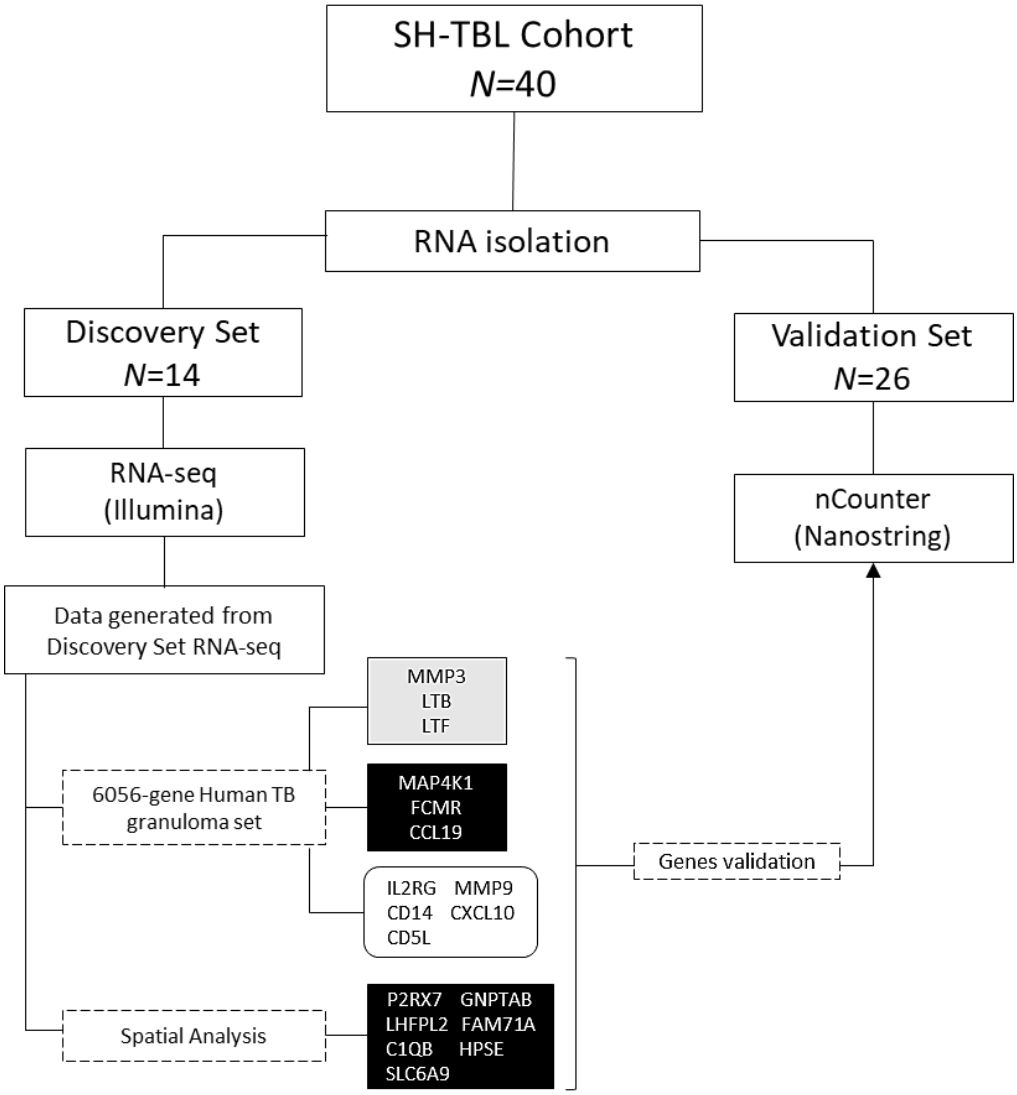
Unravelling the transcriptome of the human tuberculosis lesion and its clinical implications

Albert Despuig[1,2,&], Kaori L. Fonseca[1,2,&], Dominic Habgood-Coote[3], Álvaro Del Río-Álvarez[4], Juan Carrillo-Reixach[4,5], Lilibeth Arias[1,2], Nino Gogichadze[1,6], Aaron Goff[7], Leticia Muraro Wildner[7], Shota Gogishvili[8], Keti Nikolaishvili[8], Natalia Shubladze[8], Zaza Avaliani[8], Pere-Joan Cardona[1,2,9], Federico Martinón-Torres[2,10,11], Antonio Salas[2,11,12], Alberto Gómez-Carballa[2,10,11], Simon J Waddell[7], Carolina Armengol[4,5], Sergo Vashakidze[8,13], Myrsini Kaforou[3,14], Cristina Vilaplana[1,2,15]*

1. Experimental Tuberculosis Unit (UTE). Fundació Institut Germans Trias i Pujol (IGTP). Universitat Autònoma de Barcelona (UAB). Edifici Mar. Can Ruti Campus. Crtra. de Can Ruti, Camí de les Escoles, s/n. 08916, Badalona; Catalonia; Spain.
2. Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES). Av. Monforte de Lemos, 3-5. Pabellón 11. Planta 0. 28029, Madrid, Spain.
3. Department of Infectious Disease, Imperial College London, London, United Kingdom.
4. Childhood Liver Oncology Group, Germans Trias i Pujol Research Institute (IGTP), Program of Predictive and Personalized Medicine of Cancer (PMPCC), Edifici Muntanya. Can Ruti Campus. Ctra. de Can Ruti, camí de les Escoles, s/n, 08916 Badalona, Catalonia, Spain.
5. Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREhD). Av. Monforte de Lemos, 3-5. Pabellón 11. Planta 0. 28029, Madrid, Spain.
6. UCBL, UnivLyon, Université Claude Bernard Lyon 1 (UCBL1), 69100 Villeurbanne, France
7. Global Health and Infection, Brighton and Sussex Medical School, University of Sussex, Falmer, Brighton, BN1 9PX, UK.
8. National Center for Tuberculosis and Lung Diseases (NCTLD). 50, Maruashvili Str. 0101 Tbilisi, Georgia.
9. Hospital Universitari Germans Trias i Pujol (HUGTIP). Microbiological dept. Crtra. Del Canyet, s/n. 08916. Badalona, Catalonia; Spain.
10. Genetics, Vaccines and Pediatric Infectious Diseases Research Group (GENVIP), Instituto de Investigación Sanitaria de Santiago (IDIS) and Universidad de Santiago de Compostela (USC), 15706, Galicia, Spain.
11. Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela (SERGAS), 15706, Galicia, Spain.
12. Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), 15706, Galicia, Spain.
13. The University of Georgia, Tbilisi, Georgia.
14. Centre for Paediatrics and Child Health, Imperial College London, United Kingdom
15. Direcció Clínica Territorial de Malalties Infeccioses i Salut Internacional. Gerència territorial Metropolitana Nord, Institut Català de la Salut. Can Ruti Campus. Crtra. del Canyet, s/n. 08916, Badalona; Catalonia; Spain

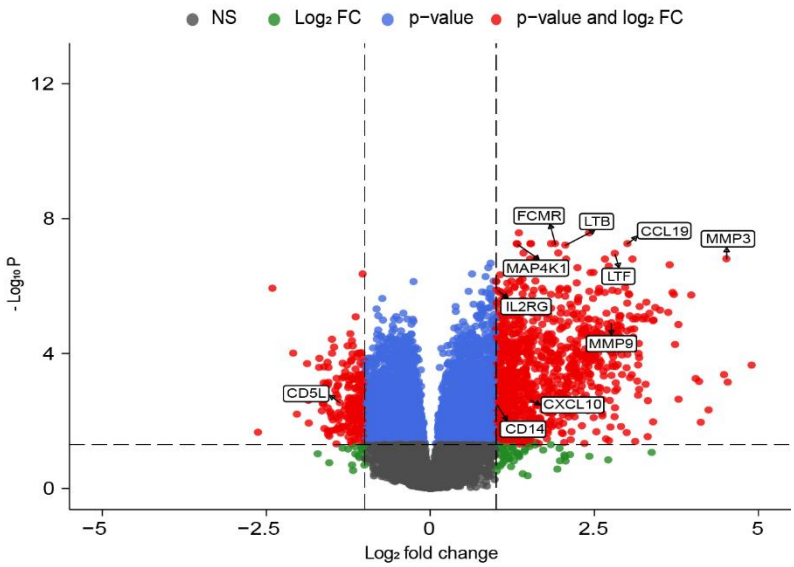
Corresponding Author*:
Cristina Vilaplana, MD, PhD
ORCID: https://orcid.org/0000-0002-2808-7270
Principal Investigator
Tuberculosis Experimental Unit (UTE)
Fundació Institut Germans Trias i Pujol (IGTP)
Campus Can Ruti. Crtra. De Can Ruti, Camí de les Escoles, s/n.
Badalona, Barcelona, 08916
Catalonia (Spain)
Telephone: +34 93 033 0527
cvilaplana@igtp.cat

[&]These authors contributed equally: Albert Despuig and Kaori L. Fonseca

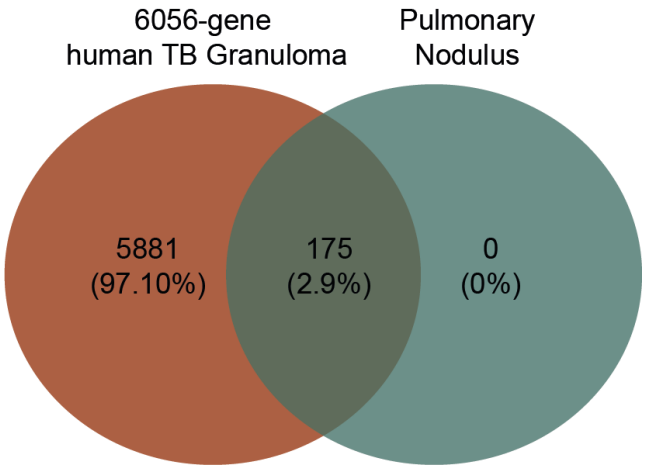


Supplementary Figure 1. Discovery Set genes of interest selection criteria and inclusion in Validation Set. After obtaining total-RNA, patients were divided into the Discovery Set (14 patients) and the Validation Set (26 patients). RNA-seq was performed on the Discovery Set, and after bioinformatic analysis, two streams of analysis generated: the TB granuloma lesion *vs* healthy tissue; and Central Lesion (CL), Internal Wall (IW), External Wall (EW) *vs* healthy tissue. Following a statistical and pathophysiological rationale, the genes better including the main pathophysiological features of the Discovery Set and the top-ranked genes from each signature were down-selected. In grey are represented the genes selected accordingly to a pathophysiological rationale after the GSEA analysis (pro-inflammatory and tissue remodeling gene sets); in black, the genes

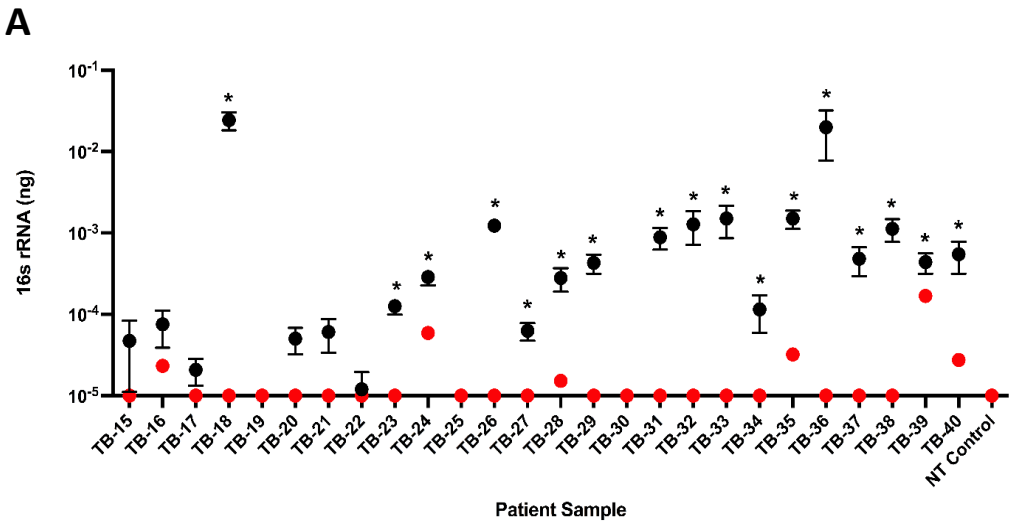
selected according to a statistical rationale from the top 20 SDE genes (logFC >+1/-1, baseMean >50); in white, the genes encoding described circulant prognostic/diagnostic biomarkers with identified features in TB pathogenesis. The listed genes of interest derived from the two signatures were validated in Validation Set samples.

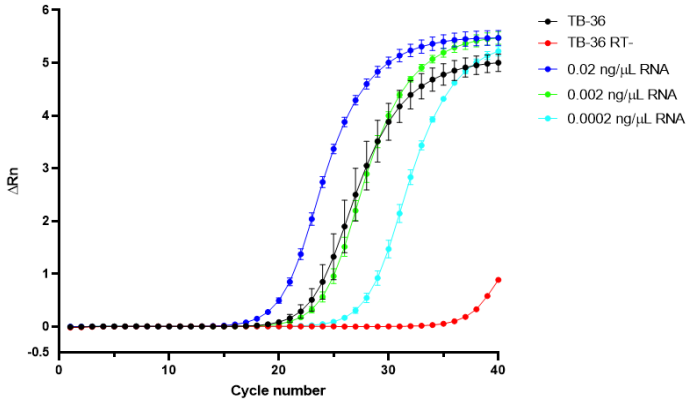


Supplementary Figure 2. Differential expression in the Discovery dataset of the granuloma tissue vs HT. 11 genes have been highlighted on this plot. Volcano plot of the granuloma *vs* healthy tissue differential expression in Discovery Set (log2 FC versus -log10 adjusted *p*-value), highlighting the TB lesion 11 genes. Significance cut-off was set at Benjamini-Hochberg adjusted *p*-value ≤ 0.05(gray color as non-significant gene) (blue color as DEGs). Log2 FC cut-off (green) was set at +1/-1, as per gene of interest selection criteria. Both *p*-value and log2 FC cut-off (red).



Supplementary Figure 3. Venn diagram of gene intersection between 6056-gene human TB granuloma set and the DEG from pulmonary nodulus *vs* healthy tissue comparison. The outcome of gene intersection between both gene sets revealed that PN had no specific genes as compared with granuloma. A total of 175 DGES genes encountered in PN overlapped with the 6056-gene human TB granuloma.





C

Name	Primer sequence
H37Rv 16SF	5'-GTGATCTGCCCTGCACTTC-3'
H37Rv 16SR	5'-ATCCCACACCGCTAAAGCG-3'

Supplementary Figure 4. MBL assay of *Mtb* 16S rRNA abundance from lesion tissue samples from 27 patient. (A) *Mtb* 16S rRNA abundance (ng) in patient samples estimated from an *Mtb* RNA standard curve. The average measurements from three independent technical replicates are marked with standard deviation (black) in comparison to RT-minus controls from the same samples (red). *Mtb* positive samples above the detection limit of the assay are marked with an asterisk. Values below the no template control (NT) have been adjusted to $1x10^{-5}$ to visualise. **(B)** Representative RT-qPCR amplification plot for sample TB-36, showing amplification of 16S rRNA (black) over 40 cycles in comparison to RT-minus sample control (red) and *Mtb* H37Rv RNA dilution series markers. **(C)** 16S rRNA *Mtb* complex-specific primer sequences[53] used at a final concentration of 200nM.