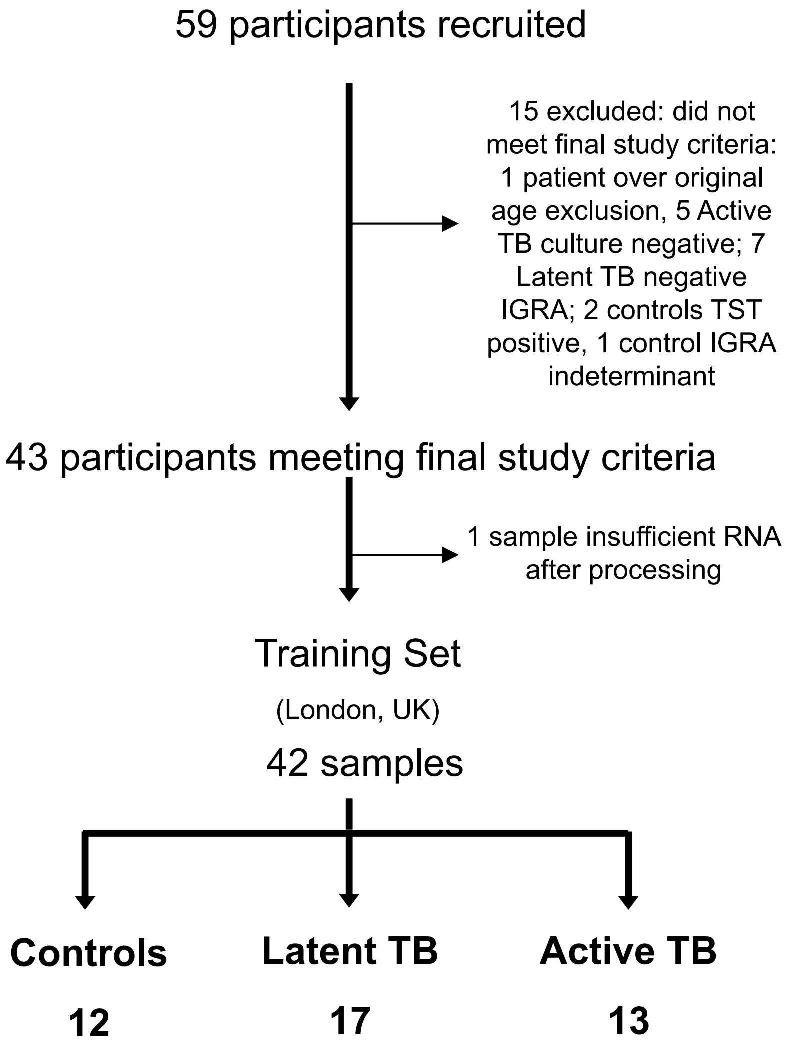
Supplementary Figure 1a



59 participants recruited

15 excluded: did not meet final study criteria: 1 patient over original age exclusion, 5 Active TB culture negative; 7 Latent TB negative IGRA; 2 controls TST positive, 1 control IGRA indeterminant

43 participants meeting final study criteria

1 sample insufficient RNA after processing

Training Set

(London, UK)

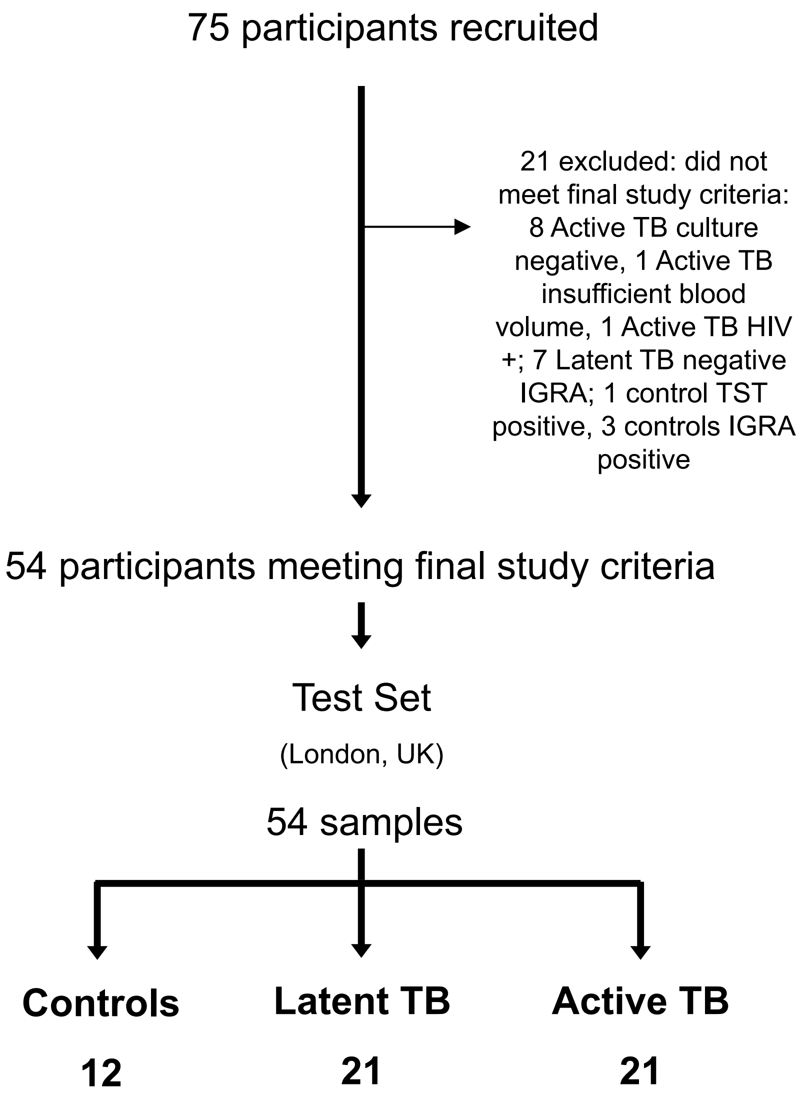
42 samples

Controls **12**

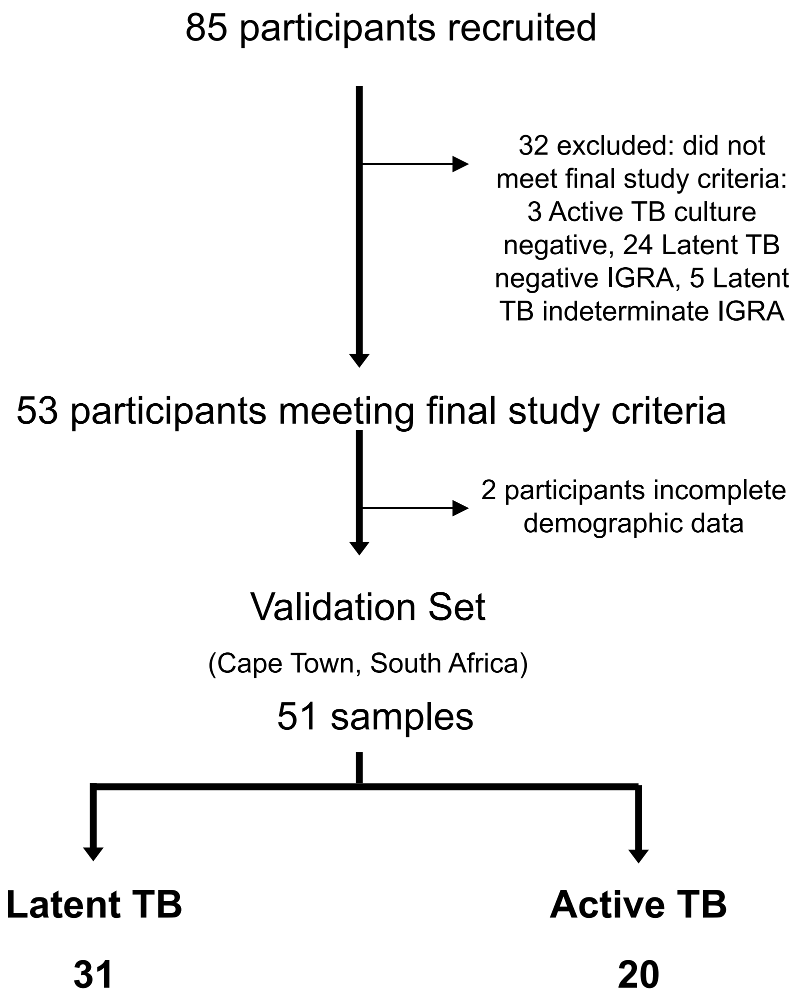
Latent TB **17**

Active TB **13**

Supplementary Figure 1b


75 participants recruited
21 excluded: did not meet final study criteria: 8 Active TB culture negative, 1 Active TB insufficient blood volume, 1 Active TB HIV +; 7 Latent TB negative IGRA; 1 control TST positive, 3 controls IGRA positive
54 participants meeting final study criteria
Test Set
(London, UK)
54 samples
Controls
12
Latent TB
21
Active TB
21

Supplementary Figure 1c



85 participants recruited

32 excluded: did not meet final study criteria: 3 Active TB culture negative, 24 Latent TB negative IGRA, 5 Latent TB indeterminate IGRA

53 participants meeting final study criteria

2 participants incomplete demographic data

Validation Set

(Cape Town, South Africa)

51 samples

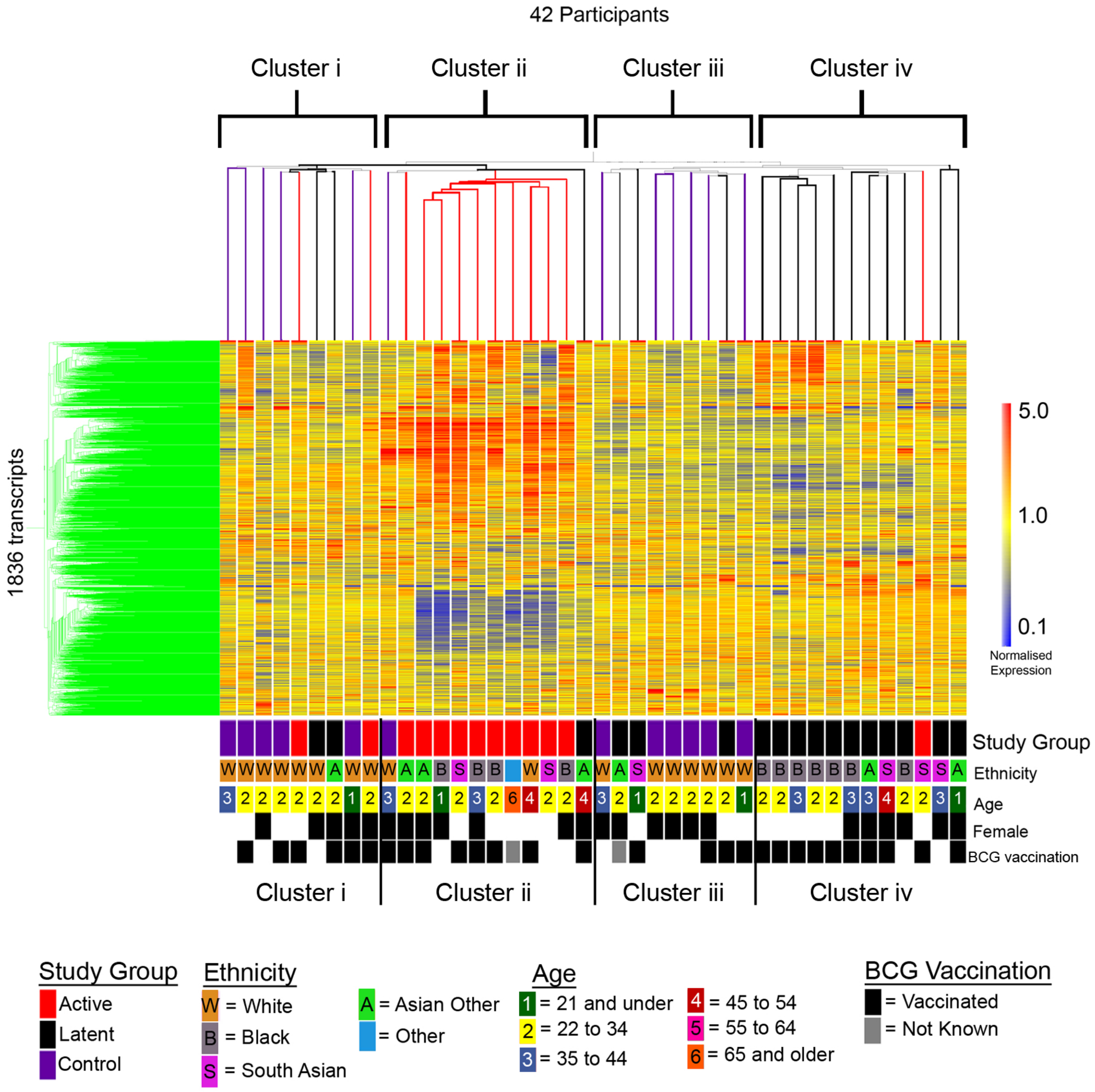
Latent TB

31

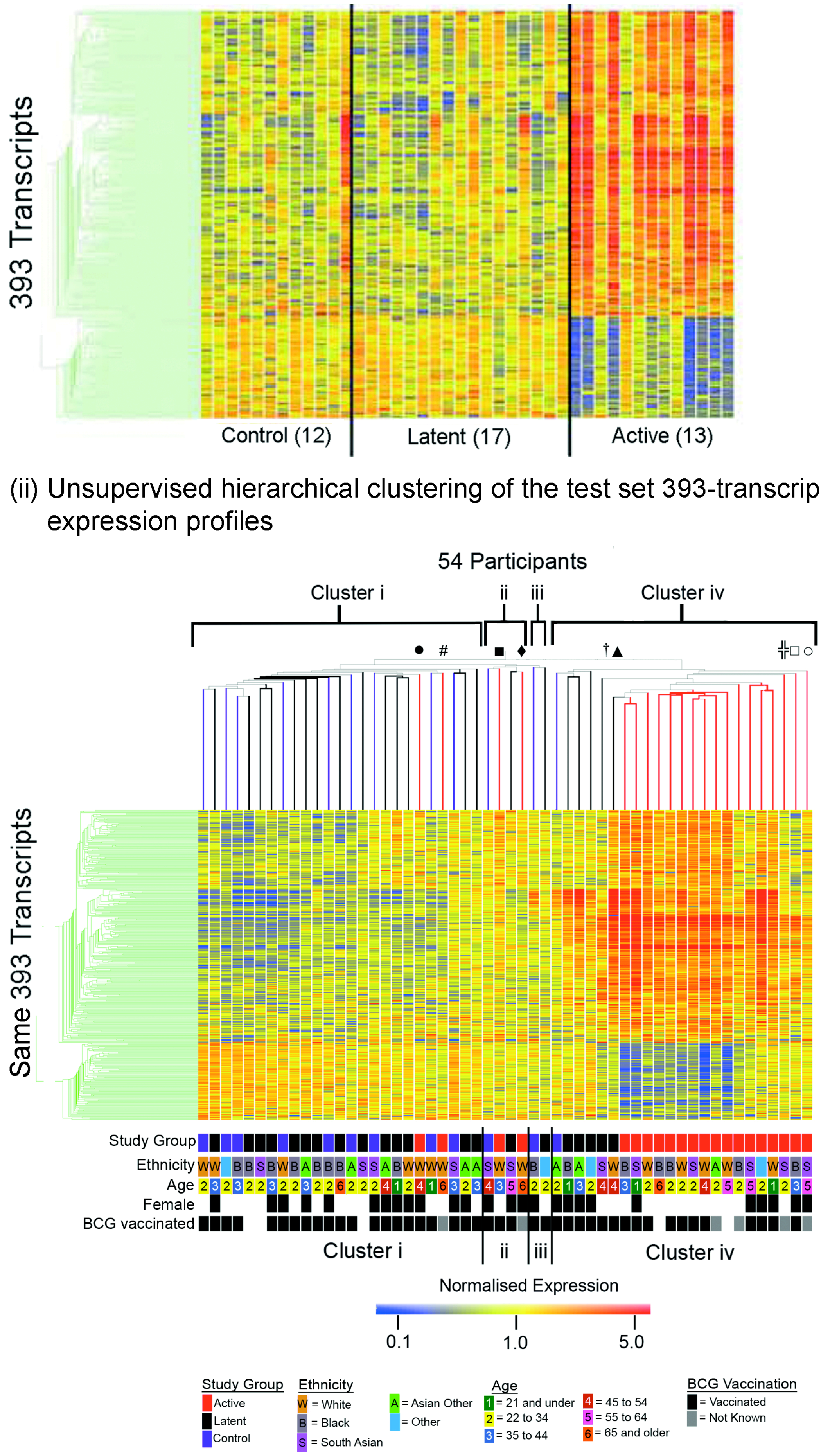
Active TB

20

Supplementary Figure 2a: Unsupervised hierarchical clustering of 1836-transcript expression
profiles



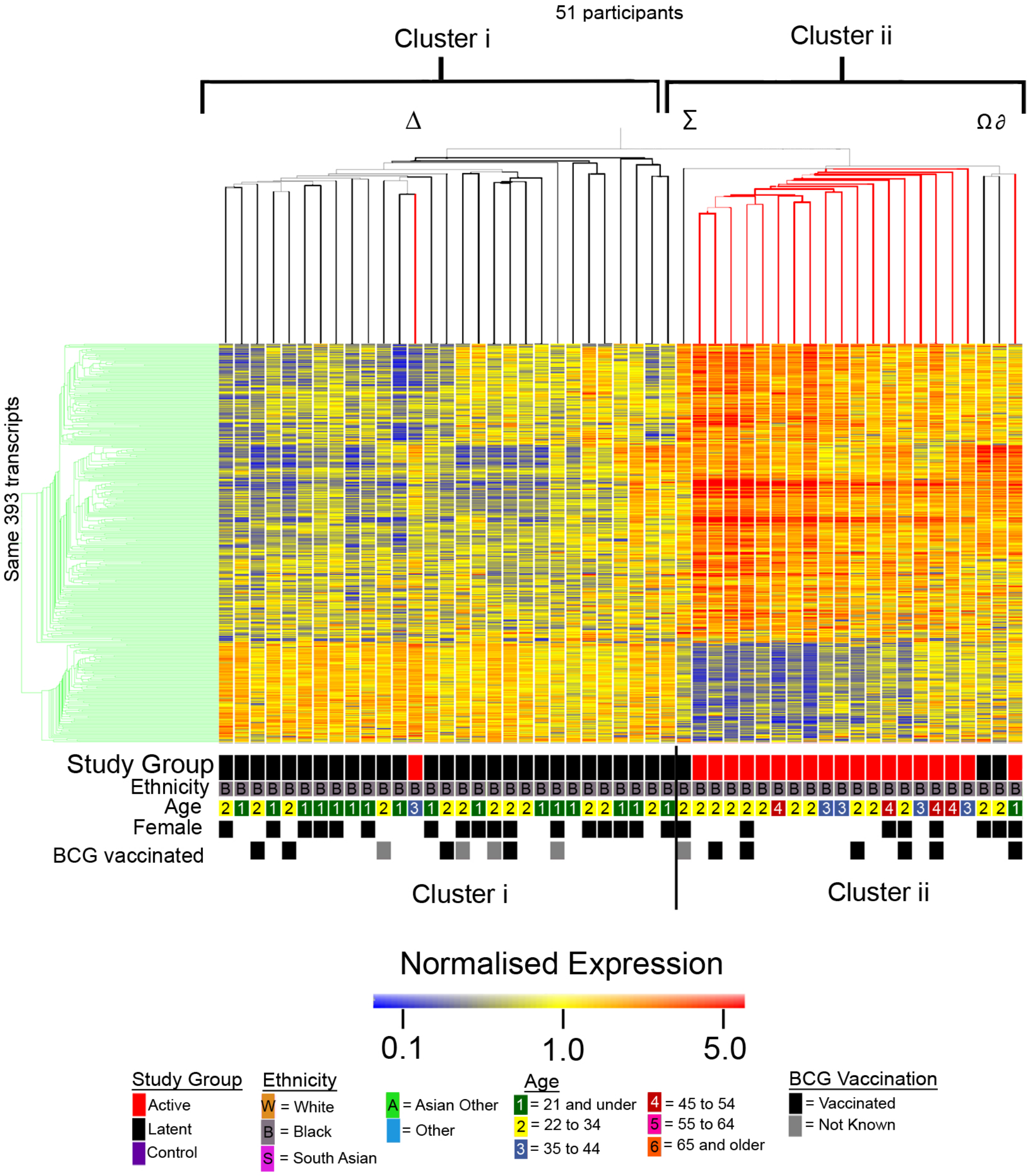
Supplementary Figure 2b: A distinct whole blood 393-gene transcriptional signature of active TB



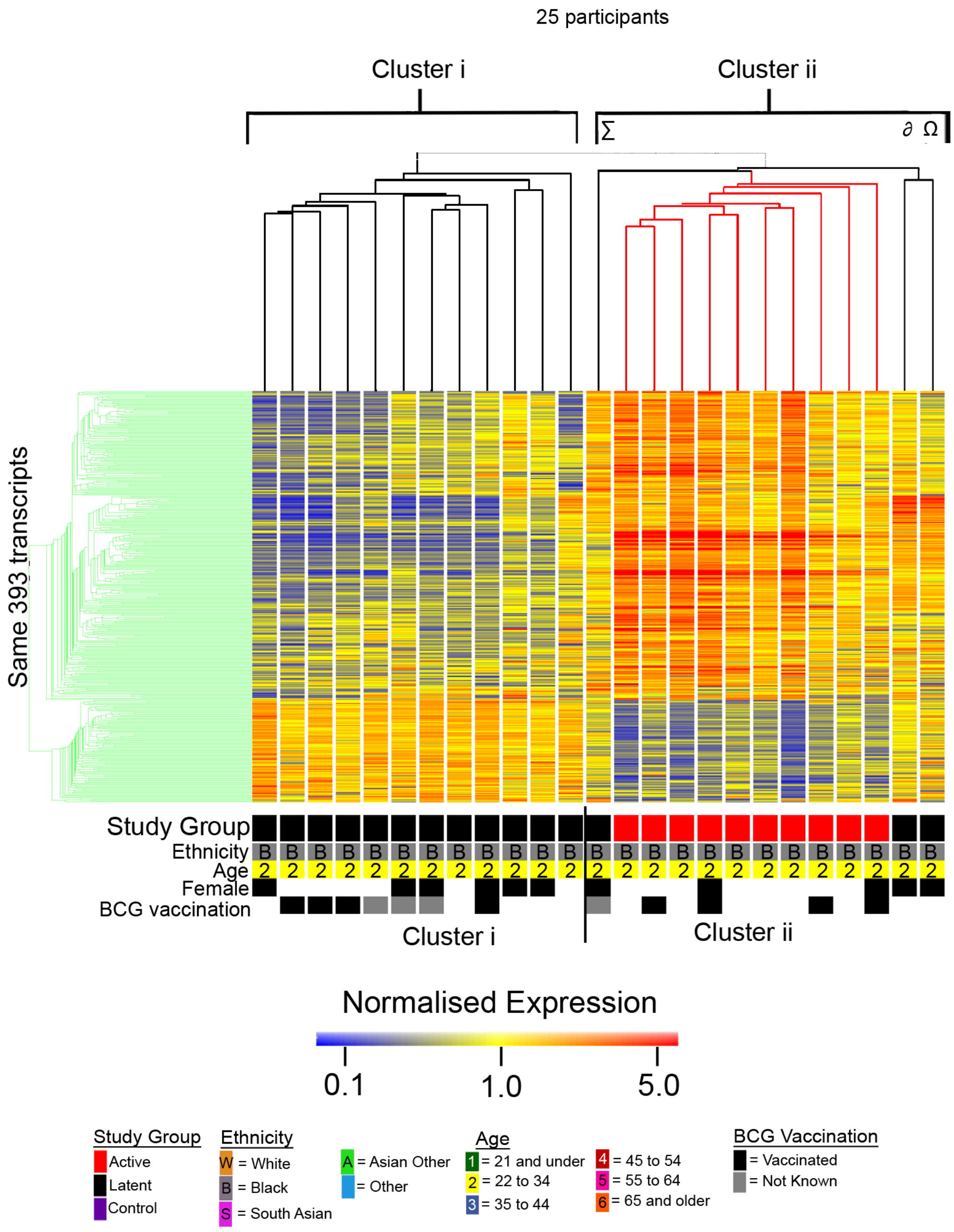
(i) The 393-gene transcriptional signature of active TB in the Training Set organized by hierarchical clustering

(ii) Unsupervised hierarchical clustering of the test set 393-transcript expression profiles

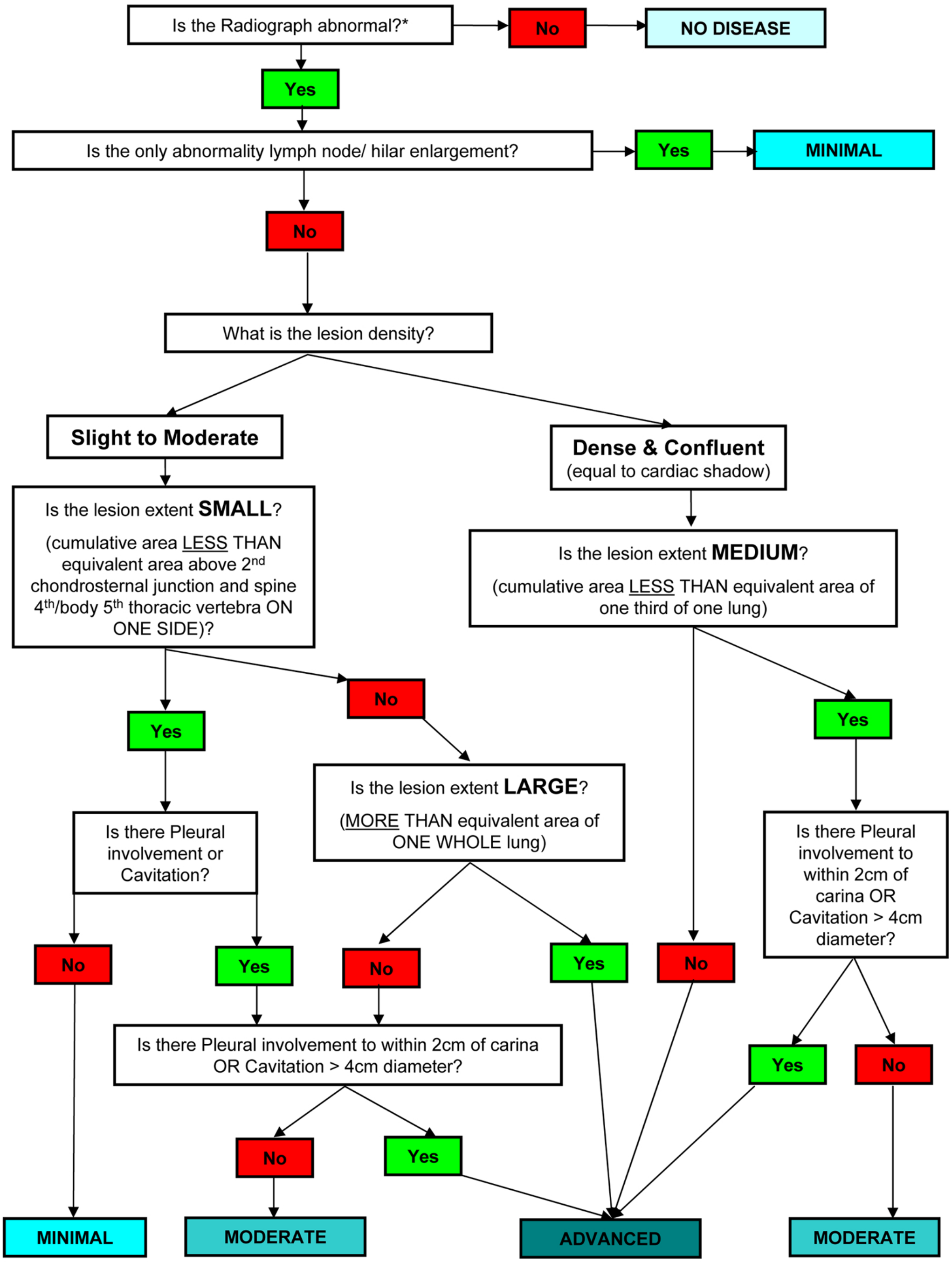
Supplementary Figure 2c: Unsupervised hierarchical clustering of the South African validation set 393-transcript expression profiles



Supplementary Figure 2d: Unsupervised hierarchical clustering of only those participants in the validation set aged between 22 and 34



Supplementary Figure 3a: The decision tree grading system developed for use in assessing the radiographic extent of disease.

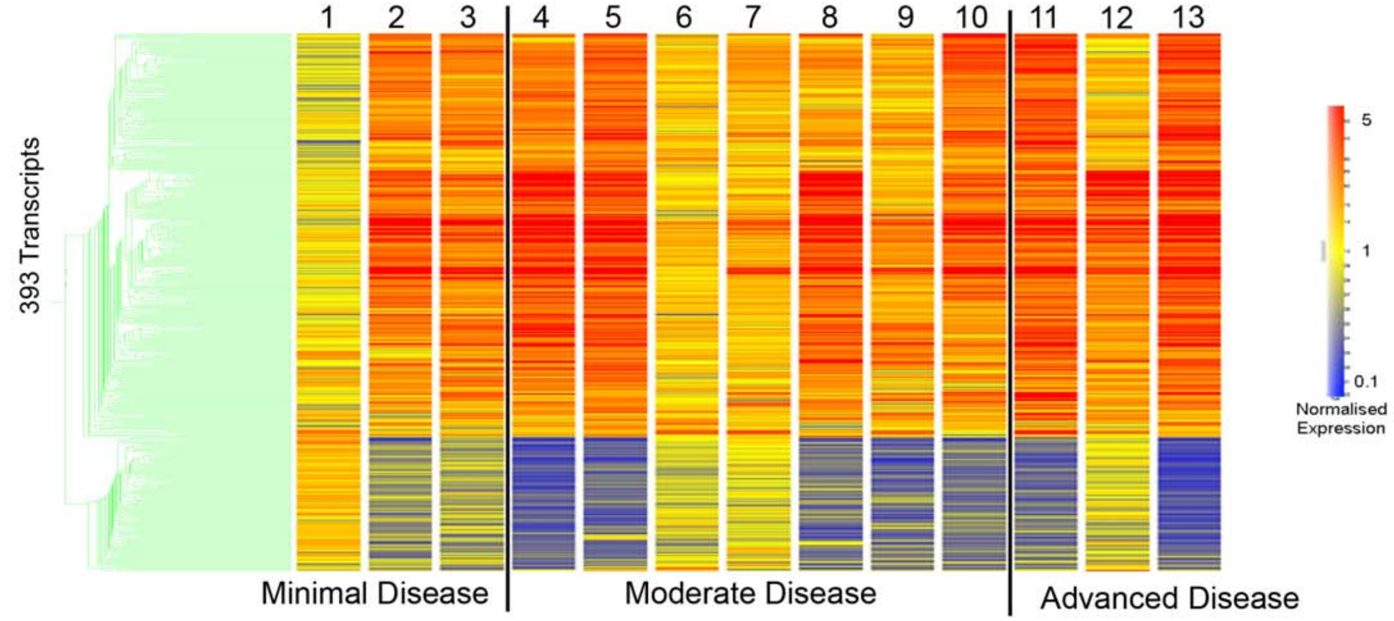

Is the Radiograph abnormal?*
No
NO DISEASE
Yes
Is the only abnormality lymph node/ hilar enlargement?
Yes
MINIMAL
No
What is the lesion density?
Slight to Moderate
Dense & Confluent
(equal to cardiac shadow)
Is the lesion extent SMALL?
(cumulative area LESS THAN equivalent area above 2nd chondrosternal junction and spine 4th/body 5th thoracic vertebra ON ONE SIDE)?
Is the lesion extent MEDIUM?
(cumulative area LESS THAN equivalent area of one third of one lung)
Yes
No
Yes
Is the lesion extent LARGE?
(MORE THAN equivalent area of ONE WHOLE lung)
Is there Pleural involvement or Cavitation?
Is there Pleural involvement to within 2cm of carina OR Cavitation > 4cm diameter?
No
Yes
No
Yes
No
Is there Pleural involvement to within 2cm of carina OR Cavitation > 4cm diameter?
Yes
No
No
Yes
MINIMAL
MODERATE
ADVANCED
MODERATE

*excluding non-pathological abnormality e.g. azygos lobe, raised hemidiaphragm

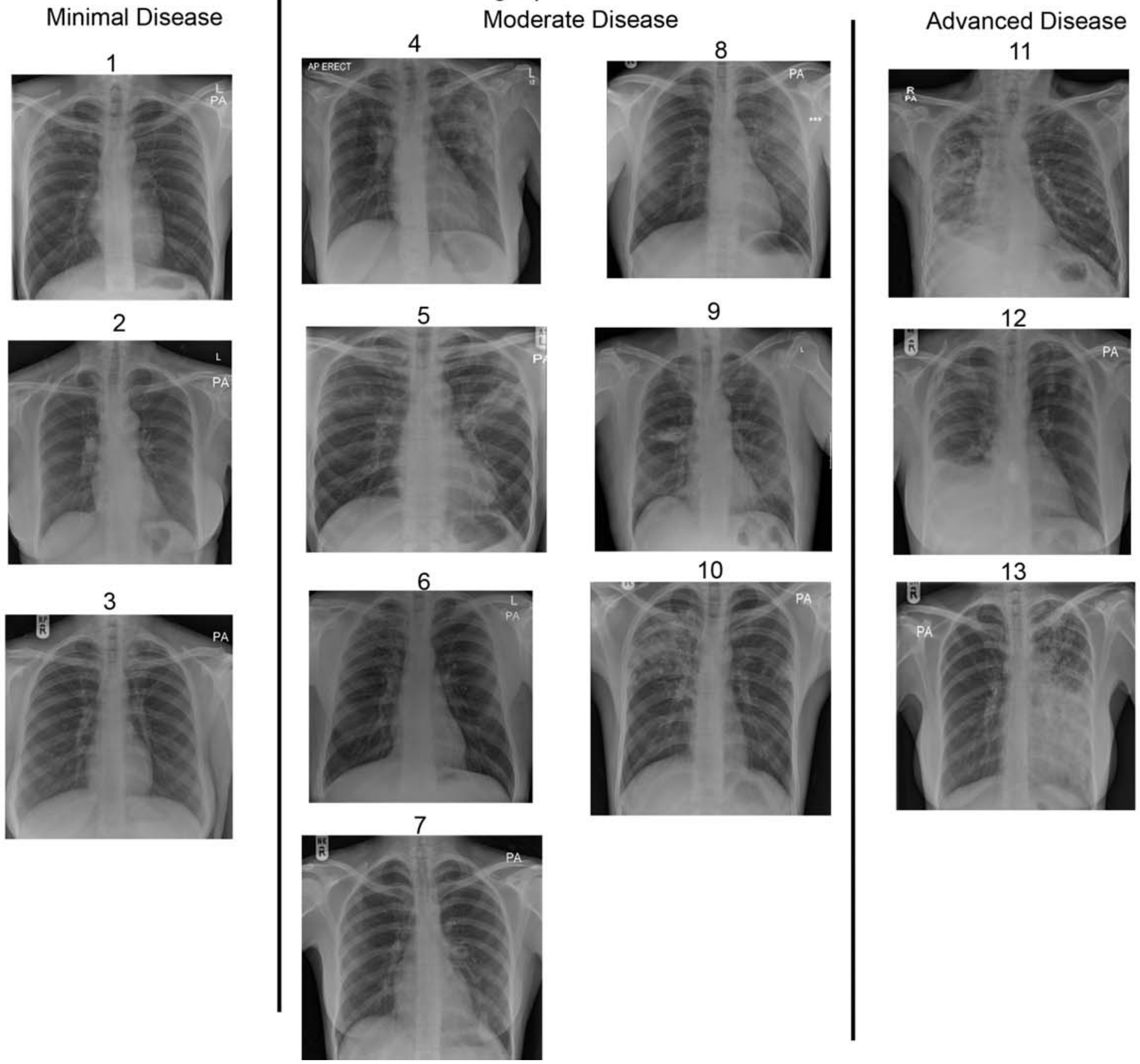
Developed and expanded from the text description definitions given in (Falk, O'Connor et al. 1969).

Supplementary Figure 3b: Comparison of Radiographic extent of Disease with 393 Transcript Profile

Active Patients From Training Set (UK) (13)

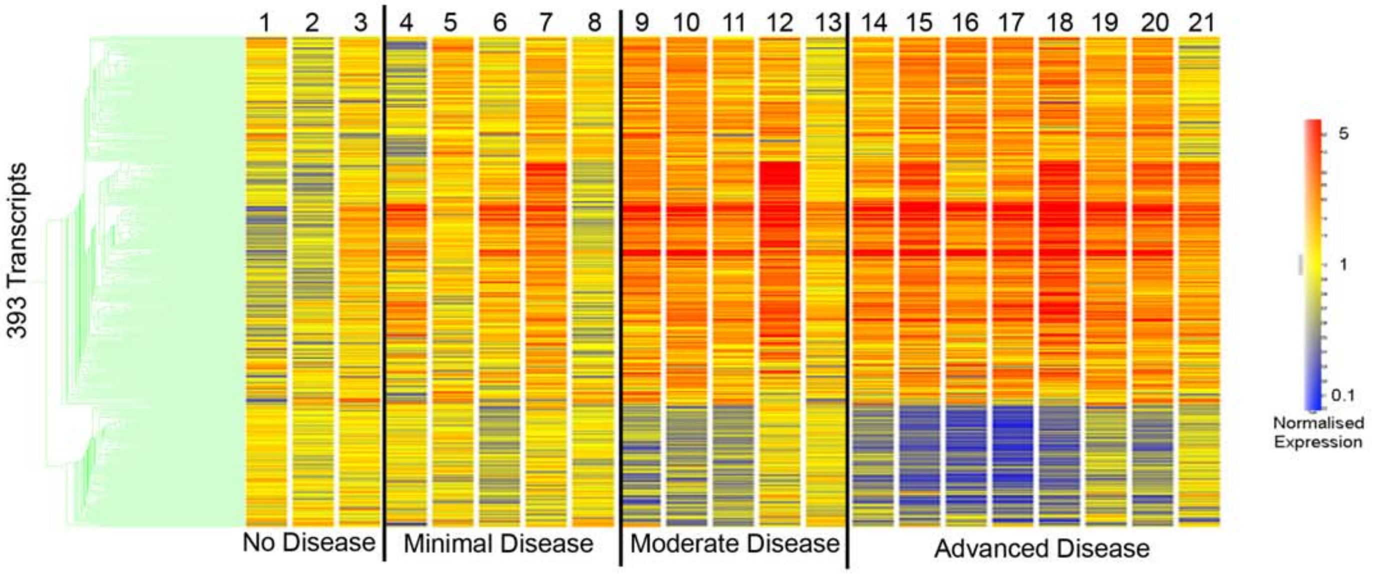


Radiographic extent of disease

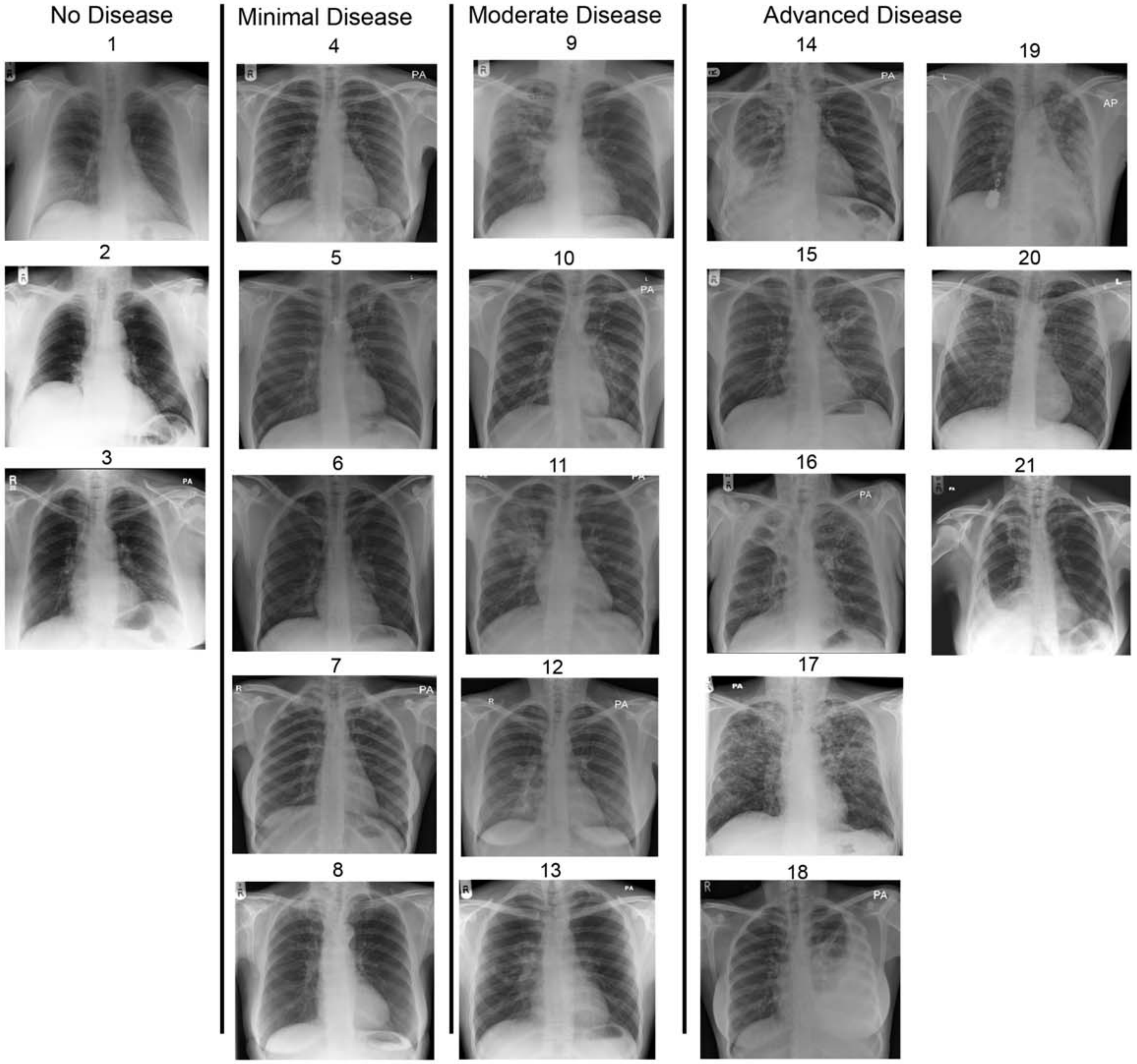


Supplementary Figure 3c: Comparison of Radiographic extent of Disease with 393 Transcript Profile

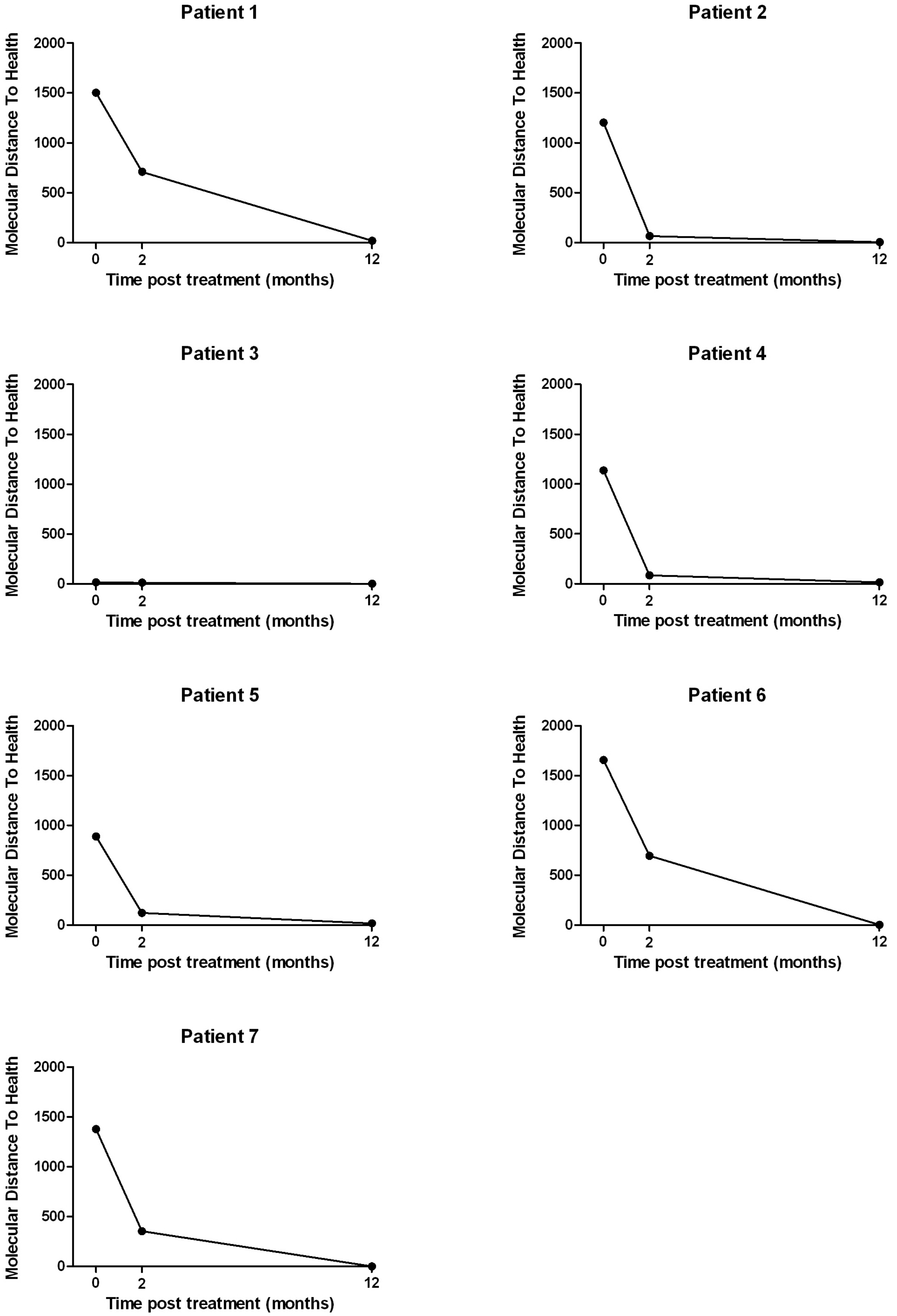
Active Patients From Test Set (UK) (21)



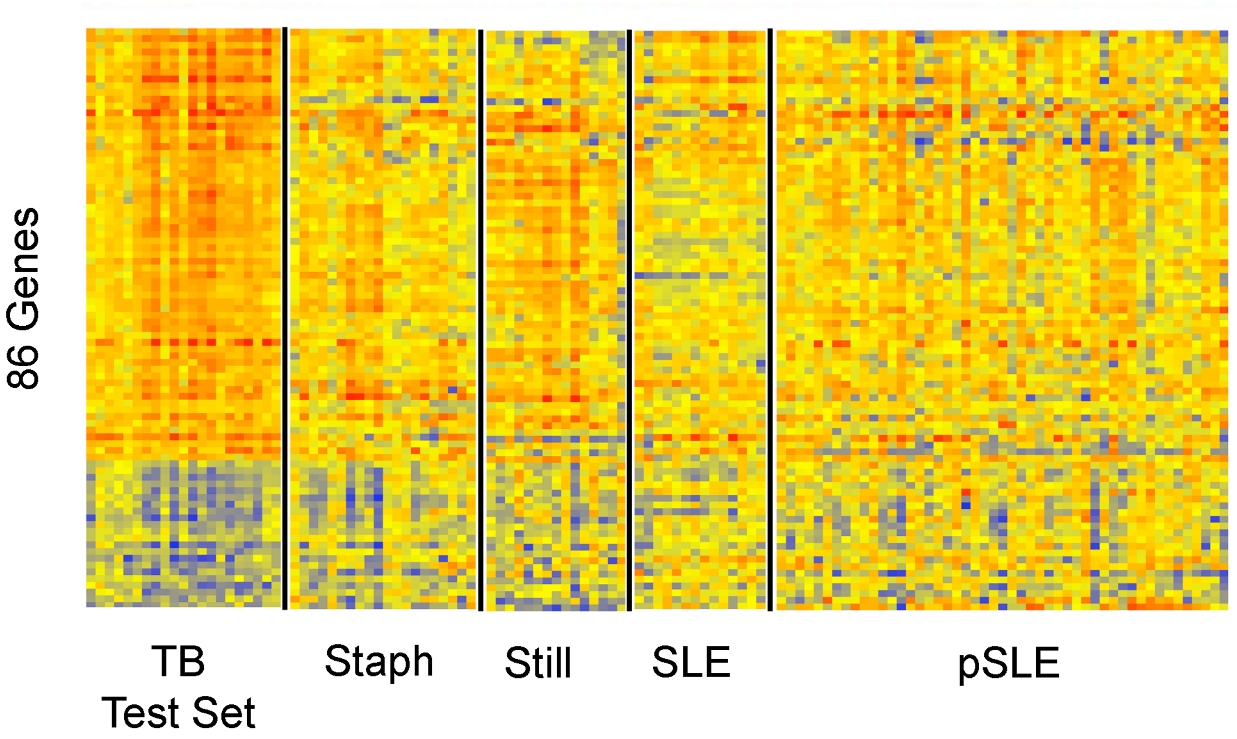
Radiographic extent of disease



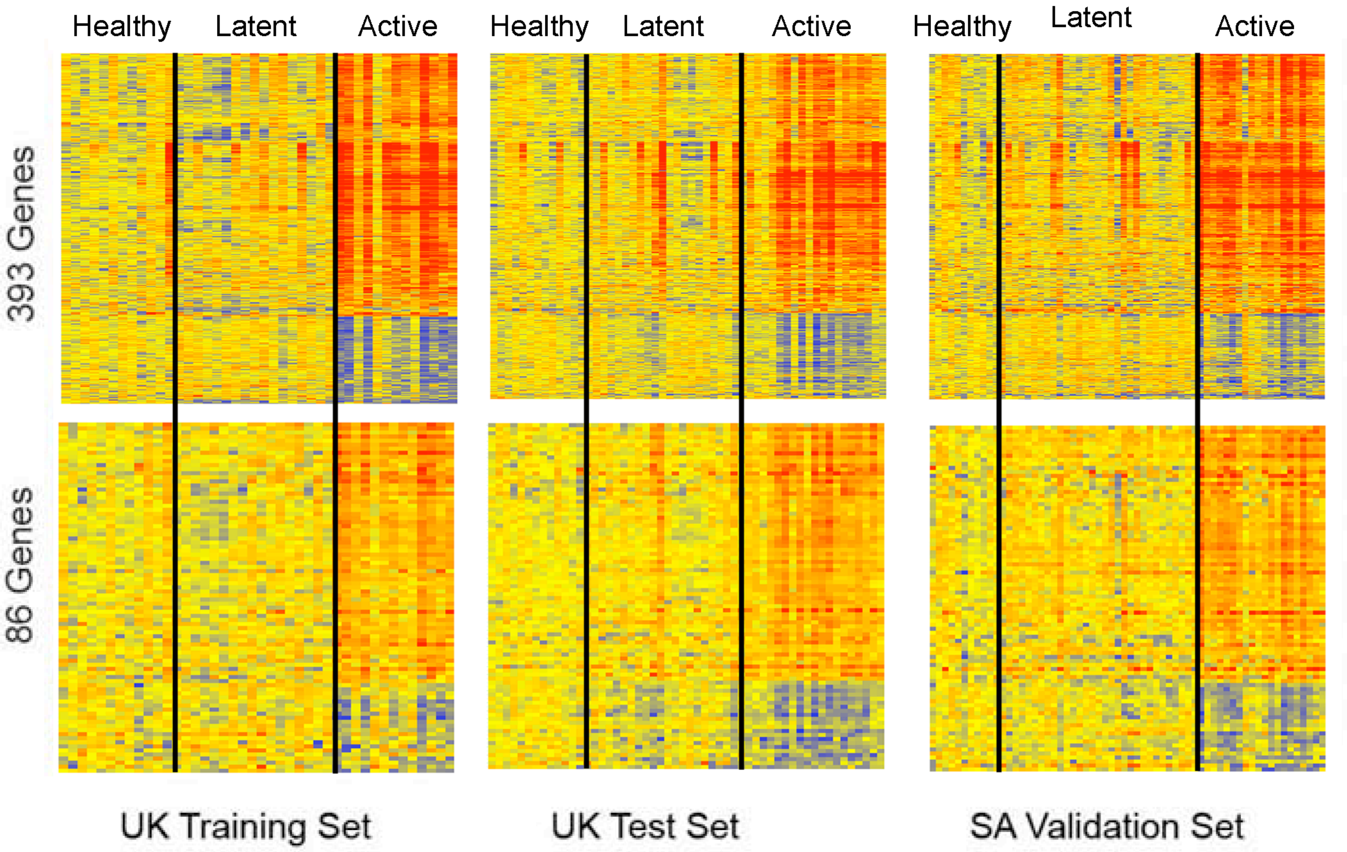
Supplementary Figure 4



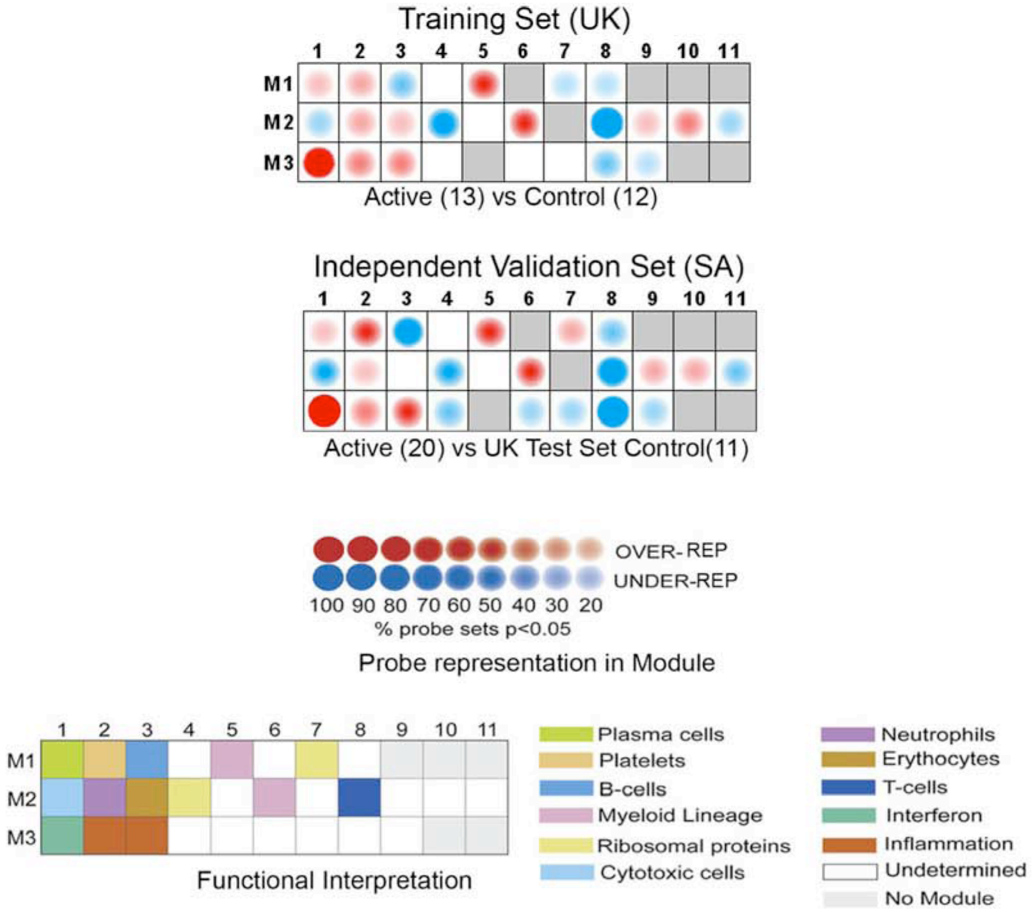
Supplementary figure 5



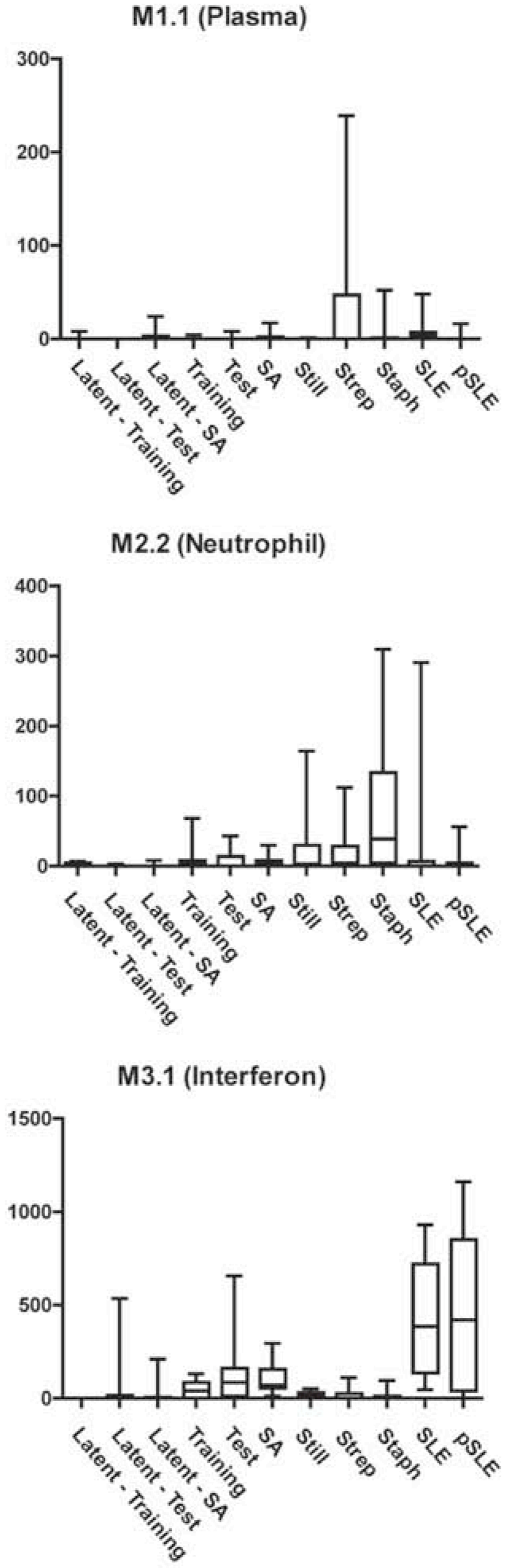
Supplementary figure 6



Supplementary Figure 7



Supplementary figure 8

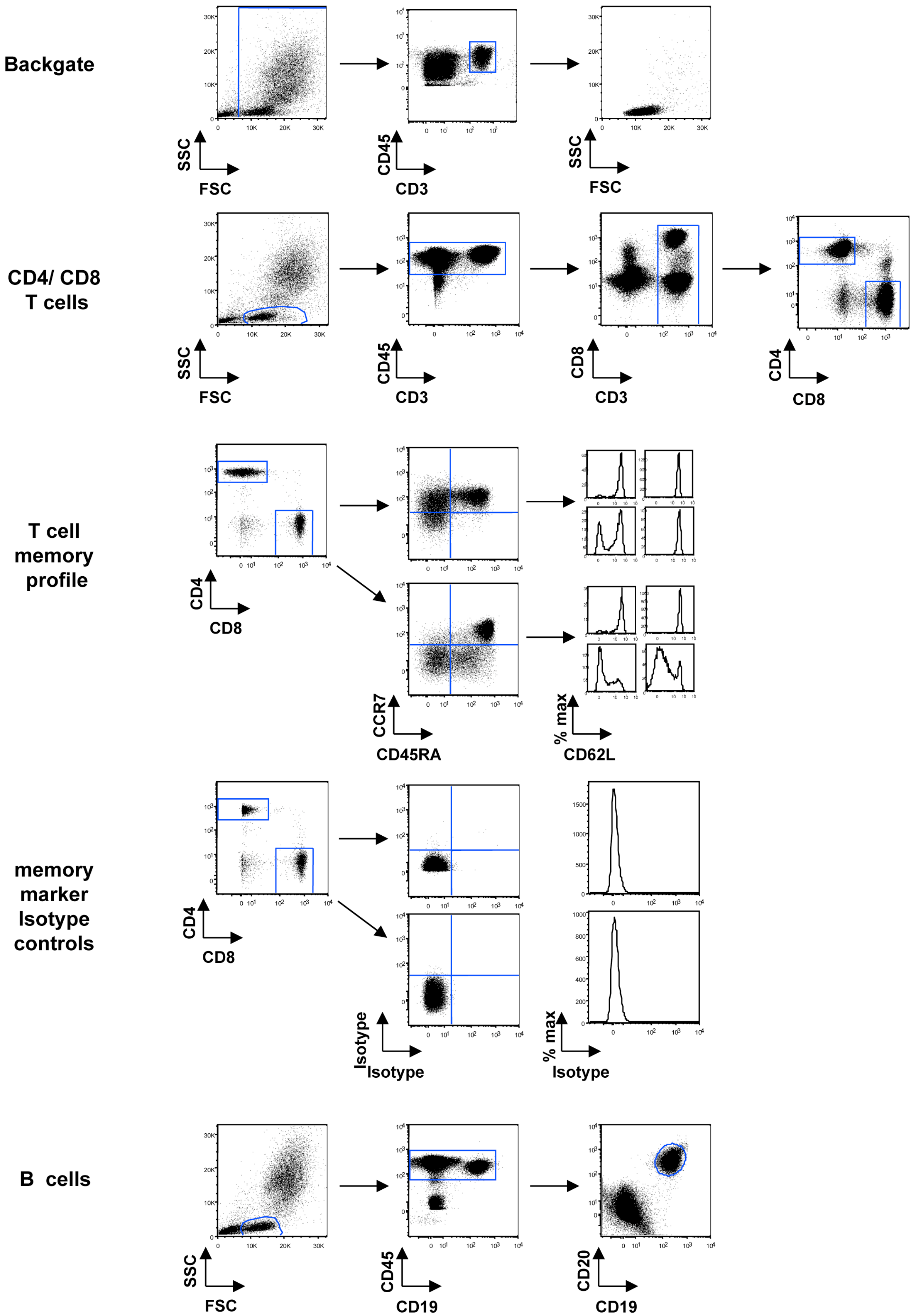


M1.1 (Plasma)

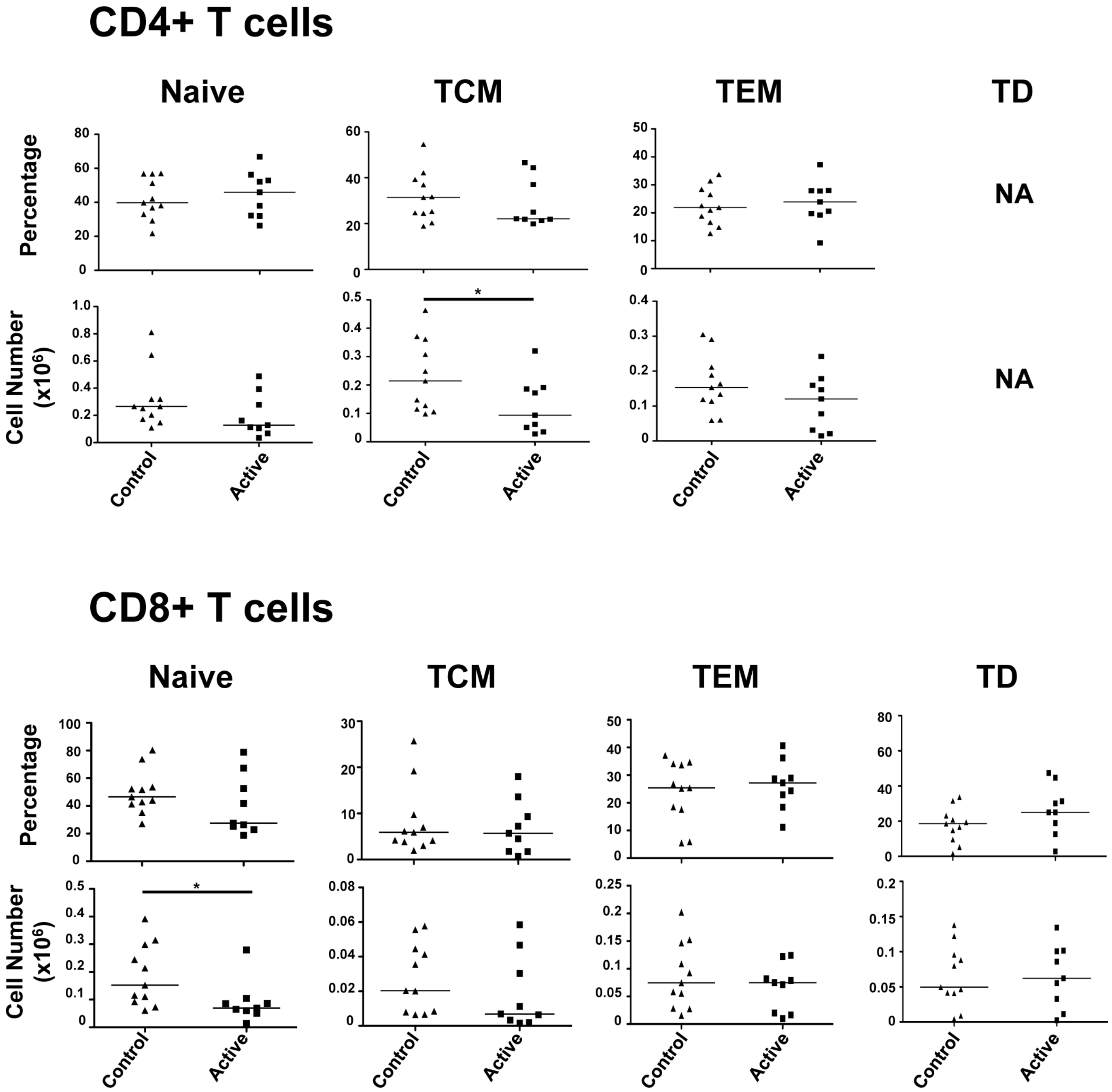
M2.2 (Neutrophil)

M3.1 (Interferon)

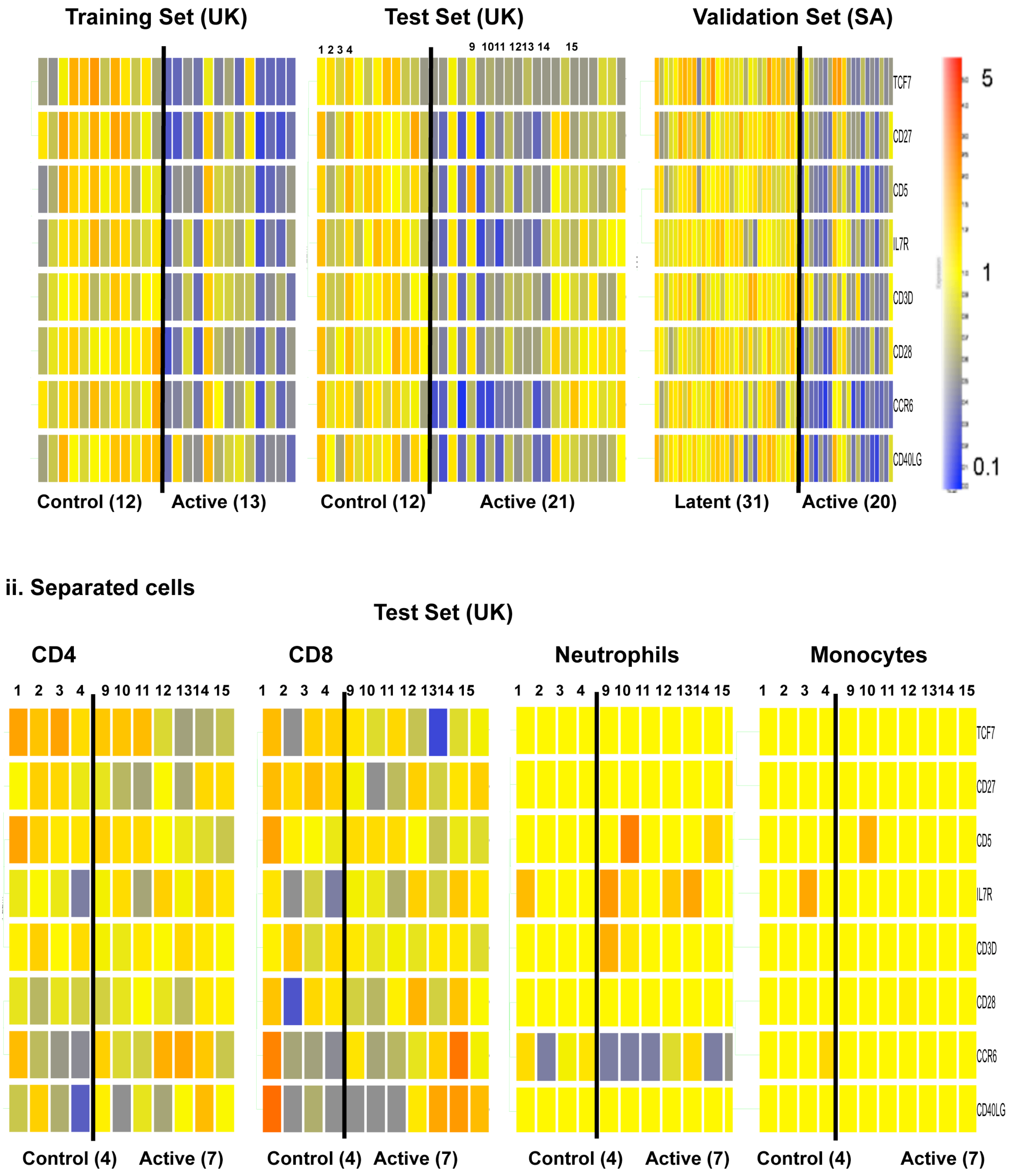
Supplementary Figure 9a



Supplementary Figure 9b



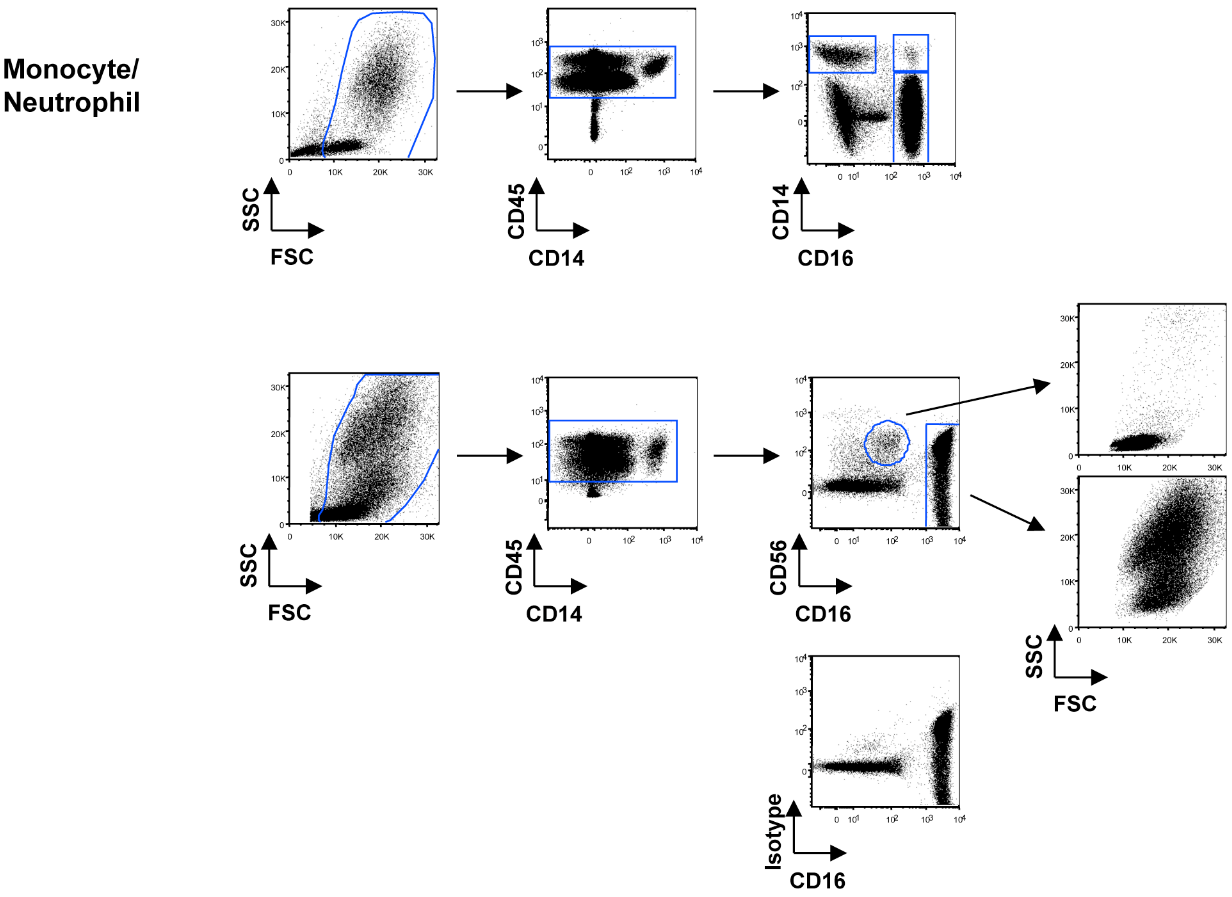
Supplementary Figure 9c T cell genes



i. Whole Blood

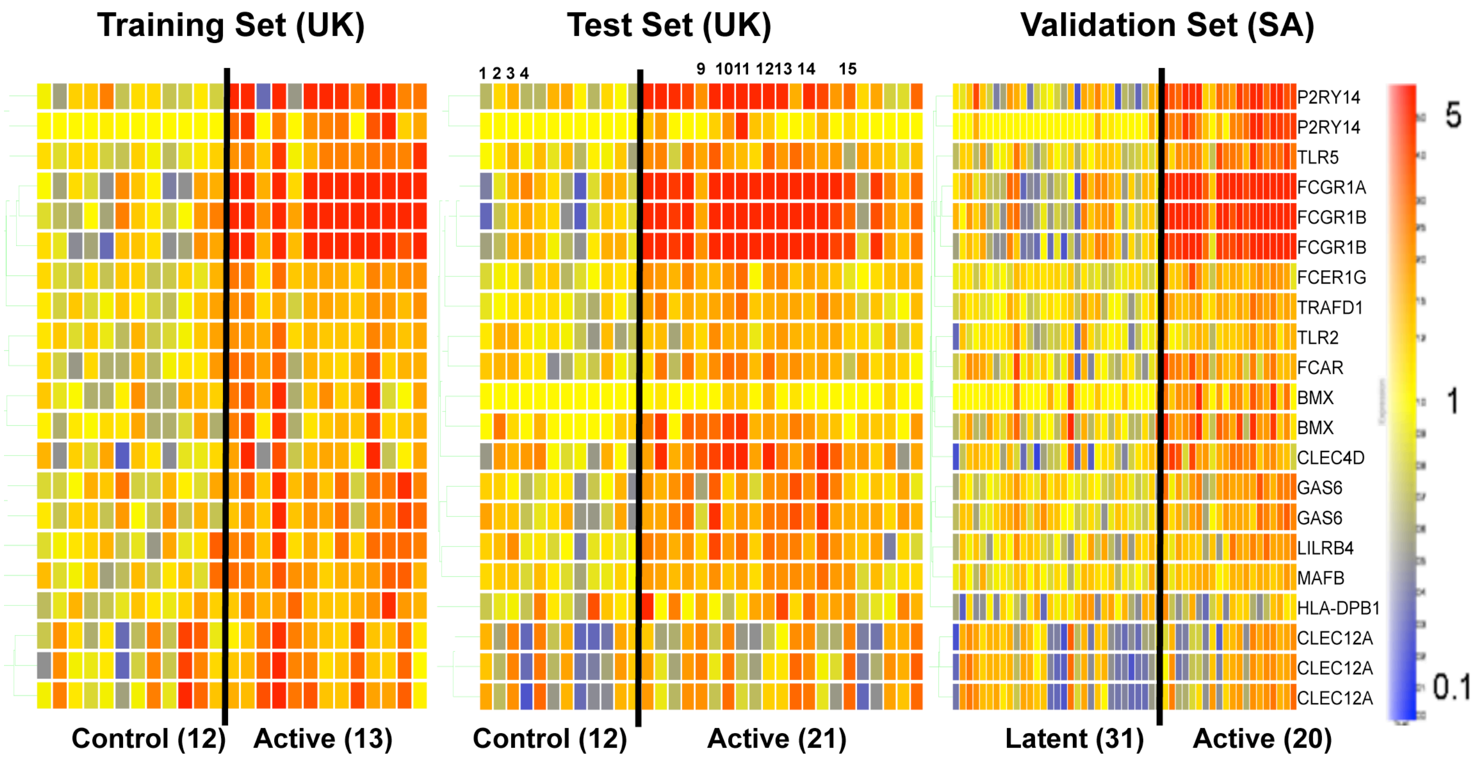
ii. Separated cells

Supplementary Figure 10a

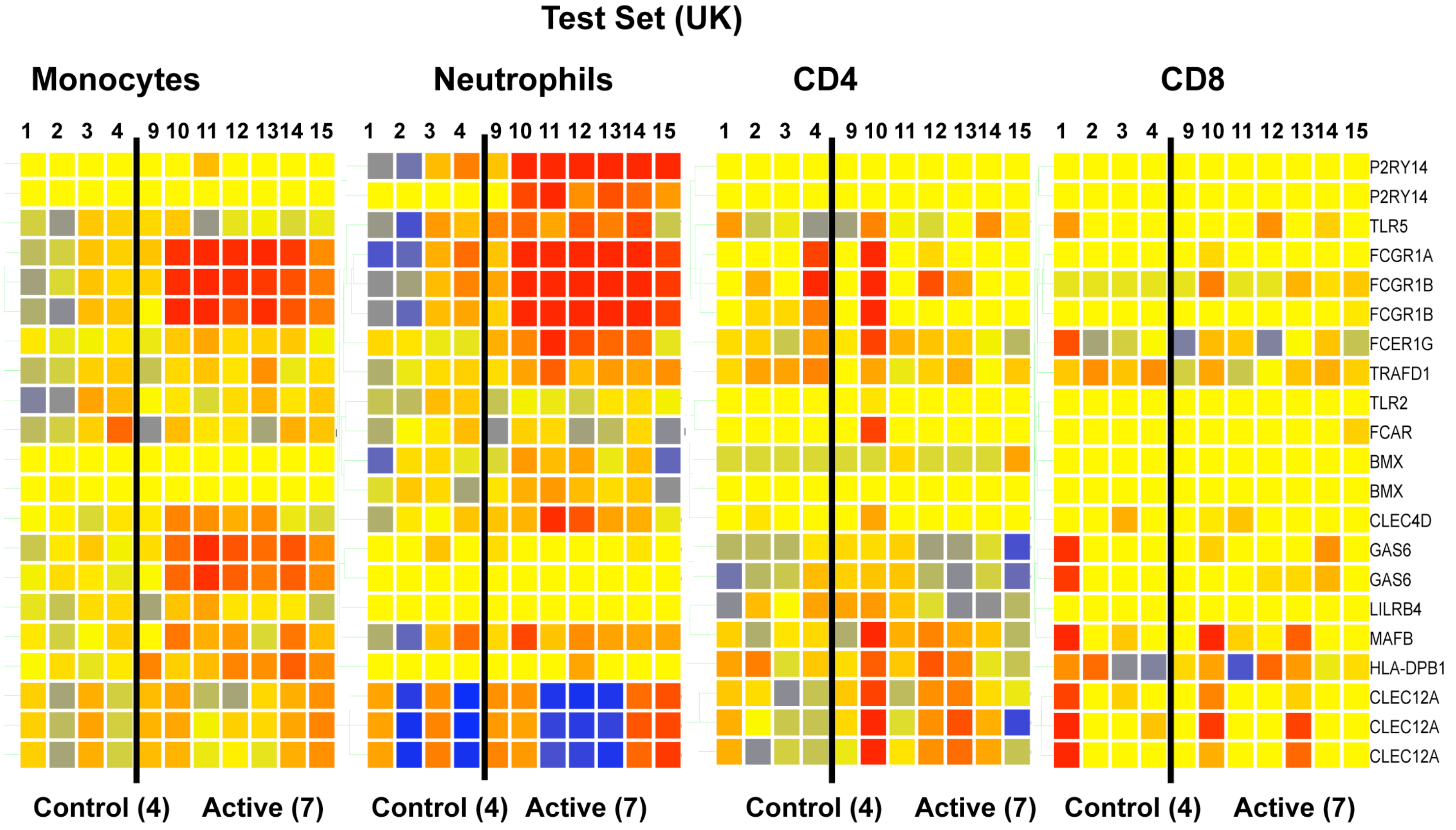


Supplementary Figure 10b Myeloid genes

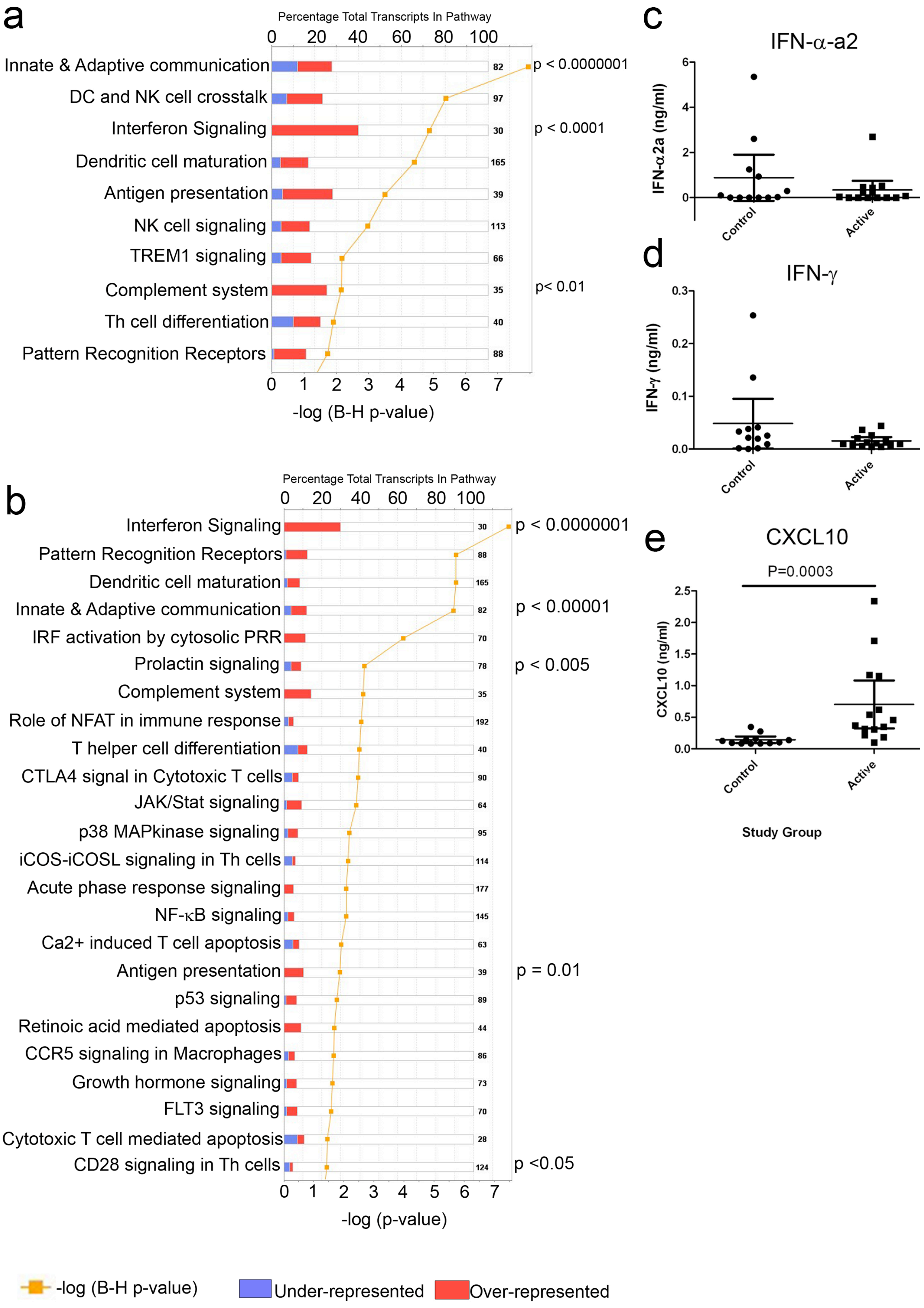
i. Whole Blood



ii. Separated cells



Supplemtary Figure 11



Supplementary Table 1 – Demographic characteristics

a Training Set

		Control	Latent	Active
Number		12	17	13
Age (mean/range, years)		30 (20-35)	33 (19-52)	33 (21-72)
Gender	Male	4 (33%)	8 (47%)	7 (54%)
	Female	8 (67%)	9 (53%)	6 (46%)
Country of Origin	UK Born	5 (42%)	1 (6%)	3 (23%)
	Non-UK Born	7 (58%)	16 (94%)	10 (77%)
Ethnicity	Black	0	7 (41%)	4 (31%)
	South Asian	0	3 (18%)	3 (23%)
	Asian Other	0	5 (29%)	2 (15%)
	White	12 (100%)	2 (12%)	3 (23%)
	Other	0	0	1 (8%)

b Test Set

		Control	Latent	Active
Number		12	21	21
Age (mean/range, years)		31 (21-49)	36 (19-68)	42 (18-78)
Gender	Male	5 (42%)	10 (48%)	13 (62%)
	Female	7 (58%)	11 (52%)	8 (38%)
Country of Origin	UK Born	3 (5%)	1 (5%)	6 (29%)
	Non-UK Born	9 (95%)	20 (95%)	15 (71%)
Ethnicity	Black	3 (25%)	7 (33%)	5 (24%)
	South Asian	3 (25%)	4 (19%)	6 (29%)
	Asian Other	2 (17%)	5 (24%)	1 (5%)
	White	3 (25%)	3 (14%)	8 (38%)
	Other	1 (8%)	2 (10%)	1 (5%)

c Validation set

		Latent	Active
Number		31	20
Age (mean/range, years)		22 (18-33)	34 (21-48)
Gender	Male	11 (35%)	15 (75%)
	Female	20 (65%)	5 (25%)
Country of Origin	UK Born	0	0
	Non-UK Born	31 (100%)	20 (100%)
Ethnicity	Black	31 (100%)	20 (100%)
	South Asian	0	0
	Asian Other	0	0
	White	0	0
	Other	0	0

Supplementary Table 2 – Clinical Characteristics

a Training Set

		Control	Latent	Active
Number		12	17	13
TST (median/range, mm)		0 (0-13)	20 (12-37)	20 (4-25)
BCG vaccinated	Yes	6 (50%)	13 (76%)	9 (69%)
	No	6 (50%)	3 (18%)	3 (23%)
	Not Known	0 (0%)	1 (6%)	1 (8%)
Smear status	Not done	12 (100%)	16 (94%)	0 (0%)
	Smear negative	n/a	1 (6%)	4 (31%)
	Smear positive	n/a	0 (0%)	9 (69%)

b Test Set

		Control	Latent	Active
Number		12	21	21
TST (median/range, mm)		7.5 (0-14)	21 (7-45)	20.5 (0-38)
BCG vaccinated	Yes	12 (100)	8 (47%)	7 (54%)
	No	0 (0%)	9 (53%)	6 (46%)
	Not Known	0 (0%)	1 (6%)	3 (23%)
Smear status	Not done	12 (100%)	20 (90%)	0 (0%)
	Smear negative	n/a	1 (1%)	12 (57%)
	Smear positive	n/a	0 (0%)	9 (43%)

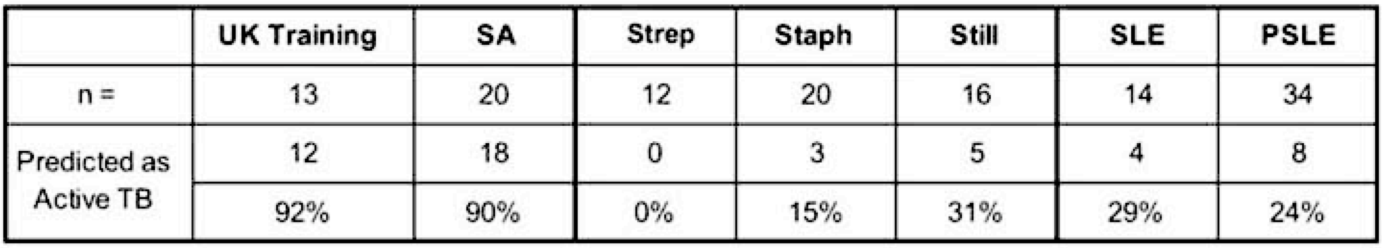
c Validation set

		Latent	Active
Number		31	20
TST (median/range, mm)		14 (0-24)	Not done
BCG vaccinated	Yes	8 (47%)	7 (54%)
	No	9 (53%)	6 (46%)
	Not Known	1 (6%)	3 (23%)
Smear status	Not done	31 (100%)	0 (0%)
	Smear negative	n/a	1 (5%)
	Smear positive	n/a	19 (95%)

Supplementary Table 4: Sensitivity and Specificity of the 393 transcript list for the detection of Active Pulmonary Tuberculosis.

	Sensitivity (95% CI)	Specificity (95% CI)	Indeterminate	p-value
Training Set UK (n = 41)	**91.7 %** (61.5 – 99.8 %)	**96.6 %** (82.2 – 99.9 %)	**0 %**	<0.0001
Test Set UK (n = 54)	**61.7 %** (43.0 – 85.4 %)	**93.8 %** (79.2 – 99.2 %)	**1.9 %**	<0.0001
Validation Set South Africa (n=51)	**94.1 %** (71.3 - 99.9 %)	**96.7 %** (82.8 - 99.9 %)	**7.8 %**	<0.0001

Supplementary Table 7: Specificity of the 86 Gene list for detection of Tuberculosis



	UK Training	SA	Strep	Staph	Still	SLE	PSLE
n =	13	20	12	20	16	14	34
Predicted as Active TB	12	18	0	3	5	4	8
	92%	90%	0%	15%	31%	29%	24%

Supplementary Figure Legends

Supplementary Figure 1. Formation of the Training, Test and Validation Sets. Each cohort was not only independently recruited, but all stages of RNA processing and microarray analysis were also performed completely independently. (**a**) The recruitment of the Training Set cohort in London, UK; (**b**) The recruitment of the independent Test Set cohort in London, UK. (**c**) The recruitment of the independent Validation Set cohort in Cape Town, South Africa.

Supplementary Figure 2. Hierarchical clustering of patient profiles. **a**, RNA was extracted initially from whole blood of the Training Set patient samples and processed as described in Methods. Resulting data were filtered to remove transcripts that were not detected (α=0.01) and had less than two-fold deviation in normalized expression from the median of all samples in greater than 10% of the samples constituting the dataset. This unsupervised filtering, independent of knowledge of sample classification (study group), yielded a list of 1836 transcripts. The 1836 transcript expression profiles for the Training Set were subjected to unsupervised hierarchical clustering by Spearman correlation with average linkage to create a condition tree (along the upper edge of the heatmap), which revealed a distinct signature within the active TB group. This 1836 transcript list was then used to identify signature genes that were significantly differentially expressed among groups (Kruskal-Wallis ANOVA, with the false discovery rate equal to 0.01 using the Benjamini-Hochberg multiple testing correction). This resulted in a 393-transcript signature. **b,** A distinct whole blood 393-gene transcriptional signature of active TB. (**i),** 393-transcripts differentially expressed in whole blood of active and latent TB patients and healthy controls (Training Set) organized by hierarchical clustering (Pearson Correlation with average linkage). **(ii)**, Test Set, ordered by hierarchical clustering (Spearman correlation with average linkage) creating a condition tree, upper horizontal edge of heatmap; study grouping (clinical phenotype) coloured blocks at each profile base. Heatmap rows = genes, columns = participants. Symbols (filled) indicate outliers (Study Group: Red+Symbols = mis-classified as Active TB; Black+Symbols = Latent TB clustering with Active TB); **c,** The 393-transcript expression profiles for the validation set clustered by Spearman correlation with average linkage. **d,** The 393-transcript patient expression profiles for only those aged 22 to 34 years old, since the mean age of the latent TB patients was significantly younger than that of the active TB patients in the Validation Set. This shows that clustering of the 393-transcript profiles is independent of age. Patient clusters can be compared with the clinical and demographic parameters displayed in blocks underneath each profile along the lower edge of the heatmap. A key is provided at the bottom of the figure. Clusters were divided evenly according to distance.

Supplementary Figure 3. A comparison of the transcriptional signature of Active TB with the radiographic extent of disease. **a,** The classification scheme used to grade chest radiographs according to extent of disease. **b,** The 393 transcript expression profiles for all 13 Active TB patients in the Training Set, along with their corresponding chest radiograph taken at the time of diagnosis, with both grouped according to X-ray Grade as per the classification scheme. The expression profile and radiograph of a given patient is given the same numerical indicator. **c,** The 393 transcript expression profiles and chest radiographs for the 21 active TB patients in the Test Set.

Supplementary Figure 4. The transcriptional signature of active TB is diminished in each patient during successful treatment. "Molecular Distance to Health" for each patient was calculated for 7 patients with active TB sampled at 0, 2 and 12 months following the initiation of anti-mycobacterial treatment, and shown for each timepoint.

Supplementary Figure 5. Derivation of 86-transcript TB-specific signature by significance analysis. TB UK Test set and other 4 datasets compared with each of their own controls (TB Test Set); Staphylococcus (Staph), Still's disease (Still), Adult (SLE) and paediatric SLE (pSLE) patients were used to generate a TB-specific signature using significance analysis of each disease against its healthy controls. This resulted in an 86-transcript TB-specific list shown here.

Supplementary Figure 6. Comparisons of the expression of 393-transcript and 86-transcript signatures in the three TB datasets. (UK Training set, UK Test set and SA validation set).

Supplementary Figure 7. Whole blood modular transcriptional signature of active TB in Training and Validation South Africa Sets. Gene expression (active TB versus healthy controls) mapped within a pre-defined modular framework. Spot intensity (red = increased, blue = decreased) indicates transcript abundance. Functional interpretations previously determined by unbiased literature profiling shown by colour-coded grid.

Supplementary Figure 8. Molecular Distance to health for individual modules in TB and different diseases. This is shown for Plasma cell module (M1.1), Neutrophil module (M2.2) and Interferon module (M3.1) in blood from patients with active TB (Training, Test and Validation South Africa, SA); Latent TB (Latent, Training, Test and Validation South Africa, SA); Still's Disease (Still), Group A Streptococcus (Strep), Staphylococcus (Staph) and adult SLE (SLE) and pediatric SLE (pSLE).

Supplementary Figure 9. Analysis of lymphocytes in blood of active TB patients and controls. **a,** Shown are flow cytometric gating strategies used to analyse whole blood from Test Set healthy controls and active TB patients for T cells and B cells. The top row of panels shows the backgating strategy used to determine the lymphocyte FSC/SSC gate used in subsequent gating. A large FSC/SSC gate was set initially (left panel) and then analysed for CD45 vs CD3. CD45CD3 cells were gated (middle panel) and their FSC/SSC profile determined (right panel). This profile was then used to determine an appropriate lymphocyte FSC/SSC gate (see second row, left hand panel). This backgating procedure was also carried out gating on $CD45^+CD19^+$ (B cells) to ensure these cells were included in the lymphocyte gate (not shown). The second row of panels shows the gating strategy used to identify T cell populations. A lymphocyte FSC/SSC gate was set and these cells assessed for CD45 vs CD3 (2nd panel from left). $CD45^+$ cells were then gated and assessed for CD3 vs CD8. $CD3^+$ T cells were gated and assessed for CD4 and CD8 expression. $CD4^+$ and $CD8^+$ subsets were then gated. Rows 3-6 show the gating strategy used to define T cell memory subsets. CD4 and CD8 T cells gated as in row 2 were assessed for CD45RA vs CCR7 expression and a quadrant set based on isotype controls (rows 5 & 6) to define naïve ($CD45RA^{+}CCR7^{+}$), central memory (CD45RA-$CCR7^{+}$), effector memory ($CD45RA^{-}CCR7^{-}$) and in the case of $CD8^{+}$ T cells, terminally differentiated effector ($CD45RA^{+}CCR7^{-}$) T cells. These subsets were also assessed for CD62L expression. The bottom row of panels shows the strategy used to gate B cells. A lymphocyte FSC/SSC gate was set and cells assessed for CD45 vs CD19. $CD45^{+}$ cells were gated and assessed for CD19 and CD20. B cells were defined as $CD19^{+}CD20^{+}$. **b**, Whole blood from 11 test set healthy controls (Control) and 9 test set active TB patients (Active) was analysed by multi-parameter flow cytometry for T cell memory populations. Full flow cytometry gating strategy is shown in Supplementary Figure 5a. Graphs show pooled data of all individuals for percentages of naïve, central memory (TCM), effector memory (TEM) and terminally differentiated effector (TD, $CD8^{+}$ T cells only) cell subsets (top row, each group) and cell numbers ($x10^{6}$/ml) for each cell subset (bottom row, each group). Each symbol represents an individual patient. Horizontal line represents the median. **c**, (**i**) T cell transcript abundance in whole blood samples from active TB (Training, Test and Validation Sets); and (**ii**) expression in separated blood leucocyte populations from Test Set whole blood. Gene abundance/expression is shown as compared to the median of the healthy controls (labelled as in Figure 1). The expression profiles of a given patient are given the same numerical indicator. Error bars = median, * = $p<0.05$, Mann-Whitney test.

Supplementary Figure 10. Analysis of myeloid cells in blood of active TB patients and controls. (a) Shown are flow cytometric gating strategies used to analyse whole blood from test set healthy controls and active TB patients for monocytes and neutrophils. A large FSC/SSC gate was set (top row, left panel) and was then analysed for CD45 vs CD14. $CD45^{+}$ cells were gated (middle panel) and assessed for CD14 vs CD16. Monocytes were defined as $CD14^{+}$, inflammatory monocytes as $CD14^{+}CD16^{+}$ and neutrophils as $CD16^{+}$. Also shown in this figure is the gating strategy used to assess possible overlap between $CD16^{+}$ neutrophils and CD16 expressing NK cells. A large FSC/SSC gate was set to encompass both neutrophils and NK cells. $CD45^{+}$ cells were then assessed for CD16 vs CD56 (NK cell marker). $CD16^{+}$ neutrophils expressed high levels of CD16 and not CD56 (as shown by isotype control plot, bottom panel). $CD56^{+}$ NK cells expressed intermediate levels of CD16 and did not overlap with CD16hi cells. $CD56^{+}$CD16int cells and CD16hi cells had different FSC/SSC properties. (**b**) Myeloid gene (**i**) transcript abundance in whole blood samples from active TB (Training, Test and Validation Sets); and (**ii**) expression in separated blood leucocyte populations from Test Set blood. Gene abundance/expression is shown as compared to the median of the healthy controls (labelled as in Figure 1). Numbers shown in the Test Set and the separated populations correspond to individual patients.

Supplementary Figure 11. IFN inducible gene expression is dominant in the TB transcriptional signature. Ingenuity Pathways analysis of **a,** 1836-transcript and **b,** the 393-transcript signature. The probability (as a -log of the p-value calculated by Fischer's Exact test, with Benjamini-Hochberg multiple testing correction) that each canonical biological pathway is significantly over-represented is indicated by the orange squares.

The solid coloured bars represent the percentage of the total number of genes comprising that pathway (given in bold at the right hand edge of each bar) present in the analysed gene list. The colour of the bar indicates the abundance of those transcripts in the whole blood of patients with active TB compared with healthy controls in the training set. Serum levels of: **c,** interferon-alpha 2a (IFN-α 2a), and **d**, IFN-γ) and **e,** CXCL10 (IP10) are shown here for the 12 healthy controls and 13 patients with active TB used for the training set microarray analyses. No significant difference was observed between groups for either IFN-α 2a and IFN-γ using two-tailed Mann-Whitney test, but was observed for CXCL10. The horizontal line indicates the mean for each group and the whiskers indicate the 95% confidence interval.