Overlapping patterns and unique differences: a study into immunological variation within and between populations

Dorst, M.M.A.R. van

Chapter 5

Lifestyle score is associated with cellular immune profiles in healthy Tanzanian adults

Jeremia J. Pyuza¶, Marloes M.A.R. van Dorst¶, Koen Stam, Linda Wammes, Marion König, Vesla I. Kullaya, Yvonne Kruize, Wesley Huisman, Nikuntufya Andongolile, Anastazia Ngowi, Elichilia R. Shao, Alex Mremi, Pancras C.W. Hogendoorn, Sia E. Msuya, Simon P. Jochems, Wouter A.A. de Steenhuijsen Piters, Maria Yazdanbakhsh
¶ Contributed equally

Adapted from:
Brain, Behavior, & Immunity- Health. doi: 10.1016/j.bbih.2024.100863

ABSTRACT

Immune system and vaccine responses vary across geographical locations worldwide, not only between high and low-middle income countries (LMICs), but also between rural and urban populations within the same country. Lifestyle factors such as housing conditions, exposure to microorganisms and parasites and diet are associated with rural-and urban-living. However, the relationships between these lifestyle factors and immune profiles have not been mapped in detail. Here, we profiled the immune system of 100 healthy Tanzanians living across four rural/urban areas using mass cytometry. We developed a lifestyle score based on an individual's household assets, housing condition and recent dietary history and studied the association with cellular immune profiles. Seventeen out of 80 immune cell clusters were associated with living location or lifestyle score, with eight identifiable only using lifestyle score. Individuals with low lifestyle score, most of whom live in rural settings, showed higher frequencies of NK cells, plasmablasts, atypical memory B cells, T helper 2 cells, regulatory T cells and activated CD4+ T effector memory cells expressing CD38, HLA-DR and CTLA-4. In contrast, those with high lifestyle score, most of whom live in urban areas, showed a less activated state of the immune system illustrated by higher frequencies of naïve CD8+ T cells. Using an elastic net machine learning model, we identified cellular immune signatures most associated with lifestyle score. Assuming a link between these immune profiles and vaccine responses, these signatures may inform us on the cellular mechanisms underlying poor responses to vaccines, but also reduced autoimmunity and allergies in low- and middle-income countries.

INTRODUCTION

Variation in the immune system have been observed across populations in low and middle-income countries (LMIC) in Africa and Asia and those living in high-income countries (HIC) in Europe and the USA [1-6]. In addition, immune system variation has been observed within countries, such as in rural compared to urban areas in Senegal [2], Tanzania [7] and Indonesia [1]. The immune system of rural-living individuals in LMICs shows higher memory, activated and regulatory immune profiles, characterized by among others regulatory T cells and T helper 2 cells (Th2 cells), compared to urban-living individuals [1, 2, 8, 9]. At the same time, reduced vaccine performance has been observed in populations living in LMICs, in particular in rural areas [4, 10, 11]. Moreover, it is known that in these same populations, there are less diseases of affluence, such as allergies or auto-immunities, where unchecked inflammation is a strong contributor [4, 11-19].

Several factors determine the immune profile of an individual, including genetic and demographic factors, such as age and sex, as well as environmental factors, including exposure to microorganisms and parasites, type of housing and dietary history [20, 21]. While genetics plays an important role in immune system variation during early childhood, this influence wanes with age due to cumulative exposure to environmental factors, including pathogens [20, 22, 23]. This has been illustrated in individuals chronically infected with helminths, who exhibit skewed baseline immune profiles, characterized by higher frequencies of Th2, regulatory T cells and higher expression of activation and inhibitory markers such as cytotoxic T lymphocyte-associated protein 4 (CTLA-4), HLA-DR and programmed cell death protein 1 (PD-1) on T cells [24-26]. Furthermore, individuals infected with cytomegalovirus (CMV) show a disproportionately higher activation state of the immune system and an increased frequency of memory cells [27, 28].

Socioeconomic status (SES) is intertwined with housing quality, nutritional status and access to healthcare [29, 30]. These factors contribute to infection risk and, therefore, propel the vicious circle of infection/infestation, which strongly impacts the immune system [18, 29-33]. The type of diet can also be linked to variation in immune profile, as was demonstrated in a recent study in Tanzania [7]. In this study, rural-living Tanzanians harbored a more anti-inflammatory immune profile that correlated with higher levels of plant-derived flavonoid apigenin found in food mostly eaten in rural settings [7]. Therefore, taken together, there is evidence for links between living environments such as housing, exposure to microorganisms and parasites,

SES including individual assets and diet and immune system variation in LMICs.

Although the immune profiles of urban- and rural-living individuals have been directly compared, a more granular assessment of lifestyles irrespective of living location is lacking, as individuals living in rural areas may exhibit an urban lifestyle and vice versa. We hypothesized that a more refined measurement of lifestyle including housing status, assets (e.g. car, bicycle motorcycle or radio), and dietary history (i.e. frequency of consumption of common dietary products) will allow us to better explain immune variation previously related to rural or urban living location. Especially, we aim to more precisely define immune signatures in individuals exhibiting immune hypo-responsiveness. Such information can have an impact on both communicable and non-communicable diseases, as a poor immune response to vaccines will affect susceptibility to vaccine-preventable infections, while poor responses to (self-)antigens can lead to fewer allergies or autoimmune diseases in rural-living individuals.

Therefore, we not only used mass cytometry to obtain a highly granular immune profile but also surveyed lifestyle variation among Tanzanian adults recruited from two rural and two urban locations to maximize lifestyle variation using a detailed questionnaire of housing conditions, assets and recent dietary history. We present a lifestyle score based on these questionnaire data, which places individuals on the spectrum ranging from rural to urban lifestyle. We used this lifestyle score to explain immune profile variation in Tanzanian adults living in rural and urban areas and contrasted this with immune signatures from urban-living Europeans. In addition, we utilized a machine learning model to define combined immune signatures most strongly associated with the lifestyle score.

MATERIALS AND METHODS

Study design

This observational study was conducted between September and October 2022 as part of the CapTan study. A total of 203 healthy Tanzanian participants aged between 18 to 35 years were included from two urban locations (Urban Arusha and Urban Moshi) and two rural locations (Rural Moshi and Mwanga) in northern Tanzania (**Figure 1A**).

The study was approved both at a local level by the Ethical Board of the Kilimanjaro Christian Medical University College (No. 2588) and at the national level by the Tanzania National Ethical Committee Board (NIMR/

HQ/R.8a/Vol.IX/4089). In addition, samples collected from ten Dutch 18 to 30-year-old adults enrolled between January 2022 and September 2022 were included in the TINO study (ClinicalTrials.gov, reference no. NCT06039527). The study was approved by the Ethics Committee of Leiden University Medical Center (NL77841.058.21).

Description of study areas

Arusha City (1400m above sea level; 617,631 inhabitants [34]) is the administrative, business, commercial and educational centre of the Arusha region, as it accommodates most diplomatic and international activities. Due to these important regional functions, there is high diversity in ethnicity, economic status and lifestyle. Maasai, Meru and Chagga are the most common ethnicities. Most people living in Arusha City have access to good sanitation with the availability of clean, treated water. However, some people are slum dwellers, i.e. living in the city but practicing a rural lifestyle. Most people are self-employed or office employees in the government and private sectors [34].

Kilimanjaro region has about 1.9 million inhabitants [34] across seven different districts, three of which are included in this study (Moshi City, Rural Moshi and Mwanga). Moshi City (referred to as Urban Moshi) (700-950m above sea level; 331,733 inhabitants [34]) is the administrative, commercial and educational center of the Kilimanjaro region. Most people live a Western lifestyle and have good general sanitation and access to clean water. The main ethnicities are Chagga and Pare. Formal business is the main activity, followed by government and public employment, while few people are involved in agricultural and entrepreneurial activities [34].

People in Rural Moshi (535,803 inhabitants [34]) are mainly involved in agricultural activities. Some people have access to clean water, while few use borehole water sources. People live in large family units and their main economic activities are subsistence farming and animal husbandry. The main ethnicity is Chagga and people follow Chagga traditions, such as drinking local brew from banana/plantain.

The population of Mwanga district (684m above sea level; 148,763 inhabitants [34]) is mainly active in irrigation, subsistence farming and animal husbandry. The primary water sources are boreholes, rivers and dams, with only few people having access to tap water. Like Rural Moshi, people live in large family units. The main ethnicity is Pare, with few Chagga.

Europeans were recruited in the area around Leiden, an urban centre in The Netherlands. European individuals were Dutch.

Participant screening and enrollment
In rural communities, study information was given through community leaders and announcements during mass gatherings in mosques, churches and during village meetings. In urban communities, study information was distributed using leaflets and through community leaders, office announcements and university gatherings. Eligible participants (age 18-35 years and permanent residency of a given location) were asked to enroll in the study. Following informed consent, 230 participants were voluntarily screened for in- and exclusion criteria. Exclusion criteria were pregnancy, lactation, having acute or chronic diseases, being HIV-positive, recent use of antibiotics, use of antimalarials and use of tuberculostatic drugs. Participants were screened for HIV infection (SDBIOLINE HIV-1/2 3.0kit, LOT:03ADG020A), malaria (Malaria Ag p.f/Pan, Ref: 05FK60, LOT:05EDG018A) and soil-transmitted helminth such as hookworms (Ancylostoma duodenale and Necator americanus), Trichuris trichiura, Ascaris lumbricoides, Strongyloides stercoralis and Schistosoma mansoni using Kato-Katz or Schistosoma haematobium (POC-CCA, butch no:220701075). Furthermore, hemoglobin levels were measured (HemoCue Hb 301(CE:1450820055) and random blood glucose was assessed (ACCU-CHECK glucose test strips, Roche Diabetic care,06993761001). Weight and height were measured using a well-calibrated machine (RGZ-160, made from China), and last, blood pressure was measured using OMRON(SN:202111007949V). After nurse counseling, HIV-positive individuals who had low or high blood pressure (≤90/60mmHg and ≥140/90mmHg, respectively) or had high blood glucose (≥7.1mmol/L fasting or ≥11.1mmol/L random glucose) were excluded and guided for further actions. People diagnosed with schistosomiasis or soil-transmitted helminth infections were treated with praziquantel and albendazole, respectively according to Tanzanian treatment guidelines. Based on exclusion criteria, 27 of 230 participants were excluded.

All questionnaires and clinical samples were collected by a trained study team, consisting of medical doctors, nurses and laboratory scientists. Data from Tanzanian individuals were collected using the cloud-based electronic data collection system REDCap, with a server hosted at the Kilimanjaro Clinical Research Institute in Tanzania. Data from Dutch participants were collected in a Castor database, with a server hosted in The Netherlands.

Lifestyle questionnaire
Questionnaires adopted from the Tanzania Demographic and Health Survey and Malaria Indicator Survey (TDHS-MIS) and previously published work conducted in Tanzania, focused on diet in relation to metabolic profiles and inflammatory status [7, 54] were used to collect data on basic demographics, wealth (house construction, general hygiene, land/animal/livestock/non-productive asset ownership) and (recent) food history. Combined, the collected information on wealth and food history was considered reflective of one's 'lifestyle'. Among others, our questionnaire included questions on the material used to construct the house's floor, roof and walls, the source of water, the type of toilet and available cooking facilities. We assessed the number of milk cows, cattle, goats, sheep, horses and poultry owned and inquiries were made on land ownership and possession of non-productive assets, such as radios, televisions, computers, refrigerators and ironing tools (whether powered by charcoal or electricity), watches, motorcycles, trucks, animal-drawn carts, generators and motorboats. As diet was recently found to shape immune responses in a Tanzanian population [7], we additionally collected data on recent food history. We specifically focused on the frequency of various food types participants consume per week, including ugali (stiff porridge), plantain, rice, potatoes, meat, fish, beans/peas, green vegetables, cabbage, fruits and local beer.

PBMC isolation and cryopreservation
Blood was collected in sodium heparin tubes from 189 of 203 participants. PBMC isolation and cryopreservation were performed as previously described [1]. 27 Samples were excluded due to low blood quality, technical problems during PBMC isolation or low cell counts. The remaining 162 cryopreserved PBMC samples were transported from Moshi, Tanzania, to Leiden, The Netherlands, using a liquid nitrogen dry vapor shipper. Out of these samples, we selected 100 individuals (25 per location) for immune phenotyping based on age, sex and educational level. Apart from these variables, baseline demographics for the total cohort and the mass cytometry cohort were comparable (**Table 1 and Table S1**).

Mass cytometry antibody staining
Antibody panels were designed to phenotype immune cells ex vivo. Details on antibodies used are listed in **Table S4**. Antibodies were conjugated to metal using 100µg of purified antibody combined with either the Maxpar X8 or MCP9 Antibody Labelling Kit (Fluidigm), as per the manufacturer's instructions. Conjugated antibodies were then stored in 200µl of Antibody Stabilizer PBS (CANDOR Bioscience GmbH) at 4°C. Titration of all antibodies was conducted on PBMC samples.

On the day of staining, cryopreserved PBMCs were thawed with 20% FCS/2mM Mg2+/1:10,000 benzonase/RPMI medium at 37°C and washed twice with 10% FCS/RPMI medium. For phenotyping, 3 × 106 cells per sample were prepared according to the Maxpar Nuclear Antigen Staining Protocol V2 (Fluidigm). PBMCs were washed with Maxpar staining buffer and centrifuged at 400g for 5 minutes in 5-ml Eppendorf tubes. Study samples were randomized over seven batches and for each batch up to 17 samples were barcoded. To barcode the samples, the cells were resuspended in 50µl of Maxpar staining buffer and 50µl of a barcode mix targeting β2-microglobulin (B2M) was added to each sample, employing a 6-choose-3 scheme using 106cadmium (Cd), 110Cd, 111Cd, 112Cd, 114Cd and 116Cd. After a 30-minute room temperature incubation and a wash with Maxpar Staining Buffer, the cells were centrifuged, the supernatant was removed and the cells were resuspended in Maxpar staining buffer and pooled into one tube for each batch.

Subsequently, cells were treated with 5ml (about 0.17 oz) of 500× diluted Cell-ID Intercalator-103Rh (Fluidigm) for 15 minutes to identify dead cells. After washing with staining buffer, cells were incubated with 20µl Human TruStain FcX Fc receptor blocking solution (BioLegend) and 130µl of staining buffer at room temperature for 5 minutes. Next, 150µl of a freshly prepared surface antibody cocktail was added for another 30-minute room-temperature incubation. After a double wash with staining buffer, cells were fixed with 1.6% PFA in 5ml PBS for 10 minutes. Post-centrifugation, cells underwent fixation and permeabilization using the eBioscience Foxp3/Transcription Factor Staining Buffer Set from eBioscience, followed by incubation with Human TruStain FcX receptor blocker. An intranuclear antibody cocktail was then added and the cells were incubated for an additional 30 minutes. After washing with permeabilization buffer and staining buffer, cells were fixed with 1.6% PFA in 5ml PBS for 10 minutes. Finally, cells are stained with 1000× diluted Cell-ID Intercalator-Ir (Fluidigm) in Maxpar Fix and Perm Buffer at room temperature for 1h and stored in RPMI 20% FCS 10% DMSO at -80°C until acquisition.

Mass cytometry data acquisition
All barcoded samples within one batch were acquired simultaneously. Cells were measured using a Helios mass cytometer (Fluidigm) and calibrated as per Fluidigm's guidelines. Before measurement, cells underwent counting, washing with Milli-Q water, straining and then were suspended at a concentration of 1.0×10^6 cells/ml in a solution containing 10% EQ Four Element Calibration Beads from Fluidigm and Milli-Q water. Data acquisition in mass cytometry was performed using dual-count mode and with noise

reduction. Various channels were used, including those for antibody detection, intercalators (103Rh, 191Ir, 193Ir), calibration beads (140Ce, 151Eu, 153Eu, 165Ho, 175Lu) and for tracking background/contamination (133Cs, 138Ba, 206Pb). Post-acquisition, the mass bead signal was used to standardize short-term signal variations, using the EQ passport P13H2302 as a reference throughout each experiment. When necessary, normalized FCS files were merged using Helios software, while retaining the beads.

Data analysis
All data preprocessing and statistics were performed in R v4.2.2 and RStudio Server v2022.03.999. All p-values were corrected for multiple testing using the Benjamini-Hochberg procedure (and referred to as q-values). P-/q-values<0.05 were considered statistically significant.

Data preprocessing
First, cells were automatically gated based on Gaussian parameters (CyTOFClean R-package; v1.03beta; https://github.com/JimboMahoney/cytofclean). Next, automatic gating was applied to select for intact/DNA+- (191Ir and 193Ir channels), CD45+- (89Y) and live cells (live/dead staining) (openCyto v2.10.1 R-package). All automatically set gates were manually inspected. Samples were compensated and debarcoded (CATALYST v1.22.0 R-package). Data were transformed using a hyperbolic arcsinh-transformation with a cofactor of 5 for downstream processing. Next, reference samples collected from healthy European adults included in each individual batch were used to train a CytoNorm-model (CytoNorm v0.0.17 R-package; CytoNorm.train-function; nQ = 101; goal = 'mean'; k = 10; limit = 0-8). The trained model was applied to all samples, adjusting for batch effects (CytoNorm.normalize-function).

Cell clustering
Cells were subjected to flowSOM-clustering (15 × 15 hexagonal grid; rlen=100; kohonen v3.0.11 R-package), followed by metaclustering at k = 80 clusters using the hierarchical clustering (factoextra v1.0.7 R-package, hcut-function, distance = 'ward.D2'). The clustering map was trained on 100k cells per sample, the remaining cells were mapped to the trained map (predict.kohonen-function). Cell clusters were annotated at subset-level by an expert immunologist. Cell labels were further refined by incorporating markers that exhibit variability within a given subset in the cell label.

Lifestyle score
Multiple correspondence analysis (MCA) was applied to categorical questionnaire data (38 manually curated lifestyle-related questions; 21 on

assets, 11 on food and 6 on housing) for all 203 Tanzanian participants (FactoMineR v2.7 R-package, MCA-function). Missing values are imputed using mode imputation. Principle component (PC) 1 was defined as 'lifestyle score', as this component, per definition, explained most variance across lifestyle questionnaire data. Coordinates of samples and variable categories were visualized in biplots. In addition, (cumulative) variable category contributions for lifestyle score were extracted and shown.

Statistical analyses

To understand the overall structure of the data, cells were placed on a two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE) map using the Fit-SNE algorithm v1.2.1 (https://github.com/KlugerLab/Fit-SNE/blob/master/fast_tsne.R). Fit-SNE was performed on a down-sampled dataset including 1,500 cells per sample (max_iter = 1,000; learning rate = n cells/12; perplexity = n cells/100).

To compare the frequency of cell clusters across rural and urban Tanzanian locations, we employed a generalized linear mixed model (binomial = 'family'; link = 'logit'; lme4 R-package v1.1-31). The number of cells in each cell cluster (as a fraction of total CD45+ cells per sample) was considered the dependent variable. We fit two models to assess the overall effect of location. Model 1 included (scaled) age and sex as fixed explanatory variables and 'sample ID' as a random intercept. 'Sample ID' was included as a random effect to deal with any under- or overdispersion due to the binomial model. Model 2 was the same as model 1, except that 'location' was added as a fixed explanatory variable. ANOVA tests were used to assess whether location (model 2) significantly improved model fit compared to model 1. Significant models (after correction for multiple testing using Benjamini-Hochberg) were subjected to pairwise comparisons between locations using the emmeans v1.8.5 R-package (Tukey post hoc test). The associations between cell cluster frequency and lifestyle score were also assessed using GLMMs, including lifestyle score, (scaled) age and sex as fixed explanatory variables and 'sample ID' as a random intercept. For sensitivity analyses, we fitted an additional 'combined' GLMM, including both location and lifestyle (LS) (as well as age (scaled) and sex) as fixed effects and sample ID as random effect. Model fit (using Akaike Information Criterion [AIC]) of the 'combined' GLMM was compared to same model, after removing either location or lifestyle score, to assess the relative importance of these variables to performance cluster-specific models.

Elastic net machine learning modelling

To identify a combined immune 'endotype' most associated with variation in lifestyle score, we fit an elastic net machine learning model (tidymodels v1.1.1 R-package, glmnet-engine). Scaled age, sex and cell frequencies of all 80 clusters were included as predictors and lifestyle score was included as an outcome variable. Data was randomly split into train (80%) and test (20%) data (stratified for living location). Model tuning was performed on training data using 2,000 bootstrapped data samples, optimizing penalty and mixture parameters. The best model was identified based on the highest explained variance (R2) between observed and predicted lifestyle score (penalty = 0.788, mixture = 0.1). The final model was applied to both training and testing data to generate final estimates of model fit (R2). Variable importance was assessed using the vip v0.4.1 R-package. Feature stability was assessed by extracting all features from the models fitted with the optimized tuning parameters across bootstrap datasets (n = 2,000). The number of times a feature was selected was used as a measure for feature stability.

RESULTS

Characteristics of the study population

The Tanzanian study population consisted of 203 adults recruited from four geographical locations in northern Tanzania, including two urban locations, Arusha and Moshi Urban and two rural locations, Moshi Rural and Mwanga (**Figure 1A**). These four locations were categorized as rural and urban based on the National Bureau of Statistics and the 2022 Census [34]. Detailed information on housing, assets and food history was collected using questionnaires [7, 35] (**Figure 1B**).

From these 203 individuals (**Table S1**), PBMC samples of 100 individuals were included for mass cytometry analyses (n = 100; n = 25 from each site in four sites) (**Table 1**). The median age was 25.0 years (interquartile range [IQR], 23-29 years). The prevalence of parasitic infections was 7% and these infections were detected only in individuals from rural areas (**Table 1**). As a comparator cohort, PBMC samples from ten Dutch individuals recruited in Leiden, The Netherlands (median age 29 [IQR 27-30], 50% female) were acquired using mass cytometry (referred to as 'urban European').

Cellular immune profiles differ between rural- and urban-living Tanzanian adults.

To characterize the cellular immune profiles between rural- and urban-living individuals, peripheral blood mononuclear cells (PBMCs) were stained with a

panel of 37 metal-tagged antibodies. The processed single-cell level dataset contained 69.6 million live CD45+ cells, which allowed the identification of six major immune lineages, including B cells, CD4+ T cells, CD8+ T cells, innate lymphoid cells (ILCs), myeloid cells and unconventional T cells (including γδ T cells) (**Figure 1C**). Clustering analyses using self-organizing maps (SOM), followed by hierarchical clustering resulted in 80 distinct immune cell clusters (**Figure S1** and **Table S2**). Cell clusters were annotated at subset-level by an expert immunologist. Cell labels were further refined by incorporating markers that exhibit variability within a given subset in the cell label. Using Generalized Linear Mixed Models (GLMMs), we identified nine clusters which were significantly different between the four locations, after adjusting for age and sex (**Figure 1D-E**).

The CD4+ T cell lineage was composed of 28 cell clusters, of which 5 significantly differed across locations. Th2 cells (cluster 51) represented the strongest rural signal, where we observed significantly higher frequencies in rural-living locations (especially rural Moshi) compared to urban-living individuals (median 0.7% of total CD45+ cells across rural sites compared to 0.3% and 0.2% in urban Tanzanians and Europeans, respectively). Rural-living individuals additionally showed a significantly higher frequencies of three cell clusters of CD4+ T cells. These clusters included CD161dim PD-1dim CTLA-4+ CD4+ T effector memory (Tem) cells (cluster 46), CD4+ Tem cells expressing CD38, CD161, CTLA-4 and PD-1 (cluster 79) and HLA-DRdim PD-1+ KLRG-1+ CD4+ Tem cells (cluster 72). In contrast, the CD27+ CD28+ CD45RO+ CD127+ CD4+ T central memory (Tcm) cell cluster (cluster 53) was higher in urban compared to rural-living individuals (**Figure 1E**).

> Description figure 1. A) Map of study sites in Tanzania and in The Netherlands. B) Graphical representation of sample numbers and the study design. C-D) t-distributed Stochastic Neighbor Embedding (t-SNE) visualizations (*n* = 1500 random cells/individual); cells are coloured according to lineage (C) or significant cell cluster (D). E) Differential cell frequencies between rural and urban Tanzanian regions. Boxplots represent the 25th and 75th percentiles (lower and upper boundaries of boxes, respectively), the median (middle horizontal line) and measurements that fall within 1.5 times the interquartile range (IQR; distance between 25th and 75th percentiles; whiskers). Only clusters showing a significant effect of 'location' (across Tanzanian sites) were shown. The significance of 'location' was assessed using analysis of variance (ANOVA)-tests comparing a full (location, age [scaled] and sex [fixed effects] and sample ID [random effect]) and a simpler model, which was the same as the full model, except that we removed 'location' from the model. ANOVA p-values were corrected for multiple testing using the Benjamini-Hochberg method and referred to as q-values. Asterisks denote statistical significance (*, q ≤ 0.05; **, q ≤ 0.01; ***, q ≤ 0.001). The statistical significance of differences between each location was assessed using the *emmeans()*-function (Tukey *post hoc* test). Urban Europeans were included in the figure for visual comparisons and were not included in statistical tests.

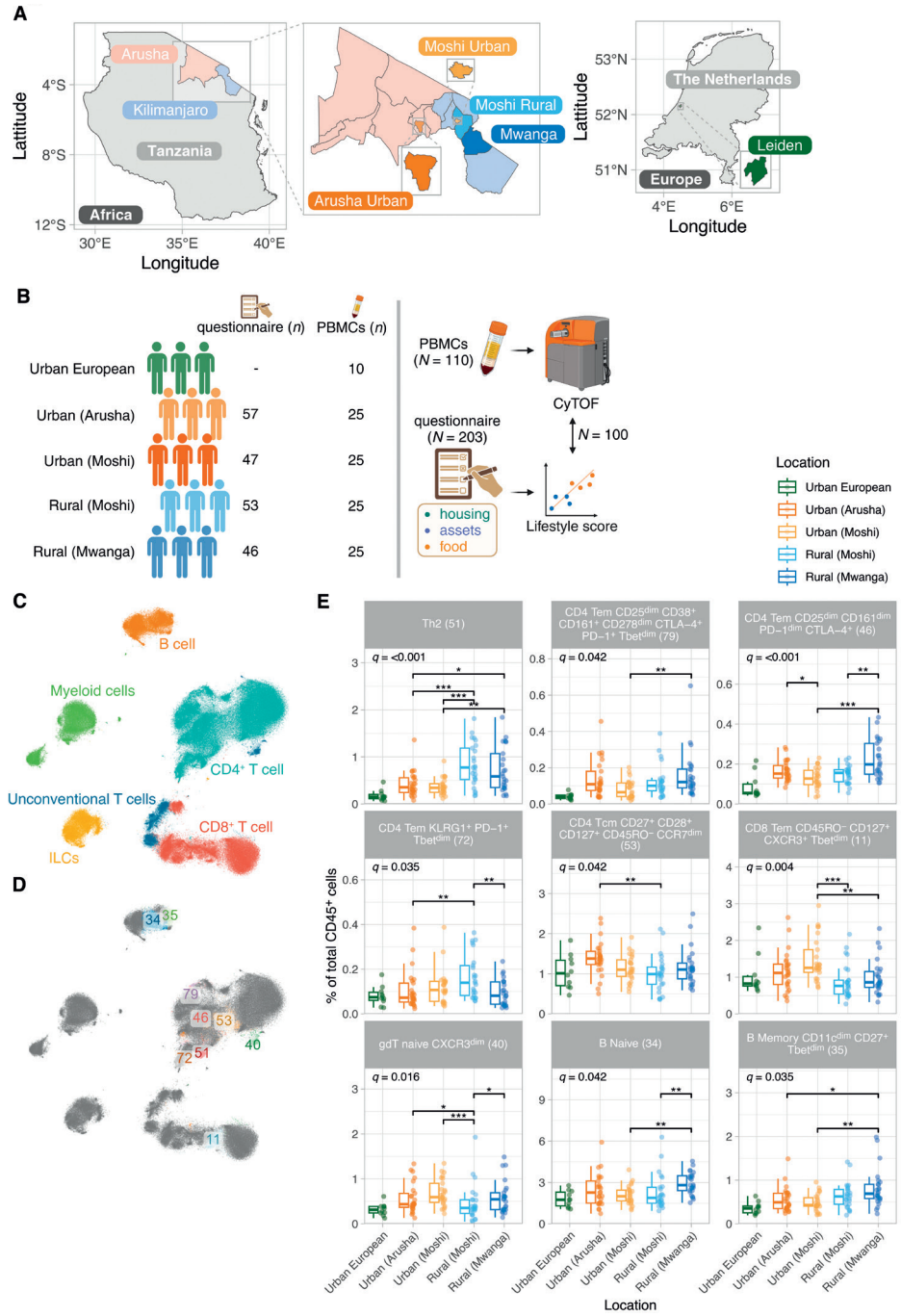

Figure 1. Mass cytometry immune profiles differ across individuals living in rural (Moshi Rural and Mwanga) and urban (Arusha and Moshi Urban) regions.

Table 1. Baseline characteristics of the study population (*N* = 100).

Variable	Overall, N = 100	Urban Arusha, N = 25	Urban Moshi, N = 25	Rural Moshi, N = 25	Rural Mwanga, N = 25	p-value
Sex, female	53 (53%)	14 (56%)	14 (56%)	13 (52%)	12 (48%)	0.932
Age	25.0 (23.0, 29.0)	25.0 (23.0, 30.0)	25.0 (24.0, 27.0)	24.0 (22.0, 27.0)	25.0 (22.0, 31.0)	0.686
Age categories						0.955
18-25	56 (56%)	13 (52%)	14 (56%)	15 (60%)	14 (56%)	
26-36	44 (44%)	12 (48%)	11 (44%)	10 (40%)	11 (44%)	
BMI	22.8 (20.5, 26.0)	21.8 (19.0, 26.8)	24.1 (22.9, 28.4)	22.3 (20.3, 26.7)	22.4 (21.3, 24.6)	0.243
Missing	1	1	0	0	0	
BMI classification						0.591
<18.5	7 (7.1%)	3 (13%)	2 (8.0%)	1 (4.0%)	1 (4.0%)	
18.5-24.9	60 (61%)	14 (58%)	13 (52%)	15 (60%)	18 (72%)	
25.0-29.9	16 (16%)	2 (8.3%)	5 (20%)	4 (16%)	5 (20%)	
>30	16 (16%)	5 (21%)	5 (20%)	5 (20%)	1 (4.0%)	
Missing	1	1	0	0	0	
Systolic blood pressure (mmHg)	119 (110, 125)	110 (109, 120)	110 (100, 119)	121 (112, 130)	123 (119, 128)	<0.001
Missing	1	1	0	0	0	
Diastolic blood pressure (mmHg)	73 (70, 79)	70 (70, 77)	69 (64, 72)	78 (70, 80)	78 (74, 80)	<0.001
Missing	1	1	0	0	0	
Hemoglobin level g/dl	14.35 (13.30, 16.50)	14.00 (13.30, 16.60)	13.80 (12.40, 15.60)	14.20 (13.70, 16.00)	15.20 (13.80, 16.60)	0.223
Random blood sugar, mmol-1^^	5.20 (4.60, 5.95)	4.90 (4.40, 5.50)	5.20 (4.70, 6.23)	5.20 (4.10, 5.50)	5.80 (4.90, 6.50)	0.053
Missing	1	0	1	0	0	
Highest level of education						<0.001
Primary	30 (30%)	0 (0%)	0 (0%)	13 (52%)	17 (68%)	
Secondary	24 (24%)	6 (24%)	0 (0%)	10 (40%)	8 (32%)	
College	15 (15%)	12 (48%)	1 (4.0%)	2 (8.0%)	0 (0%)	
University	31 (31%)	7 (28%)	24 (96%)	0 (0%)	0 (0%)	
Malaria	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Missing	1	0	1	0	0	
Helminth infection[a]	7 (7.0%)	0 (0%)	0 (0%)	2 (8.0%)	5 (20%)	0.015
Schistosomiasis[b]	3 (3.0%)	0 (0%)	0 (0%)	0 (0%)	3 (12%)	0.057
Missing	1	1	0	0	0	
Insurance status	31 (31%)	13 (52%)	15 (60%)	3 (12%)	0 (0%)	<0.001
Occupation						<0.001
Farming	20 (20%)	0 (0%)	1 (4.0%)	5 (20%)	14 (56%)	
Elementary occupation	28 (28%)	5 (20%)	2 (8.0%)	16 (64%)	5 (20%)	
Student	23 (23%)	5 (20%)	15 (60%)	2 (8.0%)	1 (4.0%)	
Employed/business owner	20 (20%)	10 (40%)	5 (20%)	2 (8.0%)	3 (12%)	
Not employed	9 (9.0%)	5 (20%)	2 (8.0%)	0 (0%)	2 (8.0%)	

N = 100 participants. Values represent number of participants (percentage of total) and median (interquartile range [IQR]) for categorical and continuous variables, respectively. Comparisons between locations were performed using Fisher's exact, chi-squared and Mann–Whitney U-test for categorical and continuous variables, respectively. [a] Stool was tested for helminths using the Kato-Katz method, testing for *Schistosoma haematobium*, *Schistosoma mansoni*, *Ascaris Lumbricoides*, hookworm and *Trichuris trichuria*. [b] Tested for schistosomiasis using the POC-CCA method, testing for *Schistosoma haematobium* and *Schistosoma manso*

Within the CD8+ T cell lineage, 1 out of 15 CD8+ T cell clusters significantly differed across locations. This cluster was characterized by recently activated CD8+ Tem cells expressing CXCR3 and T-bet (cluster 11), which showed higher frequencies in urban compared to both rural locations (**Figure 1E**). Furthermore, within the gamma delta (γδ) T cell lineage (containing 7 clusters), naïve γδ T cells expressing CXCR3 (cluster 40) were significantly higher in frequency in urban living compared to both rural-living individuals. Finally, within the B cell lineage, we observed significantly higher frequencies of classical naive B cells (cluster 34) and atypical memory B cells expressing CD11c and Tbet (cluster 35) in rural- compared to urban-living locations (**Figure 1E**). Six out of seven rural-associated clusters showed visual evidence of a rural-urban-European gradient, where cell frequencies showed a stepwise decrease from rural-to-urban and urban-to-European sites, except for cluster 40 (naïve γδ T cells). On the other hand, gradients were less clear for clusters enriched in urban Tanzanians.

Questionnaire data reveal differences in lifestyle between locations.
Within living locations, considerable variation in immune signatures was observed. Therefore, to better capture immune variation across locations, we developed a lifestyle score, which incorporates detailed questionnaire data on assets (e.g. possession of a watch, television or car), housing (i.e. materials used to construct the house) and food history (i.e. frequency of consumption of dietary products) into a single score. To obtain the lifestyle score, we applied Multiple Correspondence Analysis (MCA), a dimensionality reduction method similar to Principle Component Analysis (PCA), but for categorical data, which was applied to 38 questions (118 variable categories) collected from all 203 participants (**Table S3** and **Figure S2**). MCA clearly separated individuals based on living location, especially across principal component (PC) 1. Since the MCA was based on lifestyle questionnaire data and PC1 per definition explains most variance, PC1 was referred to as 'lifestyle score', explaining 7.8% of the variation in the questionnaire data (**Figure 2A**). Across the first two principal components, we found that spread was highest in rural- compared to urban-living individuals (variance 6.1%/5.1% and 11.3%/11.2% for PC1/PC2 scores across urban and rural sites, respectively), indicating rural people have more heterogeneous lifestyles (**Figure 2B**). Sensitivity analyses on condensed questionnaire data (collapsing rare categories and removing uninformative variables) showed that the relatively low percentage of variance explained by lifestyle score and other high-ranking principle components (**Figure S3A**) is caused by the inclusion of rarer variable categories. Removing these had no important effect on the lifestyle score (Pearson r = 0.97, p-value < 2.2 × 10-16).

We found that the lifestyle score was significantly associated with thirteen of 80 cell clusters, while none of the other principal components (PC2-PC5) showed any statistically significant associations with cell cluster frequencies (**Figure S3B**), underscoring the validity and biological relevance of the lifestyle score.

Next, we explored the most strongly contributing lifestyle score variables across questionnaire categories, including housing conditions, assets and food history. Overall, assets showed the highest cumulative contribution to the lifestyle score (53.6%), followed by housing (30.3%) and food variables (16.1%) (**Figure 2D**). Among the top 20 variables most strongly contributing to PC1, factors such as having a house with an earth/sand floor, a mud wall, no household electricity and a pit latrine as toilet were associated with low lifestyle score. Additionally, the lack of assets such as an ironing tool, refrigerator, computer, radio, car, television, or watch and not consuming potatoes was associated with a low lifestyle score. Factors associated with a high lifestyle score were a house with a flush toilet connected to a sewage/septic tank, a separate room used as a kitchen and possessing assets such as a car, a working computer and a refrigerator (**Figure 2E**).

Besides lifestyle score (PC1), we found that PC2 explained 4.1% of the variance (**Figure S3A**) and showed the highest spread across individuals living in rural Mwanga (variance across PC2 scores 15.0% compared to 2.9%-7.0% in other sites) (**Figure 2B**). Similar to PC1, variables related to assets were most important (cumulative contribution 66.0%), particularly those related to livestock farming (**Figure S3C**). PC3 through PC5 explained 3.2-3.5% of the variance (**Figure S3A**), generally showing a higher cumulative contribution of food variables (40.3-49.4%) (**Figure S3C**) compared to PC1 and PC2.

Lifestyle score association tests reveal additional immune cell clusters not previously linked to living location
We next assessed the association between lifestyle score and immune cell frequencies using GLMMs, adjusting for age and sex. We first verified that lifestyle score in individuals with matching mass cytometry data (n = 100), which was not significantly different from individuals without mass cytometry data available (**Figure S4**).

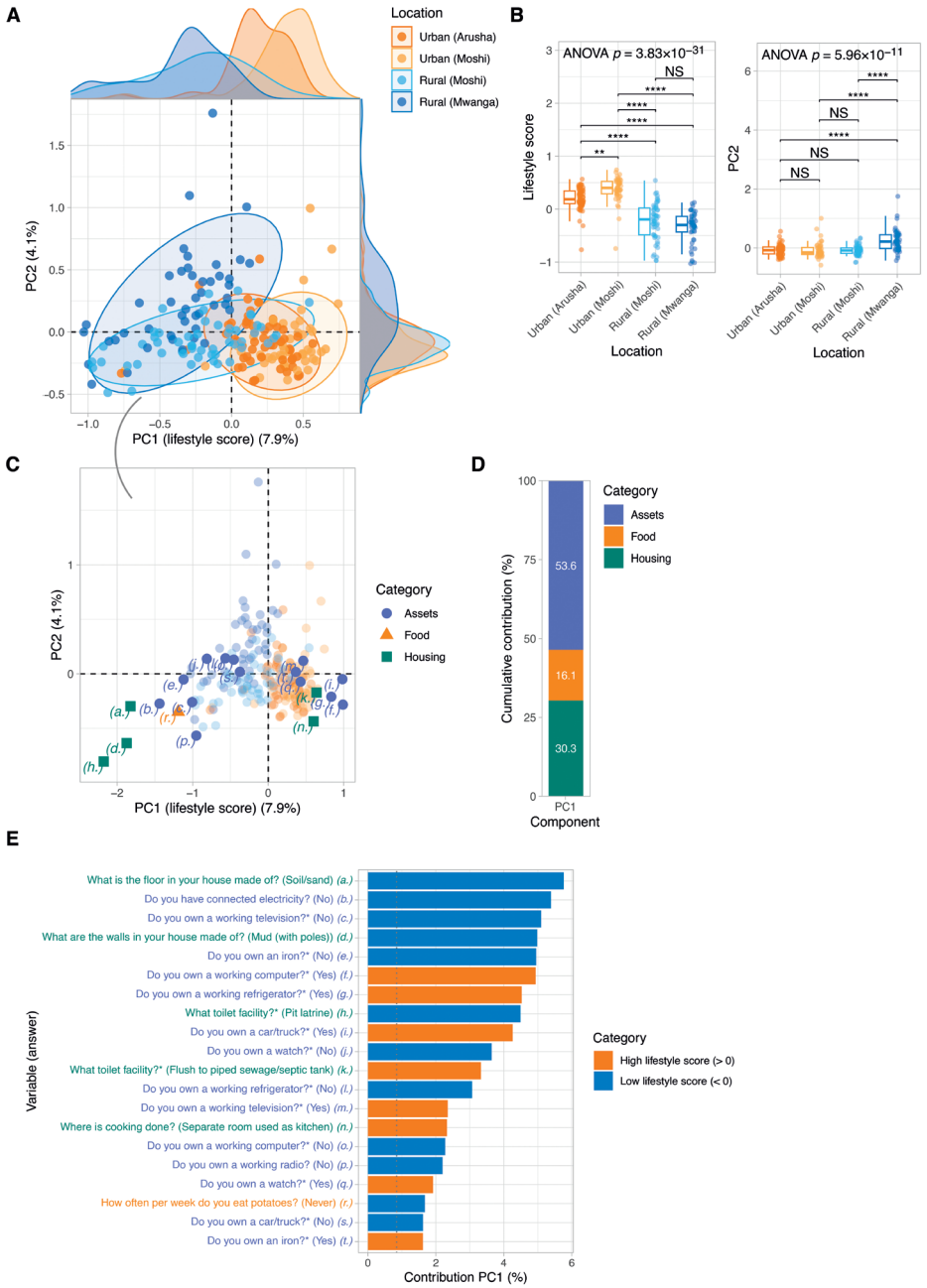

Figure 2. Multiple Correspondence Analysis (MCA) based on questionnaire data to generate lifestyle score.
A) MCA was applied to categorical questionnaire data (38 manually curated questions; 21 on assets, 11 on food and 6 on housing) (*N* = 203 individuals). Data points are coloured based on location. Ellipses reflect the data spread at a level of

confidence of 95%. Density plots show the distribution of PC1 (lifestyle score) (x-axis) and PC2 (y-axis) score. B) Comparisons of PC1 (lifestyle score) and PC2 across locations. Global significance was assessed using analysis of variance (ANOVA) and *post hoc* tests between locations were performed using Tukey HSD tests. Asterisks denote statistical significance (NS, non-significant; *, p ≤ 0.05; **, p ≤ 0.01; ***, p ≤ 0.001, p ≤ 0.0001). C) Coordinates of each variable category (a.-t.; see E) across dimensions 1 and 2. Variable categories with similar profiles are grouped together. D) Cumulative contributions (in percentage) of the variable categories by questionnaire data category (i.e. housing, assets and food). E) Contributions (in percentage) of variable categories to PC1 or lifestyle score. Bars are coloured based on whether a variable was associated with a high (> zero) or low (< zero) lifestyle score.

Overall, 13 cell clusters were associated with lifestyle score, of which 8 clusters were not identified by previous analyses where we assessed differences in immune profile between locations (**Figure 3A** and **3B**). Indeed, only one of these clusters (cluster 12; CD8+ naïve) showed a trend towards significance across locations (q = 0.055; **Figure S5**). In addition, we confirmed 5 out of 9 clusters which were previously found to significantly differ across locations, which were Th2 cells (cluster 51; GLMM; β = -0.66), two CD4+ Tem clusters that were CTLA-4+ and/or CD161+ (cluster 79 and 46; β = -0.50 and -0.28, respectively), atypical memory B cells (cluster 35; β = -0.37) (rural-living location and low lifestyle score) and a CD8+ Tem cluster (cluster 11; β = 0.32) (urban-living location and high lifestyle score) (**Figure 3C**). The additional clusters identified using the lifestyle score were two CD4+ Tem cell clusters that were associated with low lifestyle score: HLA-DR+ PD-1+ CD4+ Tem (cluster 43; β = -0.38) and regulatory T cells (cluster 75; β = -0.35). Furthermore, we identified a cluster of plasmablasts (cluster 57; β = -0.49), which was enriched in those with low lifestyle score. Last, an innate immune cell cluster of NK-cells (cluster 25; β = -0.68) was also linked to a low lifestyle score (**Figure 3D**).

In contrast, within the CD8+ T cell lineage, we identified three clusters of CD8+ T cells that were associated with high lifestyle score. These included two CD8+ naïve T cell clusters (cluster 12 and 21; β = 0.38 and 0.39, respectively) and a cluster of CD8+ Tem cells expressing CD161 and KLRG1 (cluster 38; β = 0.59). In addition, we found a positive association between higher frequencies of ILC2 (cluster 60; β = 0.33) and a high lifestyle score (**Figure 3D**). Sensitivity analyses, where we jointly modelled lifestyle score and location and compared the model fit to simpler models (excluding either lifestyle score or location), indicated that indeed using lifestyle score we can detect an additional group of clusters which we could not have detected with location alone (**Figure S6**).

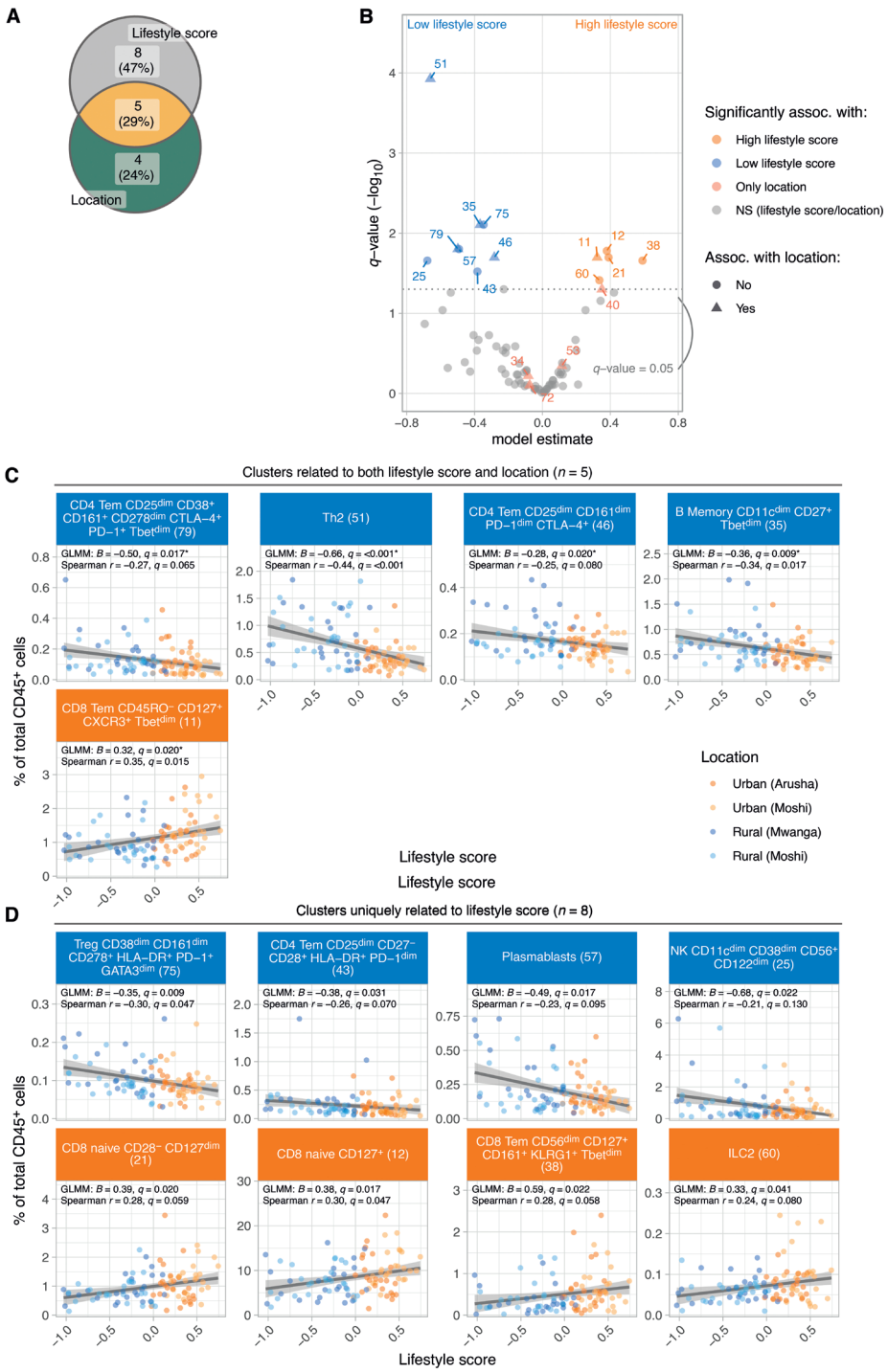

Figure 3. Lifestyle score is associated with specific immune cell clusters not identified by comparisons across locations.

< Description figure 3. A) Venn diagram indicating the number of cell clusters that show differences in cell frequencies 1) across locations [Figure 1E], 2) both across locations and lifestyle score [Figure 3C] and 3) only with lifestyle score [Figure 3D]. Eight cell clusters were uniquely associated with lifestyle score and were not identified by comparisons across sampling locations. B) Volcano plot showing differential frequency results. Results were derived from a GLMM with cell frequency as outcome variable, lifestyle score, age (scaled) and sex as fixed effects and sample ID as a random effect. Model estimates and corresponding Benjamini-Hochberg (BH)-adjusted p-values (-$\log_{10}$(q-value)) were shown. Each point represents a cluster, clusters with q-values<0.05 are coloured by association (high or low lifestyle score, or only significantly associated with location). Shapes indicate whether lifestyle-associated clusters were also detected by comparisons across sampling locations. Each point is labelled with a cluster identifier. C-D) Scatter plots showing the association between lifestyle score and cell frequency for C) clusters significantly related to both location as well as lifestyle score and D) clusters uniquely related to lifestyle score (i.e. clusters not identified as differentially abundant between locations). Data points are coloured based on location. Lines represent linear fits to the data and are included for visualization purposes only. Statistical significance was assessed using a linear mixed model including lifestyle score, age (scaled) and sex as fixed effects and sample ID as random effect. Additionally, we ran univariable Spearman correlation tests, p-values were corrected for multiple testing using the Benjamini-Hochberg method (q-value). Asterisks indicate clusters that significantly differed between locations. Only cell clusters significant in GLMMs are shown.

Machine learning modelling links a combined immune endotype with a lifestyle score

To investigate if a combination of immune cell clusters could be identified that together is associated with a lifestyle score (‘immune endotype’), a machine learning model (elastic net) was trained with lifestyle score as an outcome and cell cluster frequencies, age and sex as the predictor variables. Model training and hyperparameter tuning were performed on 80% of the data (n = 80 individuals; 2,000 bootstrapped datasets) and the model was tested on the remaining 20% of the data (n = 20 individuals) (**Figure 4A**). The model was able to predict 44.1% and 29.6% of the variance in the training and test data, respectively. Using feature importance analysis, we verified 11 of the 14 clusters that were previously associated with living location and/or lifestyle score. Compared to previous analyses, the current model is a multivariable model, estimating the contribution of each cell cluster to the prediction of lifestyle score while adjusting for all other cluster cell frequencies. Therefore, using this complementary approach, we identified three additional clusters, including CD8+ Tem cells expressing CD161 and KLRG1 (cluster 37) associated with high lifestyles score, pDCs (cluster 58) and γδ T-cells (cluster 22) related to low lifestyle score (**Figure 4B**).

Taken together the elastic net model unveiled a fairly stable (**Figure 4C**) immune endotype characterized by Th2 cells, regulatory T cells, atypical B memory cells, plasmablasts, NK, CTLA-4+ CD161+ CD4+ Tem, KLRG1+ γδ T-cells and plasmacytoid dendritic cells (pDCs) associated with a low lifestyle score. Inversely, the immune profile characterized by CD8+ naïve T cells, CXCR3+ CD127+ CD8+ Tem, two CD8+ Tem CD161+ CD56dim KLRG1+ and ILC2 is associated with a high lifestyle score (**Figure 4B**).

DISCUSSION

Here, we assessed the associations between location and/or lifestyle score and cellular immune profiles measured by mass cytometry. We found that seventeen of 80 clusters were associated with location or lifestyle score, with eight identifiable only when using lifestyle score, illustrating the ability of lifestyle score to capture immune variation. Indeed, individuals living in rural areas may exhibit an urban lifestyle and vice versa. This was further substantiated by applying a machine learning model, which identified a combined immune signature associated with lifestyle score.

We found an association between low lifestyle score and expression of activation markers such as CD38, HLA-DR and CTLA-4 on CD4+ Tem cells, along with expansion of Th2 and an increased frequency of regulatory T cells expressing CTLA-4. An increase in a specific memory T cell subsets might indicate that fewer naïve T cells are available for activation and expansion upon encounter with a new antigen. Furthermore, expression of activation/inhibitory markers on T cells can result in a reduced response to vaccines and allergens but may also explain a lower prevalence of autoimmune diseases in LMICs [19, 24, 36]. Indeed, in rural Senegalese, immune profiles were enriched for HLA-DR-expressing CD4+ T cells compared to urban-living individuals [2]. Previous studies comparing rural and urban populations in Indonesia [1, 25] and Gabon [26, 37] found that immune profiles in rural-living individuals, characterized by high frequencies of Th2 cells, T regulatory cells expressing CTLA-4, HLA-DR, ICOS or CD161 and atypical memory B cells, were strongly linked to (chronic) helminth infections [1, 25, 26].

In contrast to these previous studies, none of our participants tested positive for malaria and the prevalence of current helminth infections was very low. Therefore, we speculate that increased activation of CD4+ Tem cells, along with expansion of Th2 and higher regulatory T cell frequencies, may represent an immune footprint left behind by parasitic infection in the past or even during childhood, as have been suggested by others [24, 38, 39].

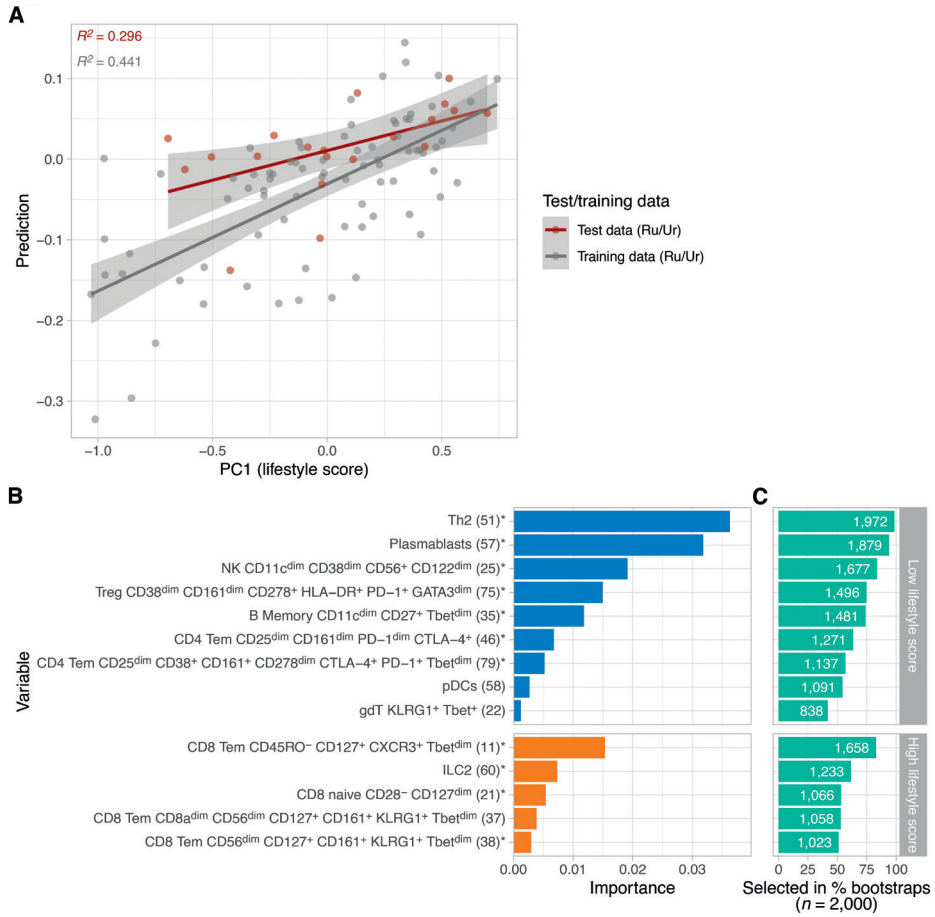

Figure 4. Machine learning model based on cell cluster frequencies can partly reconstruct lifestyle score. A) Performance of an elastic net machine learning model based on cell cluster frequencies (*n* = 80), age and sex trained to predict lifestyle score. Observed compared to predicted lifestyle score based on training (80%) and test data (20%; *n* = 5 samples per location) are shown. Using cell frequency data, we can explain ~30% of the variance in lifestyle scores (leave-out test data). B) Feature importance of all features that remained in the model after feature shrinkage/regularization. Clusters previously associated with either location or lifestyle score (*n* = 17) are indicated (*). Three clusters have not been associated with location nor lifestyle score in previous analyses. C) Feature stability across bootstraps. All features from the models fitted with the optimized tuning parameters (penalty/mixture) were extracted. The number of times a feature was selected across bootstrap samples serves as a score for stability of that feature (maximum score = 2,000).

Indeed, in 2005, the prevalence of schistosomiasis among school-aged children in two different schools located in one of the rural areas included in this study ranged between 34-70% with evidence for the presence of other soil-transmitted infections in the same setting [40]. Thus, based on their age, our study participants likely experienced a high burden of helminth infections during childhood.

Alternatively, housing conditions related to a low lifestyle score (e.g. sand or earth floors and mud-wall houses) may predispose to different commensals or exposure to bacteria and fungi and their metabolites [41], some of which have immunomodulatory properties. Poor housing conditions also attract vectors like flies, lice, ticks, mites and mosquitoes, which may directly activate the immune system through components present in their saliva, even in the absence of disease transmission [31, 42]. Furthermore, rural-living individuals closely live with livestock and as such are exposed to an additional reservoir of micro-organisms and (zoonotic) pathogens [43]. Taken together, past (parasitic) infections or unmeasured variables, such as the microbiome or exposure to vectors, are tightly linked to housing conditions. These factors may drive lifestyle-related immune variation, resulting in enrichment of Th2, regulatory T cells and activated T cells.

We found that individuals with low lifestyle score most of whom live in rural settings, display a higher frequency of plasmablasts. Plasmablasts are differentiated B cells with a short lifespan, which initiate early antibody responses during infections [44-46]. However, due to their high metabolic activity, the rapid development of short-lived plasmablasts can paradoxically impair humoral immunity by slowing down germinal centre formation. This, in turn may impair responsiveness to vaccines and reduce risk of developing allergies and autoimmunity by limiting the generation of long-lived plasma and memory B cells. Although this has been shown in the context of malaria infection [47], which is not endemic in northern Tanzania, other infectious diseases endemic in the area, may similarly induce high levels of plasmablasts, including dengue [48].

Last, we identified an association between both naïve CD8+ T cells and CD8+ Tem expressing CD161 and high lifestyle score. Although we lack immune markers to confirm, CD161+CD8+ Tem encompasses mucosal-associated invariant T cells (MAIT) cells. MAIT cells are abundant in blood and at mucosal sites and can activate dendritic cells that promote T follicular helper cells to induce mucosal antigen-specific IgA [49]. Therefore, the presence of such cells in urban-living individuals might indicate the propensity to react more strongly to antigens in a vaccine, allergens, or autoantigens. This aligns with

the results of an earlier study indicating that healthy individuals residing in urban Moshi had a higher pro-inflammatory cytokine response upon pathogen challenge in an ex vivo PBMC stimulation assay compared to those living in rural areas [7, 35]. Regarding the naïve CD8+ T cells being enriched in urban living, it has been noted that they allow new immune responses to be mounted to both infections and vaccines [50]. Their higher frequency in urban areas is in line with previous studies in Bangladeshi compared to (urban living) North American children within the first three years of life [51] as well as in Malawian compared to UK adults [52]. Reduced numbers of naïve CD8+ T cells was associated with a higher burden of intestinal worms and viral infections (e.g. CMV) in children from Bangladesh compared to those from the USA [3] and higher burden of CMV among Malawian adults [52]. Similarly, we speculate that the association between high life score and naive CD8+ T cells in our study is driven by reduced pathogen exposure in people living in urban settings due to differences in daily activities and hygiene practices compared to rural-living individuals.

The strengths of this study include the use of mass cytometry data in combination with the availability of detailed information on housing, assets and food history. Condensing this information into a single score allowed us to train a machine learning model to identify a distinct group of cell clusters (termed 'immune endotype'), which was strongly associated with lifestyle score variation. Previous studies in HICs indicated that baseline (gene-expression-based) immune endotypes exhibiting a strong pro-inflammatory profile are predictive of improved vaccine responses in young adults across multiple vaccines [53]. In a similar fashion, we speculate the immune endotypes identified in this study are linked to vaccine responses in populations living in rural or urban Africa. As such, further phenotyping of immune endotypes in varied populations, not limited to HIC, using protein-based single-cell modalities such as mass cytometry, may deepen our understanding of variation in vaccine responses or reactivity to allergens or autoantigens and their underlying mechanisms. At the same time, using lifestyle scores opens opportunities for public health experts to screen individuals prone to, for example, vaccine hypo-responsiveness, informing policymakers on preventative measures, such as repeated vaccination. These interventions could target these high-risk individuals, potentially improving vaccine efficacy and public health outcomes. Since those mounting reduced vaccine responses are the very same individuals that also show lower responses to allergens and auto-antigens, immune phenotyping may also unveil new ways to prevent non-communicable diseases in urban-living individuals. Our study also has limitations. Among others, we did not assess cellular immune function through stimulation assays. In addition,

future studies establishing direct links between low lifestyle score and responses to vaccines, allergens and autoantigens would be of great value.

In conclusion, in this study we comprehensively assessed the association between immune profiles and location and lifestyle variables in a LMIC. Additional cell clusters were detected through a more refined measurement of lifestyle. Follow-up studies should therefore focus on the links between lifestyle score, immune signature and functional immune responses, particularly in populations where vaccine responses are expected to be reduced and in populations with the highest prevalence of diseases linked to exaggerated immune responses to allergens and autoantigens.

Acknowledgements

This work was supported by grants from the Dutch Research Organization (NWO) through the Spinoza prize awarded to Maria Yazdanbakhsh, the European Research Council (ERC) via the ERC Advanced Grant 'REVERSE' awarded to Maria Yazdanbakhsh (Grant No: 101055179), the LUMC Excellent Student Fellowship awarded to Marloes M.A.R. van Dorst and the LUMC Global PhD Fellowship awarded to Jeremia J. Pyuza. We would to like acknowledge all clinical and research staff at KCRI and KCMC in Tanzania who helped to make this study possible. We would also like to acknowledge the LUMC core facility for providing mass cytometry services. Finally, we would like to thank all volunteers who participated in this study

References

1. de Ruiter, K., et al., Helminth infections drive heterogeneity in human type 2 and regulatory cells. Science Translational Medicine, 2020. 12(524).
2. Mbow, M., et al., Changes in immunological profile as a function of urbanization and lifestyle. Immunology, 2014. 143(4): p. 569-577.
3. Wager, L.E., et al., Increased T Cell Differentiation and Cytolytic Function in Bangladeshi Compared to American Children. Frontiers in Immunology, 2019. 10.
4. Muyanja, E., et al., Immune activation alters cellular and humoral responses to yellow fever 17D vaccine (vol 124, pg 3147, 2014). Journal of Clinical Investigation, 2014. 124(10): p. 4669-4669.
5. Smolen, K.K., et al., Pattern recognition receptor-mediated cytokine response in infants across 4 continents. Journal of Allergy and Clinical Immunology, 2014. 133(3): p. 818-+.
6. de Jong, S.E., et al., Systems analysis and controlled malaria infection in Europeans and Africans elucidate naturally acquired immunity. Nature Immunology, 2021. 22(5): p. 654-+.
7. Temba, G.S., et al., Urban living in healthy Tanzanians is associated with an inflammatory status driven by dietary and metabolic changes. Nature Immunology, 2021. 22(3): p. 287-+.
8. Anuradha, R., et al., Parasite Antigen-Specific Regulation of Th1, Th2, and Th17 Responses in Infection. Journal of Immunology, 2015. 195(5): p. 2241-2250.
9. Kemp, K., B.D. Akanmori, and L. Hviid, West African donors have high percentages of activated cytokine producing T cells that are prone to apoptosis. Clinical and Experimental Immunology, 2001. 126(1): p. 69-75.
10. Domingo, C., et al., Long-term immunity against yellow fever in children vaccinated during infancy: a longitudinal cohort study. Lancet Infectious Diseases, 2019. 19(12): p. 1363-1370.
11. van Dorst, M., et al., Immunological factors linked to geographical variation in vaccine responses. Nat Rev Immunol, 2023.
12. Tsang, J.S., et al., Improving Vaccine-Induced Immunity: Can Baseline Predict Outcome? Trends in Immunology, 2020. 41(6): p. 457-465.
13. Nehar-Belaid, D., et al., Baseline immune states (BIS) associated with vaccine responsiveness and factors that shape the BIS. Seminars in Immunology, 2023. 70.
14. Shannon, C.P., et al., Multi-Omic Data Integration Allows Baseline Immune Signatures to Predict Hepatitis B Vaccine Response in a Small Cohort. Frontiers in Immunology, 2020. 11.
15. Tsang, J.S., et al., Global Analyses of Human Immune Variation Reveal Baseline Predictors of Postvaccination Responses (vol 157, pg 499, 2014). Cell, 2014. 158(1): p. 226-226.
16. Avey, S., et al., Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses. Science Immunology, 2017. 2(14).
17. Okada, H., et al., The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. Clinical and Experimental Immunology, 2010. 160(1): p. 1-9.
18. Murdaca, G., et al., Hygiene hypothesis and autoimmune diseases: A narrative review of clinical evidences and mechanisms. Autoimmunity Reviews, 2021. 20(7).

19. Bach, J.F., Mechanisms of disease: The effect of infections on susceptibility to autoimmune and allergic diseases. New England Journal of Medicine, 2002. 347(12): p. 911-920.
20. Brodin, P., et al., Variation in the human immune system is largely driven by non-heritable influences. Cell, 2015. 160(1-2): p. 37-47.
21. Klein, S.L. and K.L. Flanagan, Sex differences in immune responses. Nat Rev Immunol, 2016. 16(10): p. 626-38.
22. Brodin, P. and M.M. Davis, Human immune system variation. Nat Rev Immunol, 2017. 17(1): p. 21-29.
23. Liston, A., et al., Human immune diversity: from evolution to modernity. Nat Immunol, 2021. 22(12): p. 1479-1489.
24. Lubyayi, L., et al., Infection-exposure in infancy is associated with reduced allergy-related disease in later childhood in a Ugandan cohort. Elife, 2021. 10.
25. Wammes, L.J., et al., Community deworming alleviates geohelminth-induced immune hyporesponsiveness. Proceedings of the National Academy of Sciences of the United States of America, 2016. 113(44): p. 12526-12531.
26. Labuda, L.A., et al., A Praziquantel Treatment Study of Immune and Transcriptome Profiles in-Infected Gabonese Schoolchildren. Journal of Infectious Diseases, 2020. 222(12): p. 2103-2113.
27. Kaczorowski, K.J., et al., Continuous immunotypes describe human immune variation and predict diverse responses. Proceedings of the National Academy of Sciences of the United States of America, 2017. 114(30): p. E6097-E6106.
28. Yan, Z., et al., Aging and CMV discordance are associated with increased immune diversity between monozygotic twins. Immunity & Ageing, 2021. 18(1).
29. Carr, E.J., et al., The cellular composition of the human immune system is shaped by age and cohabitation. Nature Immunology, 2016. 17(4): p. 461-+.
30. Chakraborty, N.M., et al., Simplified Asset Indices to Measure Wealth and Equity in Health Programs: A Reliability and Validity Analysis Using Survey Data From 16 Countries. Global Health-Science and Practice, 2016. 4(1): p. 141-154.
31. Wikel, S.K., Modulation of the host immune system by ectoparasitic arthropods - Blood-feeding and tissue-dwelling arthropods manipulate host defenses to their advantage. Bioscience, 1999. 49(4): p. 311-320.
32. DHS. Wealth index 2016 [cited 2024 15.01.]; Available from: https://dhsprogram.com/topics/wealth-index/#:~:text=The%20wealth%20index%20is%20a,water%20access%20and%20sanitation%20facilities.
33. Fisk, W.J., E.A. Eliseeva, and M.J. Mendell, Association of residential dampness and mold with respiratory tract infections and bronchitis: a meta-analysis. Environmental Health, 2010. 9.
34. TNBS. Tanzania population and housing census- Tanzania National Bureau of statistics(TNBS). 2022 [cited 2024; Available from: https://www.nbs.go.tz/index.php/en/census-surveys/population-and-housing-census.
35. TDHS-MIS, Tanzania Demographic and Health Survey and Malaria Indicator Survey (TDHS-MIS) 2015-16, in Ministry of Health, Community Development, Gender, Elderly and Children (MoHCDGEC) [Tanzania Mainland], Ministry of Health (MoH) [Zanzibar], National Bureau of Statistics (NBS), Office of the Chief Government Statistician (OCGS), and ICF. 2016. Tanzania Demographic and Health Survey and Malaria Indicator Survey (TDHS-MIS) 2015-16. Dar es Salaam, Tanzania, and Rockville, Maryland, USA: MoHCDGEC, MoH, NBS, OCGS, and ICF. 2016.

36. Maizels, R.M., Parasitic helminth infections and the control of human allergic and autoimmune disorders. Clinical Microbiology and Infection, 2016. 22(6): p. 481-486.
37. van Riet, E., et al., Cellular and humoral responses to influenza in Gabonese children living in rural and semi-urban areas. Journal of Infectious Diseases, 2007. 196(11): p. 1671-1678.
38. Djuardi, Y., et al., Immunological footprint: the development of a child's immune system in environments rich in microorganisms and parasites. Parasitology, 2011. 138(12): p. 1508-1518.
39. Mpairwe, H., R. Tweyongyere, and A. Elliott, Pregnancy and helminth infections. Parasite Immunology, 2014. 36(8): p. 328-337.
40. Poggensee, G., et al., A six-year follow-up of schoolchildren for urinary and intestinal schistosomiasis and soil-transmitted helminthiasis in Northern Tanzania. Acta Tropica, 2005. 93(2): p. 131-140.
41. McCall, L.I., et al., Home chemical and microbial transitions across urbanization. Nature Microbiology, 2020. 5(1): p. 108-115.
42. Vogt, M.B., et al., Mosquito saliva alone has profound effects on the human immune system. Plos Neglected Tropical Diseases, 2018. 12(5).
43. Libera, K., et al., Selected Livestock-Associated Zoonoses as a Growing Challenge for Public Health. Infectious Disease Reports, 2022. 14(1): p. 63-81.
44. Nutt, S.L., et al., The generation of antibody-secreting plasma cells. Nature Reviews Immunology, 2015. 15(3): p. 160-171.
45. Wrammert, J., et al., Rapid and Massive Virus-Specific Plasmablast Responses during Acute Dengue Virus Infection in Humans. Journal of Virology, 2012. 86(6): p. 2911-2918.
46. Fink, K., Origin and function of circulating plasmablasts during acute viral infections. Frontiers in Immunology, 2012. 3.
47. Vijay, R., et al., Infection-induced plasmablasts are a nutrient sink that impairs humoral immunity to malaria. Nature Immunology, 2020. 21(7): p. 790-+.
48. Hertz, J.T., et al., Chikungunya and Dengue Fever among Hospitalized Febrile Patients in Northern Tanzania. American Journal of Tropical Medicine and Hygiene, 2012. 86(1): p. 171-177.
49. Pankhurst, T.E., et al., MAIT cells activate dendritic cells to promote T(FH) cell differentiation and induce humoral immunity. Cell Rep, 2023. 42(4): p. 112310.
50. Jongo, S.A., et al., Safety, Immunogenicity, and Protective Efficacy against Controlled Human Malaria Infection of Sporozoite Vaccine in Tanzanian Adults. American Journal of Tropical Medicine and Hygiene, 2018. 99(2): p. 338-349.
51. Godfrey, D.I., et al., The biology and functional importance of MAIT cells. Nature Immunology, 2019. 20(9): p. 1110-1128.
52. Ben-Smith, A., et al., Differences between naive and memory T cell phenotype in Malawian and UK adolescents: a role for Cytomegalovirus? Bmc Infectious Diseases, 2008. 8.
53. Fourati, S., et al., Pan-vaccine analysis reveals innate immune endotypes predictive of antibody responses to vaccination. Nature Immunology, 2022. 23(12): p. 1777-+.
54. TDHS-MIS, The 2022 Tanzania Demographic and Health Survey and Malaria Indicator Survey (2022 TDHS-MIS). 20

Supplementary material

Table S1. Baseline characteristics of the study population (*N* = 203).

Variable	Overall, N = 203	Urban Arusha, N = 57	Urban Moshi, N = 47	Rural Moshi, N = 46	Rural Mwanga, N = 53	p-value
Sex, female	100 (49%)	40 (70%)	26 (55%)	18 (39%)	16 (30%)	<0.001
Age	25.0 (22.0, 29.5)	25.0 (22.0, 30.0)	25.0 (23.0, 27.0)	26.0 (22.3, 31.0)	24.0 (21.0, 27.0)	0.165
Age categories						0.259
18-25	116 (57%)	30 (53%)	30 (64%)	22 (48%)	34 (64%)	
26-36	87 (43%)	27 (47%)	17 (36%)	24 (52%)	19 (36%)	
BMI	22.6 (20.5, 25.6)	22.2 (19.9, 25.8)	23.9 (22.2, 26.1)	22.4 (20.7, 25.0)	22.3 (20.3, 25.3)	0.183
Missing	1	1	0	0	0	
BMI classification						0.585
<18.5	13 (6.4%)	6 (11%)	3 (6.4%)	3 (6.5%)	1 (1.9%)	
18.5-24.9	130 (64%)	34 (61%)	27 (57%)	31 (67%)	38 (72%)	
25.0-29.9	39 (19%)	10 (18%)	11 (23%)	10 (22%)	8 (15%)	
>30	20 (9.9%)	6 (11%)	6 (13%)	2 (4.3%)	6 (11%)	
Missing	1	1	0	0	0	
Systolic blood pressure (mmHg)	120 (110, 128)	110 (109, 120)	110 (103, 120)	126 (118, 130)	122 (120, 130)	<0.001
Missing	1	1	0	0	0	
Diastolic blood pressure (mmHg)	73 (68, 80)	70 (67, 79)	70 (64, 78)	78 (72, 81)	76 (70, 80)	0.001
Missing	1	1	0	0	0	
Hemoglobin level g/dl	14.50 (13.35, 16.40)	13.90 (13.10, 15.00)	13.70 (12.30, 15.30)	15.25 (14.03, 16.58)	15.80 (14.00, 17.00)	<0.001
Random blood sugar, mmol-1^^	5.00 (4.50, 5.80)	4.80 (4.40, 5.50)	5.15 (4.53, 5.85)	5.50 (4.75, 6.20)	4.70 (3.90, 5.50)	0.002
Missing	1	0	1	0	0	
Highest level of education						<0.001
Primary	50 (25%)	4 (7.0%)	2 (4.3%)	27 (59%)	17 (32%)	
Secondary	74 (36%)	18 (32%)	11 (23%)	19 (41%)	26 (49%)	
College	40 (20%)	27 (47%)	6 (13%)	0 (0%)	7 (13%)	

Table S1. Baseline characteristics of the study population (*N* = 203) - *continued*

Variable	Overall, N = 203	Urban Arusha, N = 57	Urban Moshi, N = 47	Rural Moshi, N = 46	Rural Mwanga, N = 53	p-value
University	39 (19%)	8 (14%)	28 (60%)	0 (0%)	3 (5.7%)	
Malaria	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Missing	1	0	1	0	0	
Helminth infection[a]	8 (3.9%)	0 (0%)	0 (0%)	6 (13%)	2 (3.8%)	0.002
Schistosomiasis[b]	7 (3.5%)	2 (3.6%)	1 (2.1%)	4 (8.9%)	0 (0%)	0.098
Missing	3	2	0	1	0	
Insurance status	51 (25%)	24 (42%)	23 (50%)	0 (0%)	4 (7.5%)	<0.001
Missing	1	0	1	0	0	
Occupation						<0.001
Farming	32 (16%)	2 (3.5%)	1 (2.1%)	23 (50%)	6 (11%)	
Elementary occupation	60 (30%)	14 (25%)	7 (15%)	13 (28%)	26 (49%)	
Student	47 (23%)	12 (21%)	23 (49%)	2 (4.3%)	10 (19%)	
Employed/ business owner	34 (17%)	15 (26%)	9 (19%)	4 (8.7%)	6 (11%)	
Not employed	30 (15%)	14 (25%)	7 (15%)	4 (8.7%)	5 (9.4%)	

N = 203 participants. Values represent number of participants (percentage of total) and median (interquartile range [IQR]) for categorical and continuous variables, respectively. Comparisons between locations were performed using Fisher's exact, chi-squared and Mann–Whitney U-test for categorical and continuous variables, respectively. [a] Stool was tested for helminths using the Kato-Katz method, testing for *Schistosoma haematobium*, *Schistosoma mansoni*, *Ascaris Lumbricoides*, hookworm and *Trichuris trichuria*. [b] Tested for schistosomiasis using the POC-CCA method, testing for *Schistosoma haematobium* and *Schistosoma mansoni*.

Table S2. Overview of identified cell clusters.
See Supplemental Videos and Spreadsheets.

Table S3. Descriptives of lifestyle score variables.

Characteristic	Urban Arusha, N = 57	Urban Moshi, N = 47	Rural Moshi, N = 53	Rural Mwanga, N = 46	p-value
House floor					<0.001
Hard floor (tile, cement, concrete, wood)	57 (100%)	47 (100%)	44 (83%)	33 (72%)	
Earth/sand	0 (0%)	0 (0%)	9 (17%)	13 (28%)	
House walls					<0.001
Cement, brick or stone	56 (98%)	46 (100%)	42 (79%)	39 (85%)	
Cane, palm, trunks, bamboo	0 (0%)	0 (0%)	1 (1.9%)	0 (0%)	
Mud (with poles)	1 (1.8%)	0 (0%)	10 (19%)	7 (15%)	
Missing	0	1	0	0	
House roof					0.257
Roof tiles	2 (3.5%)	2 (4.3%)	0 (0%)	0 (0%)	
Metal sheets	55 (96%)	45 (96%)	53 (100%)	45 (98%)	
Other	0 (0%)	0 (0%)	0 (0%)	1 (2.2%)	
Water source					<0.001
Tap water	51 (89%)	45 (96%)	33 (62%)	13 (28%)	
Public standpipe	3 (5.3%)	1 (2.1%)	12 (23%)	10 (22%)	
Protected tube well or bore hole	3 (5.3%)	0 (0%)	3 (5.7%)	20 (43%)	
Spring	0 (0%)	1 (2.1%)	5 (9.4%)	0 (0%)	
Pond-water or stream	0 (0%)	0 (0%)	0 (0%)	3 (6.5%)	
Toilet facility					<0.001
Flush to piped sewage or septic tank	41 (72%)	42 (89%)	17 (32%)	3 (6.5%)	
Pour flush latrine	14 (25%)	1 (2.1%)	18 (34%)	36 (78%)	
Pit latrine	2 (3.5%)	4 (8.5%)	18 (34%)	7 (15%)	
Cooking place					<0.001
In a separate room used as kitchen	32 (56%)	31 (66%)	14 (26%)	5 (11%)	
In a separate building used as kitchen	17 (30%)	9 (19%)	38 (72%)	37 (80%)	
In a room used for living or sleeping	8 (14%)	5 (11%)	1 (1.9%)	2 (4.3%)	
Outdoors	0 (0%)	2 (4.3%)	0 (0%)	2 (4.3%)	
Total number of milk cows					0.012

Table S3. Descriptives of lifestyle score variables - *continued.*

Characteristic	Urban Arusha, N = 57	Urban Moshi, N = 47	Rural Moshi, N = 53	Rural Mwanga, N = 46	p-value
None	51 (89%)	43 (91%)	40 (75%)	40 (87%)	
1-4	6 (11%)	1 (2.1%)	11 (21%)	2 (4.3%)	
5-9	0 (0%)	2 (4.3%)	1 (1.9%)	1 (2.2%)	
10+	0 (0%)	1 (2.1%)	1 (1.9%)	3 (6.5%)	
Total number of other cattle					<0.001
None	56 (98%)	46 (98%)	45 (85%)	39 (85%)	
1-4	1 (1.8%)	1 (2.1%)	8 (15%)	2 (4.3%)	
5-9	0 (0%)	0 (0%)	0 (0%)	1 (2.2%)	
10+	0 (0%)	0 (0%)	0 (0%)	4 (8.7%)	
Total number of horses					>0.999
None	57 (100%)	47 (100%)	53 (100%)	46 (100%)	
1-4	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
5-9	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
10+	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Total number of goats					<0.001
None	53 (93%)	39 (83%)	29 (55%)	30 (65%)	
1-4	3 (5.3%)	3 (6.4%)	12 (23%)	7 (15%)	
5-9	0 (0%)	2 (4.3%)	11 (21%)	5 (11%)	
10+	1 (1.8%)	3 (6.4%)	1 (1.9%)	4 (8.7%)	
Total number of sheep					0.031
None	55 (96%)	46 (98%)	52 (98%)	38 (83%)	
1-4	0 (0%)	0 (0%)	1 (1.9%)	2 (4.3%)	
5-9	1 (1.8%)	1 (2.1%)	0 (0%)	3 (6.5%)	
10+	1 (1.8%)	0 (0%)	0 (0%)	3 (6.5%)	
Total number of chicken/poultry					<0.001
None	33 (58%)	18 (38%)	8 (15%)	19 (41%)	
1-4	2 (3.5%)	2 (4.3%)	2 (3.8%)	5 (11%)	
5-9	6 (11%)	5 (11%)	11 (21%)	5 (11%)	
10+	16 (28%)	22 (47%)	31 (60%)	17 (37%)	
Missing	0	0	1	0	
Agricultural land (hectares)					0.439

Table S3. Descriptives of lifestyle score variables - *continued.*

Characteristic	Urban Arusha, N = 57	Urban Moshi, N = 47	Rural Moshi, N = 53	Rural Mwanga, N = 46	p-value
None	39 (68%)	31 (67%)	38 (72%)	30 (65%)	
1-4	12 (21%)	10 (22%)	14 (26%)	12 (26%)	
5-9	4 (7.0%)	2 (4.3%)	0 (0%)	4 (8.7%)	
10+	2 (3.5%)	3 (6.5%)	1 (1.9%)	0 (0%)	
Missing	0	1	0	0	
Connected to electricity	54 (96%)	46 (98%)	37 (70%)	32 (70%)	<0.001
Missing	1	0	0	0	
Working radio	49 (86%)	44 (94%)	42 (79%)	37 (80%)	0.185
Working television	51 (89%)	40 (85%)	22 (42%)	25 (54%)	<0.001
Missing	0	0	1	0	
Working computer	23 (40%)	37 (79%)	4 (7.7%)	0 (0%)	<0.001
Missing	0	0	1	0	
Working refrigerator	34 (60%)	38 (81%)	8 (15%)	2 (4.3%)	<0.001
Working rechargeable battery or generator	8 (15%)	13 (28%)	4 (7.5%)	11 (24%)	0.035
Missing	2	0	0	1	
An iron (charcoal/ electric)	51 (89%)	42 (93%)	38 (72%)	20 (43%)	<0.001
Missing	0	2	0	0	
Watch	44 (77%)	44 (98%)	29 (55%)	14 (30%)	<0.001
Missing	0	2	0	0	
Mobile phone	55 (96%)	47 (100%)	53 (100%)	44 (96%)	0.283
Bicycle	11 (19%)	18 (38%)	4 (7.7%)	28 (61%)	<0.001
Missing	0	0	1	0	
Motorcycle	21 (37%)	17 (37%)	12 (23%)	24 (52%)	0.026
Missing	0	1	0	0	
Animal drawn cart	0 (0%)	1 (2.2%)	0 (0%)	1 (2.2%)	0.353
Missing	1	1	0	0	
Car or truck	19 (33%)	30 (64%)	6 (11%)	1 (2.2%)	<0.001
Boat with a motor	0 (0%)	1 (2.2%)	0 (0%)	1 (2.2%)	0.353
Missing	0	1	1	1	
Ugali (stiff porridge) (×/ week)					<0.001
0	0 (0%)	2 (4.3%)	0 (0%)	0 (0%)	

Table S3. Descriptives of lifestyle score variables - *continued.*

Characteristic	Urban Arusha, N = 57	Urban Moshi, N = 47	Rural Moshi, N = 53	Rural Mwanga, N = 46	p-value
1	6 (11%)	11 (23%)	2 (3.8%)	1 (2.2%)	
2-4	26 (46%)	23 (49%)	31 (58%)	13 (28%)	
≥5	24 (43%)	11 (23%)	20 (38%)	32 (70%)	
Missing	1	0	0	0	
Plantain (×/week)					<0.001
0	19 (35%)	13 (28%)	16 (30%)	28 (62%)	
1	27 (49%)	30 (64%)	25 (47%)	17 (38%)	
2-4	5 (9.1%)	1 (2.1%)	10 (19%)	0 (0%)	
≥5	4 (7.3%)	3 (6.4%)	2 (3.8%)	0 (0%)	
Missing	2	0	0	1	
Banana (×/week)					0.152
0	7 (13%)	4 (8.5%)	2 (3.8%)	10 (22%)	
1	27 (48%)	22 (47%)	23 (43%)	23 (50%)	
2-4	19 (34%)	18 (38%)	20 (38%)	10 (22%)	
≥5	3 (5.4%)	3 (6.4%)	8 (15%)	3 (6.5%)	
Missing	1	0	0	0	
Rice (×/week)					<0.001
0	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
1	4 (7.0%)	4 (8.5%)	19 (36%)	7 (15%)	
2-4	25 (44%)	17 (36%)	28 (53%)	18 (39%)	
≥5	28 (49%)	26 (55%)	6 (11%)	21 (46%)	
Potatoes (×/week)					0.005
0	1 (1.8%)	0 (0%)	11 (21%)	3 (6.7%)	
1	26 (46%)	21 (45%)	28 (53%)	26 (58%)	
2-4	21 (37%)	19 (40%)	11 (21%)	13 (29%)	
≥5	9 (16%)	7 (15%)	3 (5.7%)	3 (6.7%)	
Missing	0	0	0	1	
Meat (×/week)					0.008
0	1 (1.8%)	1 (2.1%)	0 (0%)	2 (4.3%)	
1	13 (23%)	5 (11%)	16 (30%)	11 (24%)	
2-4	29 (52%)	20 (43%)	31 (58%)	25 (54%)	
≥5	13 (23%)	21 (45%)	6 (11%)	8 (17%)	
Missing	1	0	0	0	

Table S3. Descriptives of lifestyle score variables - *continued.*

Characteristic	Urban Arusha, N = 57	Urban Moshi, N = 47	Rural Moshi, N = 53	Rural Mwanga, N = 46	p-value
Fish (×/week)					<0.001
0	0 (0%)	3 (6.4%)	2 (3.8%)	0 (0%)	
1	25 (44%)	26 (55%)	24 (45%)	7 (15%)	
2-4	23 (40%)	15 (32%)	26 (49%)	13 (28%)	
≥5	9 (16%)	3 (6.4%)	1 (1.9%)	26 (57%)	
Beans/peas (×/week)					0.005
0	2 (3.5%)	1 (2.1%)	1 (1.9%)	0 (0%)	
1	11 (19%)	8 (17%)	20 (38%)	3 (6.5%)	
2-4	28 (49%)	21 (45%)	20 (38%)	18 (39%)	
≥5	16 (28%)	17 (36%)	12 (23%)	25 (54%)	
Green vegetables (×/week)					0.625
0	0 (0%)	1 (2.1%)	1 (1.9%)	1 (2.2%)	
1	4 (7.0%)	5 (11%)	1 (1.9%)	2 (4.3%)	
2-4	15 (26%)	10 (21%)	15 (28%)	16 (35%)	
≥5	38 (67%)	31 (66%)	36 (68%)	27 (59%)	
Fruits (×/week)					0.003
0	0 (0%)	1 (2.1%)	1 (1.9%)	0 (0%)	
1	9 (16%)	6 (13%)	21 (40%)	13 (28%)	
2-4	15 (26%)	11 (23%)	16 (30%)	18 (39%)	
≥5	33 (58%)	29 (62%)	15 (28%)	15 (33%)	
Locally brewed beer (×/week)					0.011
0	47 (82%)	40 (85%)	33 (62%)	41 (89%)	
1	6 (11%)	6 (13%)	7 (13%)	1 (2.2%)	
2-4	2 (3.5%)	1 (2.1%)	4 (7.5%)	1 (2.2%)	
≥5	2 (3.5%)	0 (0%)	9 (17%)	3 (6.5%)	

N = 203 participants. Values represent number of participants (percentage of total). Comparisons between locations were performed using Fisher's exact or chi-squared tests. All variables (*n* = 38 variables), after mode imputation, were used to construct the lifestyle score. See Figure S2.

Table S4. Mass cytometry antibody panel.

Label	Specificity	Clone	Supplier[a]	Cat no	Lot no	End dilution	Working dilution
^{89}Y	CD45	HI30	Fluidigm	3089003B	2203476-08	200	100
^{115}In	CD278 (ICOS)	C398.4A	Biolegend	313502	22-02-2022 MK	100	50
^{141}Pr	CD196 (CCR6)	G034E3	Fluidigm	3141003A	2201583-11	100	50
^{142}Nd	CD19	HIB19	Biolegend	302202	24-06-2020	500	250
^{143}Nd	CD117 (c-Kit)	104D2	Biolegend	313223	28-01-2020	500	250
^{145}Nd	CD4	RPA-T4	Fluidigm	3145001B	2202012-07	500	250
^{146}Nd	CD8a	RPA-T8	Fluidigm	3146001B	2108701-11	500	250
^{147}Sm	CD183 (CXCR3)	G025H7	Biolegend	353733	03-01-2018	100	50
^{148}Nd	CD14	M5E2	Biolegend	301802	30-05-2022	200	100
^{149}Sm	CD25 (IL-2Ra)	2A3	Fluidigm	3149010B	2104640-07	500	250
^{150}Nd	CD185 (CXCR5)	J252D4	Biolegend	356902	10-09-2019	500	250
^{151}Eu	CD123	6H6	Fluidigm	3151001B	2112140-01	500	250
^{152}Sm	TCRγδ	11F2	Fluidigm	3152008B	2110581-20	200	100
^{153}Eu	CD7	CD7-6B7	Fluidigm	3153014B	0282010	200	100
^{154}Sm	CD163	GHI/61	Fluidigm	3154007B	3321818	100	50
^{155}Gd	CD45RA	HI100	Fluidigm	3155011B	0492003	200	100
^{156}Gd	CD294 (CRTH2)	BM16	Biolegend	350102	30-05-2022	100	50
^{158}Gd	CD122 (IL-2Rb)	TU27	Biolegend	339002	01-02-2022	500	250
^{159}Tb	CD197 (CCR7)	G043H7	Biolegend	353237	11-09-2020	200	100
^{161}Dy	KLRG1 (MAFA)	REA261	Miltenyi	130-126-458	01-02-2022	500	250
^{162}Dy	CD11c	Bu15	Fluidigm	3162005B	2111081-25	500	250
^{164}Dy	CD161	HP-3G10	Fluidigm	3164009B	2111083-25	200	100
^{165}Ho	CD127 (IL-7Ra)	AO19D5	Biolegend	351302	24-09-2020	500	250
^{167}Er	CD27	O323	Biolegend	302839	11-09-2019	500	250
^{168}Er	HLA-DR	L243	Biolegend	307651	01-02-2022	200	100
^{170}Er	CD3	UCHT1	Fluidigm	3170001B	169104	200	100
^{171}Yb	CD28	CD28.2	Biolegend	302902	01-02-2022	200	100
^{172}Yb	CD38	HIT2	Fluidigm	3172007B	2108738-17	200	100

Table S4. Mass cytometry antibody panel - *continued.*

Label	Specificity	Clone	Supplier[a]	Cat no	Lot no	End dilution	Working dilution
^{173}Yb	CD45RO	UCHL1	Biolegend	304239	11-09-2019	200	100
^{174}Yb	CD335 (NKp46)	9E2	Biolegend	331902	22-12-2020	500	250
^{175}Lu	CD279 (PD-1)	EH 12.2H7	Fluidigm	3175008B	2104621-07	500	250
^{176}Yb	CD56	NCAM16.2	Fluidigm	3176008B	2202917-03	500	250
^{209}BI	CD16	3G8	Fluidigm	3209002B	2112429-15	200	100

[a]Fluidigm, South San Francisco, CA, USA; BioLegend, San Diego, CA, USA; Miltenyi Biotech, Bergisch Gladbach, Germany. CCR, CC chemokine receptor. CD, cluster of differentiation. CRTH2, prostaglandin D2 receptor 2. CXCR, CXC chemokine receptor. HLA-DR, human leukocyte antigen-D related. IL-2R, interleukin-2 receptor. IL2RB, Interleukin-2 receptor subunit beta, IL2Ra, Interleukin-2 receptor subunit alpha, ICOS, inducible T-cell COStimulator, IL-7Rα, interleukin-7 receptor alpha. KLRG1, killer cell lectin-like receptor subfamily G member 1. MAFA, mast cell function-associated antigen. c-Kit, receptor tyrosine kinase, PD-1, programmed cell death protein 1. TCR, T cell receptor.

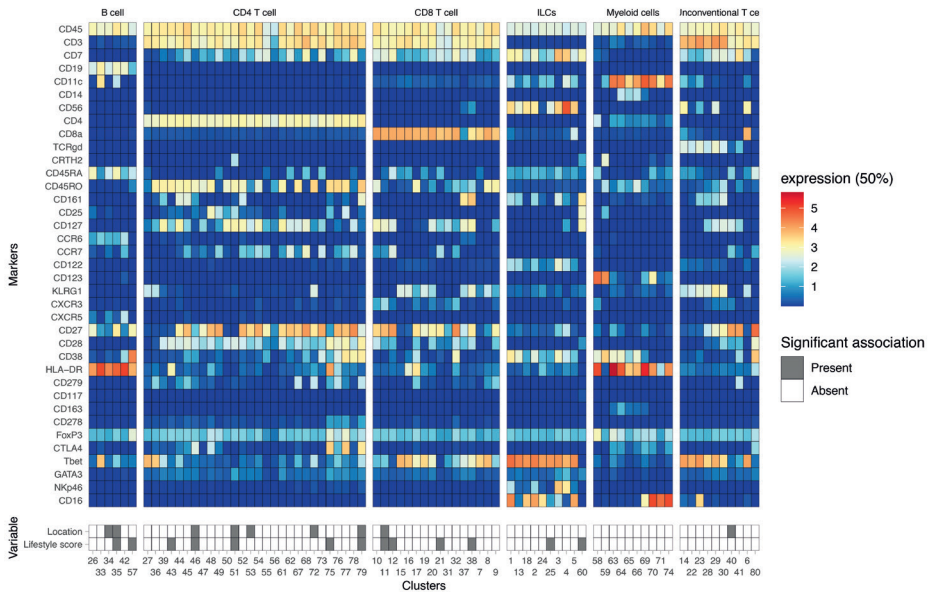

Figure S1. Heatmap showing median marker expression for each cluster. Clusters were based on SOM and hierarchical clustering. Each tile depicts the median expression of a given marker (rows) for a specific cluster (columns). The heatmap is stratified based on cell lineage. The bottom heatmap indicates which clusters were significantly associated with 1) location **(Figure 1)** and/or 2) lifestyle score **(Figure 3)**.

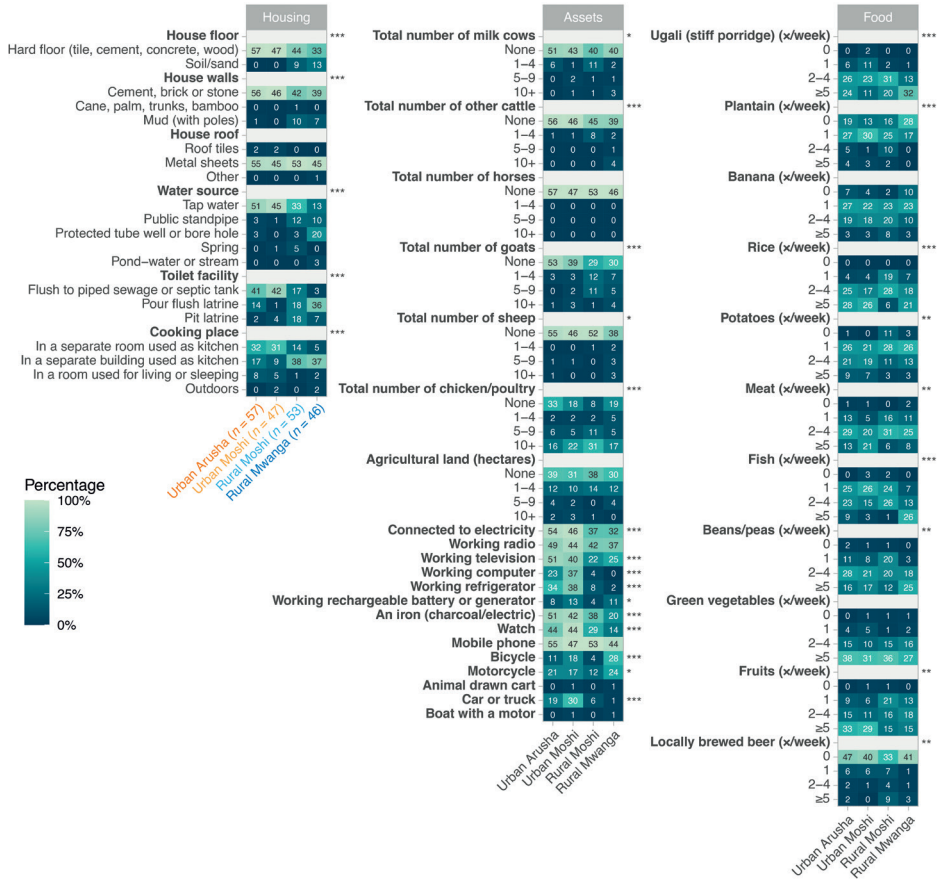

Figure S2. Heatmap visualizing lifestyle questionnaire data. *N* = 203 participants. Values represent the number of participants. Colours indicate the percentage of the total. Comparisons between locations were performed using Fisher's exact or chi-squared tests. Asterisks denote statistical significance (NS, non-significant; *, p ≤ 0.05; **, p ≤ 0.01; ***, p ≤ 0.001, p ≤ 0.0001). See **Table S3**.

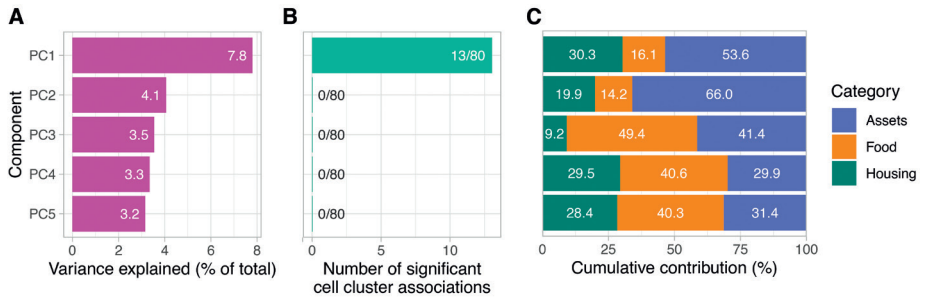

Figure S3. MCA principal component variance explained, contributions and cluster associations.
A) Variance explained (% of total) for PC1-PC5. B) Number of significant cell cluster associations with PC1 (lifestyle score) to PC5 using modelling as described in the legend of **Figure 3**. C) Cumulative contributions (in percentage) of the variable categories by questionnaire data category (i.e. housing, assets and food, *n* = 38 questions and *n* = 118 variable categories) for PC1-PC5.

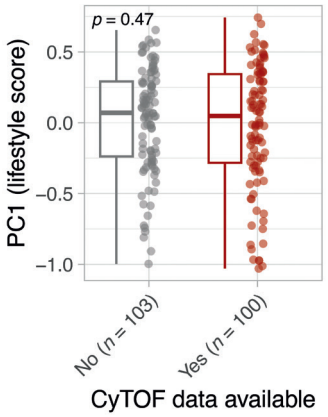

Figure S4. Boxplots showing lifestyle score for individuals with and without mass cytometry immune profiles (*n* = 100).
P-value determined using Student's t-test.

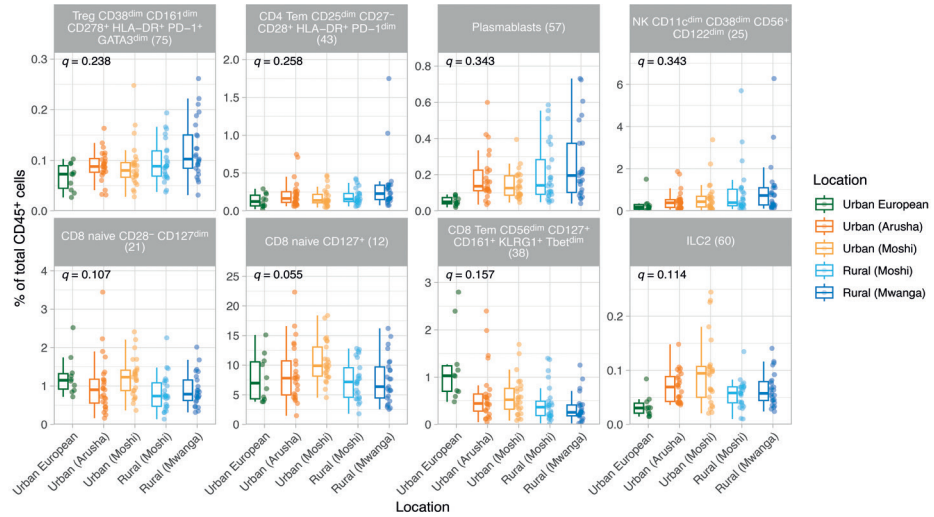

Figure S5. Cell frequencies of clusters uniquely related to lifestyle score between locations.
Cell frequencies of clusters uniquely related to lifestyle score across rural and urban Tanzanian regions and urban Europeans **(Figure 3D)**. Boxplots represent the 25th and 75th percentiles (lower and upper boundaries of boxes, respectively), the median (middle horizontal line) and measurements that fall within 1.5 times the interquartile range (IQR; distance between 25th and 75th percentiles; whiskers). Significance of 'location' was assessed using analysis of variance (ANOVA)-tests comparing a simple (age [scaled] and sex [fixed effects] and sample ID [random effect]) and a full model (simple model with location as fixed effect added). P-values were corrected for multiple testing using the Benjamini-Hochberg method and referred to as q-values. Urban Europeans were included in the figure for visual comparisons and were not included in statistical tests.

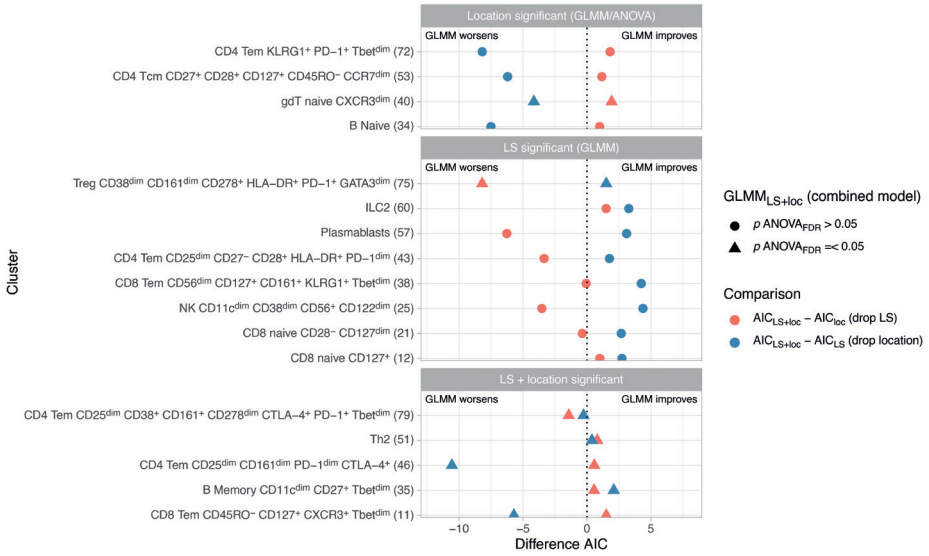

Figure S6. Sensitivity analysis comparing location- and/or lifestyle-based models.
For each of the clusters that was significant in either location- and/or lifestyle-based models (*n* = 17), we additionally fitted a joint model, including both location and lifestyle (LS) (as well as age [scaled] and sex) as fixed effects and sample ID as random effect ($GLMM_{LS+loc}$). Statistical significance of the combined effect of location and lifestyle score was assessed by comparing $GLMM_{LS+loc}$ to an 'empty model' where both location and lifestyle score were removed using ANOVA (triangles indicate significant models). Akaike Information Criterion (AIC) (measure of model fit while accounting for model complexity) was compared between the 'combined model' (AIC_{LS+loc}) and the same model from which either lifestyle score (AIC_{loc}) or location (AIC_{LS}) was removed. Clusters were grouped according to the statistics shown in **Figure 1** and **Figure 3**, i.e. location significant, LS significant or LS + location significant clusters. Dropping location or lifestyle score from the combined model for location significant and LS significant clusters, respectively, worsened the combined model, indicating that location and lifestyle score were indeed related to distinct immune cell clusters. For most of the clusters in the LS + location significant group, dropping either location or lifestyle score did not change model performance, indicating that indeed here, location and lifestyle score may be more interrelated and capture similar information.