Changes in metabolic compounds in blood predict the onset of tuberculosis

The GC6 Consortium

2016-05-24

Biomarkers of tuberculosis (TB) risk and subclinical TB progression are of paramount importance for controlling the disease. In a prospective multicohort study, serum and plasma samples were collected from household contacts of TB index cases. We have analysed metabolic profiles of samples from individuals who eventually, at a later time point, developed TB (cases) and a number of matched individuals who remained healthy (controls).

We found that several groups of metabolites are different between case and control groups, notably including amino acids (such as tryptophan and glutamine) and cortisol. Based on these findings, we developed predictive biosignatures which were then applied to a blinded validation sample set, showing the capability to predict the development of TB in high-risk individuals.

Time dependent analysis showed that the first changes allowing a discrimination between the study groups arise as early as 12-6 months before the disease is diagnosed. Our study indicates that TB progression is a slow process that may be detected early from simple serum or plasma metabolite concentrations prior to clinical TB diagnosis. In future, this will facilitate the identification of individuals at high risk of developing disease shortly after exposure or infection, allowing the development of better intervention methods.

# *Suggestions (for Stefan, Gayle, Jeroen)*

1. I have created the manuscript with one goal in mind: a complete disclosure and transparency of methods and data sets. The interested reader should be able to repeat every calculation and get precisely identical results. This has been achieved by writing the manuscript as, essentially, one large computer program. I am aware that it was a lot of time invested in a thing that only few reviewers will appreciate; however, firstly, I think that this is the right thing to do, and secondly, in the long run it will save massive amounts of time, because -- until the production stage -- everything can be quickly amended on the go.
2. The second rule that I tried to follow was maximal clarity and brevity. We have a straightforward story, so we don't want to muddle it. Please correct my English in especially to make things simpler.
3. Some sections are missing and need input from our African colleagues to be completed (in specifically, we lack additional data such as smoking status). Other sections are completed but not yet included (in especially, section in which the disease onset time is calculated). However, the main issue at hand is the logical flow of the manuscript ("the story").
4. Currently, we use "cases" and "controls". I do not particularly like it, because it always confuses people who think that the "case" group corresponds to individuals who have TB at time of sample collection. I tried to mitigate it by giving a lot of space to stressing the fact that we look at two groups of healthy individuals. Maybe we should think of a different group designation?

## Introduction

In 2014 alone, there were over globally 9 million incident cases of tuberculosis (TB) and over a million deaths (Organization and others, [2015](#ref-world2015global)). Early and sensitive diagnosis of TB is considered to be one of the main components of the global strategy to control TB (Organization and others, [2015](#ref-world2015global)). Only some infected individuals will develop active disease, but with an estimated one third of the world population being infected with Mtb, TB remains one of the major global health threats. This huge latent reservoir of TB keeps the epidemic going, where current control measures fail to eradicate this pool of latent infections or identify individuals at risk of developing active disease.

Novel TB intervention methods, especially new and improved vaccines, are expected to be the most cost-effective way to control this TB epidemic. Defining biomarkers of protective immunity or risk of disease, as well as early TB detection is crucial for the design of new vaccines to eradicate latent TB or prevent the outbreak into active disease. Identifying infected or recently exposed people with a high risk of developing active TB disease would allow stratification of these individuals into new vaccine trials. Thus drastically reducing the duration and costs of testing new vaccines.

With this in mind, the GC6-74 project was initiated in 2005 (correct?) with the goal to identify biomarkers with prognostic potential. In this effort, over XXXX individuals, who were household contacts of TB index cases were recruited in four African countries. The healthy individuals were monitored and donated blood samples at regular intervals. A number of enrolled individuals ("cases") was diagnosed with TB at a later time point. This allows to retroactively investigate the differences between these individuals and those who remained healthy ("controls") at time points predating the clinical diagnosis.

Metabolic profiling (metabonomics, MP) has been successfully applied for biomarker discovery in several diseases(Li et al., [2016](#ref-li2016blood)), including infectious diseases (Langley et al., [2013](#ref-langley2013integrated), Amaral et al. ([2013](#ref-amaral2013metabonomics)), Langley et al. ([2014](#ref-langley2014integrative))). In tuberculosis, MP of serum samples has been shown to be highly sensitive and specific for discriminating between TB patients and healthy controls (Weiner 3rd et al., [2012](#ref-weiner2012biomarkers), Frediani et al. ([2014](#ref-frediani2014plasma)), Feng et al. ([2015](#ref-feng2015analysis))). In that study, metabolites which showed a significantly different relative abundances in TB patients as compared to healthy controls included a number of amino acids (including histidine, cysteine, tryptophan and glutamine), lysophosphatidilcholines, bile acids (including glycocholate, taurocholate and derivatives) as well as immunoreactive molecules such as kynurenine and cortisol.

If MP is able to aid in discrimination of prospective TB cases from individuals who will remain healthy, then the predictive metabolites could be of one of the following, according to two hypotheses. Firstly, metabolites which remain significantly elevated in the case group throughout or at the beginning of the study may indicate risk factors. Secondly, if the difference in relative abundances is larger when the samples were collected closer to the actual TB diagnosis, then this may indicate a subclinical TB progression. In the former case, MP can aid to determine risk profile of an individual. In the latter case, changes in MP would indicate an earlier onset of TB, and could be used to detect subclinical TB before a clinical diagnosis is possible.

XXX Here a paragraph from Dan & co. on the ACS (Zak et al., [2016](#ref-zak2016blood)).

In this study, we have investigated serum and plasma samples from cases and controls in the GC6 effort; all samples were taken from individuals who at the time of sample collection were not diagnosed with TB[[1]](#footnote-1). We show that by applying MP to these samples, it is possible to predict the onset of TB in persons who could not be diagnosed with TB at the time point of sample collection. The discriminating metabolites do not form a constant signature of risk, but rather gradual rise in their abundance staring as much as a year before the clinical diagnosis.

# Results

**Table 1.** Numbers of cases and controls stratified by site and validation set.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Institution / Country | Training |  | Test |  | Total |  |
|  |  | case | control | case | control | case | control |
| **AHRI** | Armauer Hansen Research Institute AHRI, Addis Ababa, Ethiopia | 8 | 23 | 4 | 12 | 12 | 35 |
| **MAK** | Makerere University, Uganda | 7 | 23 | 4 | 16 | 11 | 39 |
| **MRC** | Medical Research Council Unit, The Gambia | 23 | 75 | 11 | 38 | 34 | 113 |
| **SUN** | Stellenbosch University, South Africa | 28 | 90 | 12 | 50 | 40 | 140 |

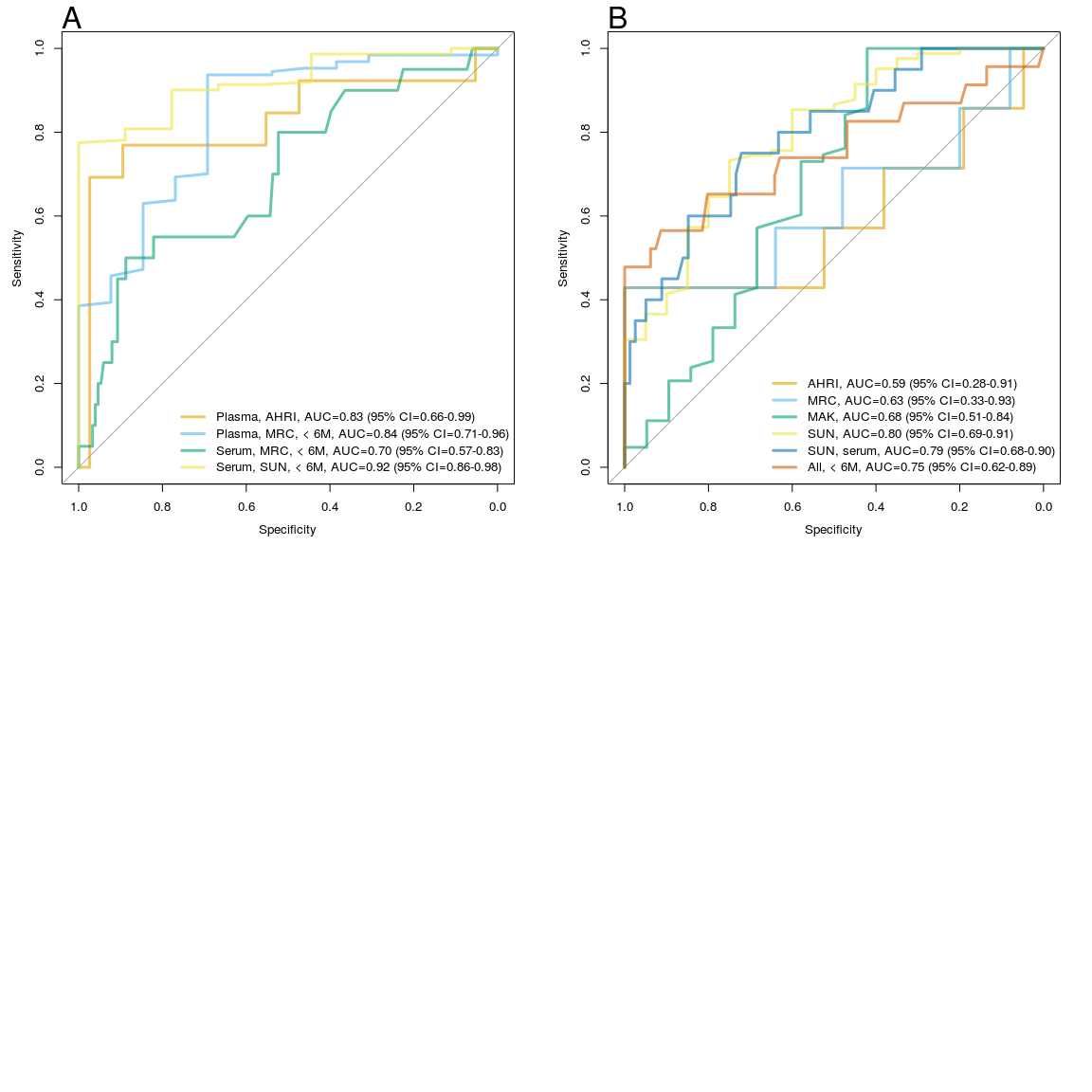
In total, 97 (XXX %) individuals who developed clinical TB and 181 samples from cases were included in this study. These samples were matched to samples from individuals who remained healthy during the follow-up examinations, resulting in the total of 751 samples analysed in this study (Table 1). For each site, two thirds of the samples were selected as a training set, and the remaining test set was blinded.

## Biosignature model building, unblinding and validation

First, we have tested the performance of the machine learning (ML) approaches within the test set only, both within and across different strata. We have created a series of random forest machine learning models (biosignatures)[[2]](#footnote-2) based on a given sample type (serum, plasma and plasma / RPMI) and sample subset (AHRI, MAK, MRC and SUN) as well as time to diagnosis (late: < 6 months; early: ≥ 6 months). We have tested the performance of these biosignatures using cross-validation within the training sets and found that all biosignatures, with the exception of the MAK cohort, performed at AUC at 0.7 or better, with samples taken at month 18 in SUN cohort achieving a performance over 0.9. In general, samples taken less than 6 months prior to the TB diagnosis were better at classification (Fig. r fig\_cv, A and Supplementary Fig. 3).

We next tested how biosignatures build on samples collected from one group (for example, SUN) validate on another group (for example, MRC). For this, we selected subgroups from test samples only and for each sub group generated a biosignature which was then tested on all remaining groups. In general, biosingatures build from late time points perform better; moreover, biosignatures applied to late time points also perform better. This is in line with the hypothesis that metabolic profiling reveals an early, subclinical TB profile.

**Figure 1.** Performance of machine learning models (biosignatures) in discriminating between cases and controls. A, receiver-operator characteristic (ROC) curves of models cross-validated within the same training set (see also Supplementary Figure 3 for all results). B, ROC curves showing the performance of the Total model on the blinded data set.

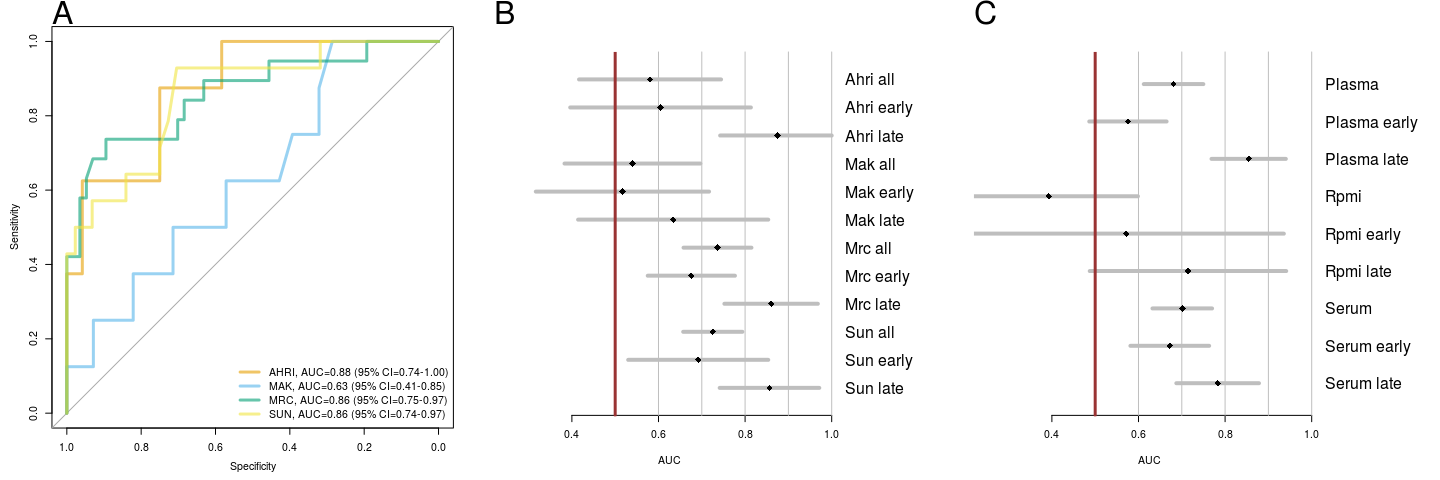


Finally, we have generated predictive models based on either all of the samples (model "Total"), selected time points ("Total BL", all baseline samples) or the different subsets. We then applied the models to the blinded test set and evaluated the results after unblinding.

In general, the performance of the models applied to the test data showed clearly the potential of the method in predicting the TB, although it was slightly lower than in the case of the training set CV alone. Again, late time points were easier to predict for most of the models (Fig. r s\_fig\_unblinding). The overall performance of the best models achieved AUC=0.8, and would ensure a 30-50% sensitivity without compromising specificity or vice versa.

We then have tested a biosignature derived from an external data set, using the previously described metabolic differences between TB patients and healthy individuals ("Metabo" model). Here, relative metabolite levels derived from serum of 136 individuals (including 44 TB patients) were used to construct a random forest model, which then was directly applied to GC6 data. We reasoned that if the signatures correspond to the advance of the disease, as suggested by the previous results -- rather than to a risk factor -- then the Metabo biosignature should show a similar performance. Indeed, that was the case (Fig. 2); the Metabo model showed a higher performance for case samples collected less than 6 months before the clinical TB diagnosis.

**Figure 2.** Predictive power of the Metabo biosignature derived from sera of TB patients and healthy individuals when applied to the GC6 data. **A**, ROC curves showing the performance of the Metabo biosignature on data sets derived from individuals less than six months prior to the diagnosis; **B**, AUC and 95% CI of the Metabo biosignature applied to complete subcohorts and subcohorts stratified by time to diagnosis; **C** AUC and 95% CI of the Metabo biosignature applied to different sample types.



XXX Here another paragraph from Seattle -- describing their model. I don't know how they would like to handle it.

## Metabolic profiles differ between cases and controls

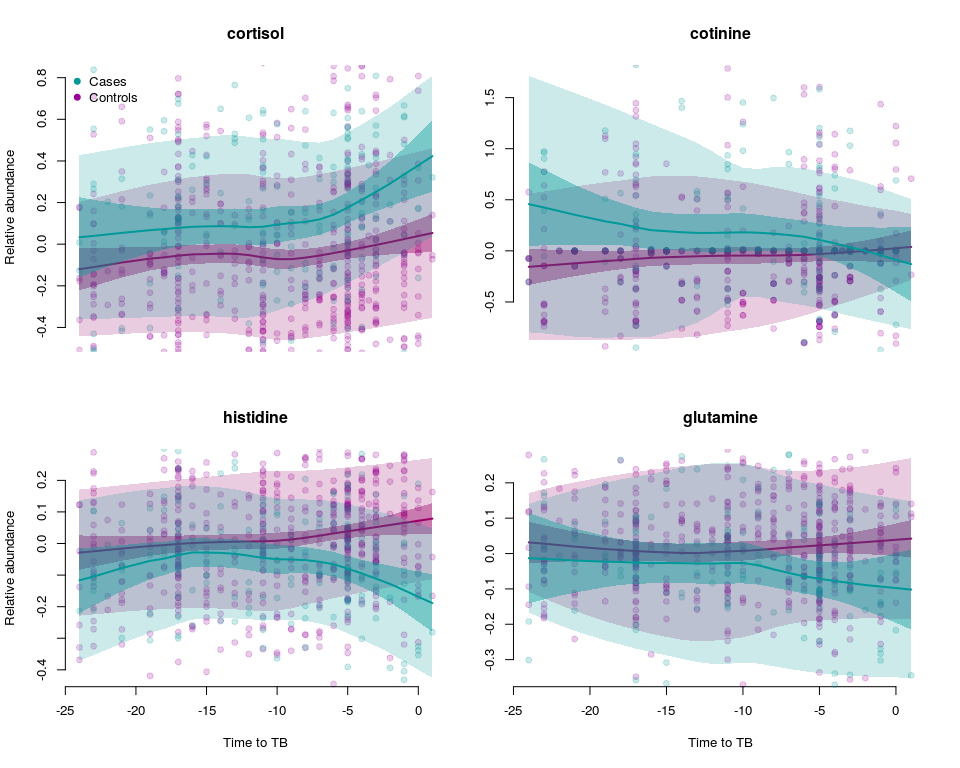
Using linear modelling, we have discerned the differences between case and control samples. We found that (depending on the model applied) between 13 and 59 compounds were significantly different between cases and controls, with the exception of plasma / RPMI samples. In total, 114 compounds were significantly different between cases and controls in at least one model. Several differences were concordant with the previously published differences between TB and healthy individuals; these included changes in several amino acids and cortisol. Unexpectedly, we did not find any statistically significant differences in the levels of kynurenine, which was one of the most prominent markers in TB vs. healthy control comparison (Table 2).

Most of the identified changes are more prominent in the months immediately leading to the disease (Fig. 3, Supplementary Fig. 1 and 2). There was one notable exception to that observation: the compound cotinine, a xenobiotic metabolite of nicotine was found at higher levels in the case group already at the baseline (Fig. 3). The presence of cotinine was correlated with patient smoking status and smoking intensity (see Supplementary Fig. 6.

**Table 2.** P-values corrected for multiple testing in the differential metabolite analysis. Table shows only the identified metabolites which achieved a q-value lower than 0.01 in any of the models tested when comparing cases to controls. Last 6M, model including only samples collected less than 6M prior to the TB diagnosis; Full, model including all samples; Serum, Plasma, Plasma / RPMI -- models including only specific sample types. Q-values smaller than 0.05 are shown in bold.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | Last 6M | Full | Plasma | Serum | Plasma / RPMI |
| M.1712 | cortisol | **0.0013** | **6.1e-06** | 0.071 | **0.0015** | 0.92 |
| M.1126 | alanine | **0.0013** | **9.9e-05** | **1.3e-05** | 0.42 | 0.97 |
| M.584 | mannose | 0.072 | **9.9e-05** | 0.093 | **0.0022** | 0.95 |
| M.553 | cotinine | 0.16 | **0.00018** | 0.63 | **0.0022** | 0.95 |
| M.43488 | N-acetylcarnosine | **0.0013** | **0.00022** | 0.13 | **0.0053** | 0.79 |
| M.22137 | homoarginine |  |  | 0.18 | **0.00027** |  |
| M.59 | histidine | **0.00048** | **0.0024** | 0.097 | 0.13 | 0.58 |
| M.12129 | beta-hydroxyisovalerate | 0.19 | **0.00051** | 0.086 | 0.08 | 0.58 |
| M.27710 | N-acetylglycine | 0.85 | **0.031** | 0.97 | **0.0018** | 0.95 |
| M.43258 | acisoga | 0.76 | 0.057 | 0.99 | **0.0022** | 0.99 |
| M.53 | glutamine | **0.0025** | **0.0048** | **0.031** | 0.27 | 0.92 |
| M.54 | tryptophan | **0.0025** | **0.012** | **0.02** | 0.69 | 0.98 |
| M.3155 | 3-ureidopropionate | **0.0028** | 0.051 | 0.68 | 0.09 | 0.95 |
| M.38661 | hydroxycotinine | 0.53 | **0.037** | 0.68 | **0.003** | 0.95 |
| M.44876 | gamma-CEHC | **0.042** | 0.1 | 0.95 | **0.003** | 0.95 |
| M.32379 | scyllo-inositol |  |  | 0.48 | **0.003** |  |
| M.33950 | N-acetylphenylalanine | **0.014** | **0.0031** | 0.47 | **0.012** | 0.58 |
| M.37202 | 4-androsten-3beta,17beta-diol disulfate (1) | 0.14 | **0.0086** | 0.73 | **0.0055** | 0.58 |
| M.42489 | 2-hydroxydecanoate | **0.006** | 0.058 | 0.29 | 0.69 | 0.48 |
| M.32675 | C-glycosyltryptophan\* | **0.008** | 0.34 | 0.86 | 0.45 | 0.73 |
| M.37063 | gamma-glutamylalanine | **0.0091** | **0.0099** | 0.14 | 0.21 | 0.65 |

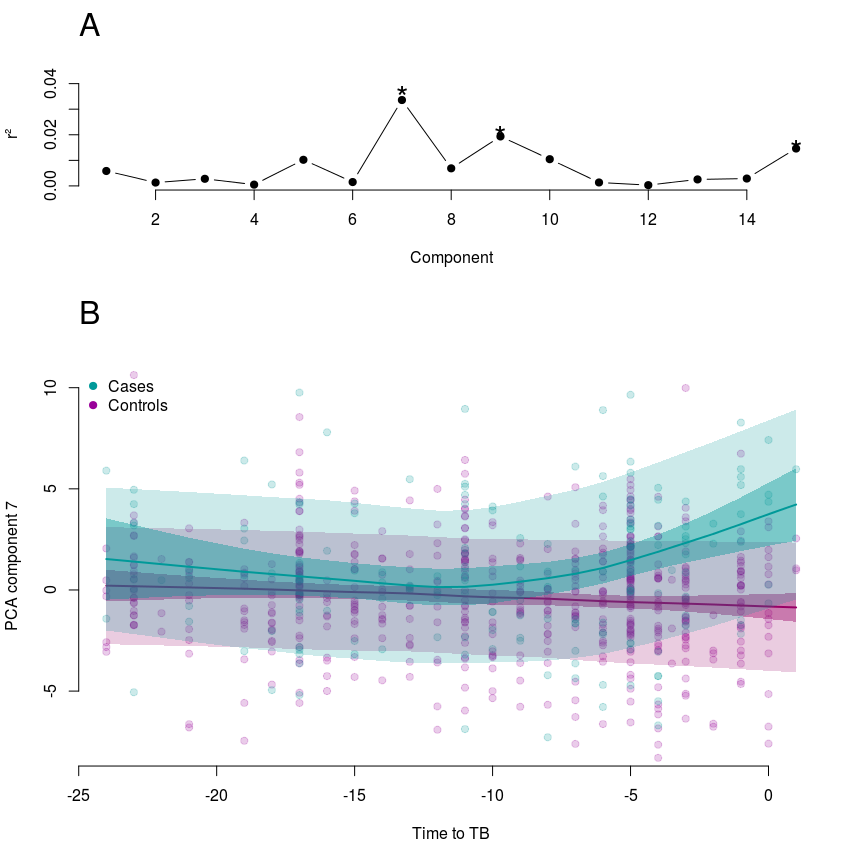
**Figure 3.** Profiles of four selected compounds and loess fits showing changes in compounds in cases and controls. Purple, cases; green, controls. Narrow, darker shades indicate 95% estimate confidence intervals, while broader, lighter shades indicate 95% prediction confidence intervals.



## Analysis of disease onset time

We next asked the question: if the changing compound concentrations in the blood of the case group individuals indicate the onset of the disease, then at what time, relative the diagnosis, a subclinical TB starts to manifestate? Firstly, for selected compounds we have fitted a segmented linear model and calculated the predicted onset time point. Secondly, we have performed a principal component analysis regression to create a compound marker.

**Figure 4.** Principal component regression analysis. A, r² for regression of the PCA components over case/control group. Stars indicate coefficients which are significant (q < 0.05). B, plot of change of PCA component 7 over time to TB.



We have found that the time of disease onset falls in a period of 12 to 6 months before the actual TB diagnosis (see fig-disease-onset)[[3]](#footnote-3). This has been confirmed by the PCA analysis, in which the first component significantly related to case/control grouping (Fig. 4, A) shows identical trend (Fig. 4, B).

# Discussion

We show that MP detects changes in the concentrations of small metabolic compounds which are predictive for developing TB. Moreover, these changes increase towards the disease onset time point and are concordant with the differences between clinical TB and healthy individuals. In fact, a model based on differences between patients and controls showed good performance in predicting the outcome. This may indicate that MP allows to monitor early, subclinical TB progression.

The number of individuals who progressed towards TB during the study (cases) was a major limiting factor in the study design. Despite a large initial cohort, less than hundred cases could be included in the study; less than expected based on the known TB incidence in these countries, especially as the individuals recruited were household contacts of TB cases. Most of these originated from two countries, South Africa and The Gambia. This explains why the SUN and MRC cohorts generally showed better performance, while the MAK (Uganda) and AHRI (Ethiopia) did not perform as well, especially in case of the test set.

We have seen that serum and plasma samples show, unexpectedly, mutually compatible profiles -- an ML model trained on one sample type can be successfully applied to another sample type. However, as sample types and cohorts were confounded, it is not possible to discern precisely the effect of cohort and sample type. Nonetheless, despite differences in samples, life style, diet etc. between the countries, we observe that a metabolic signature can work across cohorts and populations. The only group which did not follow this pattern were the samples from Uganda (MAK), as neither the serum samples nor the plasma / RPMI samples could be reliably used for prediction in that cohort. However, only two serum samples and four plasma / RPMI samples from the case group were collected less than six months prior to the disease onset, thus any conclusions based on this small sample size are not warranted.

# Methods

## Study cohorts

Samples were collected at four field sites: ... sample types ...[[4]](#footnote-4)

## Study design and sample selection

The study was planned as prospective, longitudinal, multicohort study. A third of the individuals were assigned the test group which was subsequently blinded. Data analysis was performed and revised by two independent bioinformatic teams.

To each case sample, at least three control samples were matched by age, gender and collection time point. For each site, two thirds of the samples were selected as a training set, and the remaining test set was blinded until the completion of predictive model building and creating a final prediction set (Table 1).

## Data and method availability

To ensure reproducibility of our findings, all data, scripts and software packages necessary to replicate the results and generate the figures included in this paper are provided as supplementary material. The manuscript text itself is available as a knitr (Xie, [2015](#ref-xie2015dynamic)) document.

## Machine learning

For machine learning, we used the random forest machine learning algorithm (Liaw and Wiener, [2002](#ref-liaw2002classification)), as implemented in the R package randomForest 4.6. The cross-validation within a data set was a modified leave one out procedure, in which all samples associated with an individual were removed in the cross-validation process.

## Statistical analysis

For each of the three sample types, we have removed metabolites with zero variance or without identification, and rank-normalized each metabolite, and selected only metabolites found in all three sample types. Then, we have used limma to fit a linear model which included case/control grouping, stratification grouping, time to TB and sample type, and considered the contrast of cases vs controls. We have used separate models to test the total data set for metabolites common to all three sample types, and for serum, plasma and plasma/RPMI specific metabolites.

For data standarization for visualization purposes, we have fit a model which did included all controlled variables except for the case/control grouping, and used the residuals from that model as "standardized relative abundance score".

For independent component analysis (ICA), we used fastICA in version 1.2 (Marchini et al., [2013](#ref-marchini2013fastica)).

# Supplementary tables

**Supplementary Table 1.** Sample types by cohort.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Plasma | Plasma/RPMI | Serum |
| **AHRI** | 79 | 0 | 0 |
| **MAK** | 0 | 45 | 40 |
| **MRC** | 251 | 0 | 0 |
| **SUN** | 36 | 0 | 300 |

**Supplementary Table 2.** Significant differences in metabolic profiles between cases and compounds for last collection time point (model 1).

|  |  |  |
| --- | --- | --- |
|  | BIOCHEMICAL | adj.P.Val |
| **M.59** | histidine | 0.00048 |
| **M.46500** | X - 02249 | 0.0013 |
| **M.1126** | alanine | 0.0013 |
| **M.1712** | cortisol | 0.0013 |
| **M.43488** | N-acetylcarnosine | 0.0013 |
| **M.54** | tryptophan | 0.0025 |
| **M.53** | glutamine | 0.0025 |
| **M.3155** | 3-ureidopropionate | 0.0028 |
| **M.47640** | X - 22379 | 0.003 |
| **M.42489** | 2-hydroxydecanoate | 0.006 |
| **M.32675** | C-glycosyltryptophan\* | 0.008 |
| **M.47391** | X - 22145 | 0.008 |
| **M.37063** | gamma-glutamylalanine | 0.0091 |
| **M.41377** | phenylalanyltryptophan | 0.012 |
| **M.46928** | X - 21755 | 0.012 |
| **M.35136** | 5-methyluridine (ribothymidine) | 0.012 |
| **M.46908** | X - 21739 | 0.012 |
| **M.33950** | N-acetylphenylalanine | 0.014 |
| **M.46608** | X - 11880 | 0.014 |
| **M.1302** | methionine | 0.014 |
| **M.32599** | glycocholenate sulfate\* | 0.019 |
| **M.46482** | X - 21437 | 0.019 |
| **M.1284** | threonine | 0.026 |
| **M.46997** | X - 12822 | 0.026 |
| **M.43802** | guanidinoacetate | 0.026 |
| **M.32786** | X - 11469 | 0.026 |
| **M.37207** | 4-androsten-3alpha,17alpha-diol monosulfate (2) | 0.026 |
| **M.31787** | 3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF) | 0.031 |
| **M.31591** | androsterone sulfate | 0.031 |
| **M.33935** | piperine | 0.035 |
| **M.46700** | X - 18922 | 0.041 |
| **M.46598** | X - 11438 | 0.042 |
| **M.44876** | gamma-CEHC | 0.042 |
| **M.46595** | X - 11378 | 0.045 |
| **M.46497** | X - 02269 | 0.045 |

**Supplementary Table 3.** Significant differences in metabolic profiles between cases and compounds for all collection time points (model 2).

|  |  |  |
| --- | --- | --- |
|  | BIOCHEMICAL | adj.P.Val |
| **M.47391** | X - 22145 | 3.9e-07 |
| **M.1712** | cortisol | 6.1e-06 |
| **M.1126** | alanine | 9.9e-05 |
| **M.584** | mannose | 9.9e-05 |
| **M.46608** | X - 11880 | 9.9e-05 |
| **M.553** | cotinine | 0.00018 |
| **M.43488** | N-acetylcarnosine | 0.00022 |
| **M.46515** | X - 21470 | 3e-04 |
| **M.12129** | beta-hydroxyisovalerate | 0.00051 |
| **M.46681** | X - 16935 | 0.0018 |
| **M.46626** | X - 12739 | 0.0022 |
| **M.46294** | X - 21285 | 0.0024 |
| **M.59** | histidine | 0.0024 |
| **M.47715** | X - 12681 | 0.0028 |
| **M.33950** | N-acetylphenylalanine | 0.0031 |
| **M.53** | glutamine | 0.0048 |
| **M.46595** | X - 11378 | 0.0048 |
| **M.47709** | X - 12339 | 0.0079 |
| **M.46366** | X - 18249 | 0.0086 |
| **M.37202** | 4-androsten-3beta,17beta-diol disulfate (1) | 0.0086 |
| **M.37063** | gamma-glutamylalanine | 0.0099 |
| **M.542** | 3-hydroxybutyrate (BHBA) | 0.011 |
| **M.31453** | cysteine | 0.012 |
| **M.46368** | X - 18914 | 0.012 |
| **M.54** | tryptophan | 0.012 |
| **M.1769** | cortisone | 0.013 |
| **M.32599** | glycocholenate sulfate\* | 0.014 |
| **M.18349** | indolelactate | 0.014 |
| **M.1564** | citrate | 0.014 |
| **M.46652** | X - 14658 | 0.016 |
| **M.33935** | piperine | 0.016 |
| **M.20693** | tartronate (hydroxymalonate) | 0.016 |
| **M.46390** | X - 11308 | 0.016 |
| **M.41377** | phenylalanyltryptophan | 0.016 |
| **M.2730** | gamma-glutamylglutamine | 0.016 |
| **M.46622** | X - 12792 | 0.017 |
| **M.46660** | X - 18739 | 0.017 |
| **M.32857** | X - 11540 | 0.017 |
| **M.37207** | 4-androsten-3alpha,17alpha-diol monosulfate (2) | 0.017 |
| **M.27738** | threonate | 0.018 |
| **M.44630** | 1-eicosatrienoylglycerophosphoethanolamine\* | 0.02 |
| **M.47640** | X - 22379 | 0.02 |
| **M.46471** | X - 12824 | 0.022 |
| **M.46663** | X - 11334 | 0.023 |
| **M.46928** | X - 21755 | 0.023 |
| **M.32562** | pregnen-diol disulfate\* | 0.023 |
| **M.32786** | X - 11469 | 0.025 |
| **M.46905** | X - 21736 | 0.031 |
| **M.27710** | N-acetylglycine | 0.031 |
| **M.34406** | valerylcarnitine | 0.036 |
| **M.35884** | 2-eicosatrienoylglycerophosphocholine\* | 0.036 |
| **M.46497** | X - 02269 | 0.037 |
| **M.57** | glutamate | 0.037 |
| **M.38661** | hydroxycotinine | 0.037 |
| **M.46751** | X - 12855 | 0.037 |
| **M.27447** | 1-linoleoylglycerol (1-monolinolein) | 0.037 |
| **M.37190** | 5alpha-androstan-3beta,17beta-diol disulfate | 0.04 |
| **M.46739** | laurylcarnitine\* | 0.05 |
| **M.33952** | myristoylcarnitine | 0.05 |

**Supplementary Table 4.** Significant differences in metabolic profiles between cases and compounds for all collection time points for serum metabolites.

|  |  |  |
| --- | --- | --- |
|  | BIOCHEMICAL | adj.P.Val |
| **M.47391** | X - 22145 | 2.5e-05 |
| **M.22137** | homoarginine | 0.00027 |
| **M.47964** | X - 13722 | 0.00027 |
| **M.47642** | X - 12101 | 0.00038 |
| **M.47799** | X - 15477 | 0.00069 |
| **M.1712** | cortisol | 0.0015 |
| **M.27710** | N-acetylglycine | 0.0018 |
| **M.47495** | X - 14427 | 0.0018 |
| **M.46628** | X - 12472 | 0.0018 |
| **M.46906** | X - 21737 | 0.0018 |
| **M.43258** | acisoga | 0.0022 |
| **M.584** | mannose | 0.0022 |
| **M.553** | cotinine | 0.0022 |
| **M.46608** | X - 11880 | 0.0024 |
| **M.46362** | X - 21318 | 0.0026 |
| **M.44876** | gamma-CEHC | 0.003 |
| **M.47718** | X - 12688 | 0.003 |
| **M.46515** | X - 21470 | 0.003 |
| **M.38661** | hydroxycotinine | 0.003 |
| **M.32379** | scyllo-inositol | 0.003 |
| **M.43488** | N-acetylcarnosine | 0.0053 |
| **M.37202** | 4-androsten-3beta,17beta-diol disulfate (1) | 0.0055 |
| **M.46681** | X - 16935 | 0.0055 |
| **M.47698** | X - 12122 | 0.007 |
| **M.33950** | N-acetylphenylalanine | 0.012 |
| **M.46612** | X - 12206 | 0.012 |
| **M.46626** | X - 12739 | 0.012 |
| **M.27738** | threonate | 0.012 |
| **M.47817** | X - 17327 | 0.012 |
| **M.47715** | X - 12681 | 0.013 |
| **M.46390** | X - 11308 | 0.016 |
| **M.47696** | X - 12119 | 0.02 |
| **M.33954** | glycylphenylalanine | 0.022 |
| **M.46634** | X - 12846 | 0.022 |
| **M.46624** | X - 12798 | 0.022 |
| **M.46366** | X - 18249 | 0.022 |
| **M.44552** | S-(3-hydroxypropyl)mercapturic acid (HPMA) | 0.022 |
| **M.32562** | pregnen-diol disulfate\* | 0.022 |
| **M.33963** | acetoacetate | 0.028 |
| **M.37506** | palmitoyl sphingomyelin | 0.031 |
| **M.46294** | X - 21285 | 0.032 |
| **M.31453** | cysteine | 0.033 |
| **M.47845** | X - 20674 | 0.033 |
| **M.47709** | X - 12339 | 0.033 |
| **M.41374** | phenylalanylalanine | 0.034 |
| **M.32786** | X - 11469 | 0.035 |
| **M.37190** | 5alpha-androstan-3beta,17beta-diol disulfate | 0.036 |
| **M.36103** | p-cresol sulfate | 0.037 |
| **M.46751** | X - 12855 | 0.037 |
| **M.46497** | X - 02269 | 0.037 |
| **M.46623** | X - 12729 | 0.038 |
| **M.38662** | cotinine N-oxide | 0.039 |
| **M.1769** | cortisone | 0.039 |
| **M.46368** | X - 18914 | 0.042 |
| **M.1444** | pipecolate | 0.043 |
| **M.47727** | X - 12860 | 0.044 |
| **M.47725** | X - 12802 | 0.044 |
| **M.39603** | ethyl glucuronide | 0.044 |

**Supplementary Table 5.** Significant differences in metabolic profiles between cases and compounds for all collection time points for plasma metabolites.

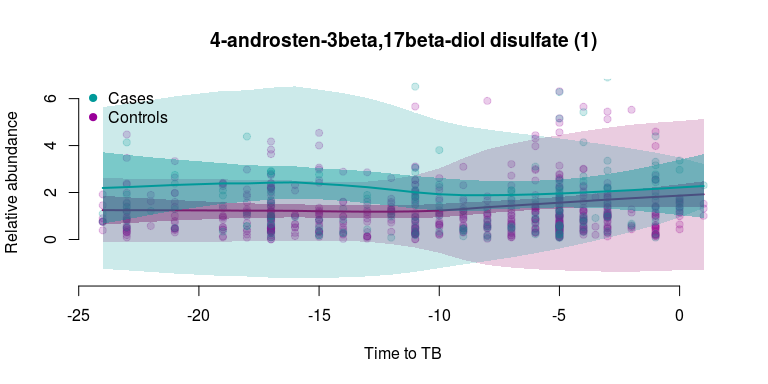
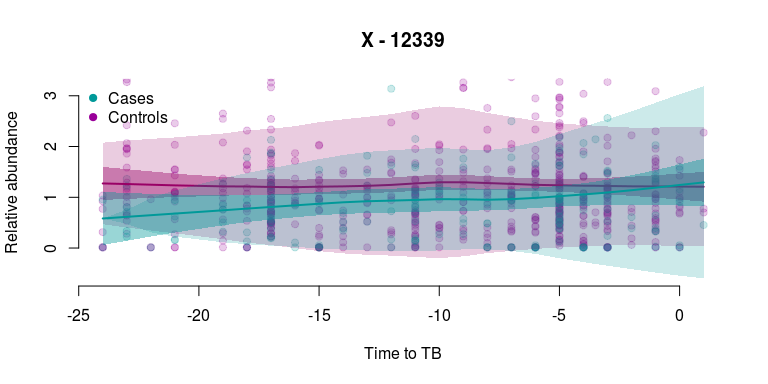
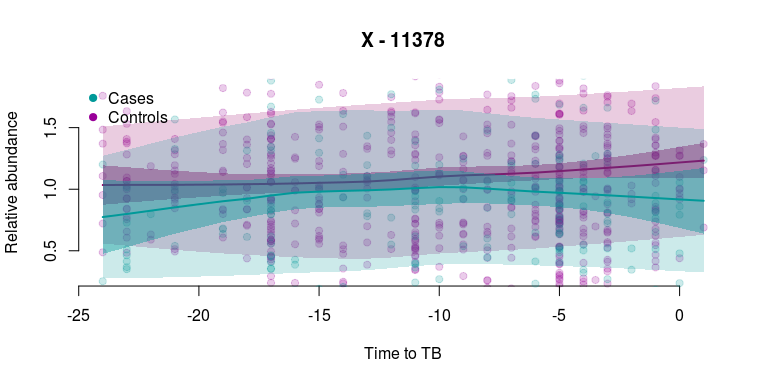
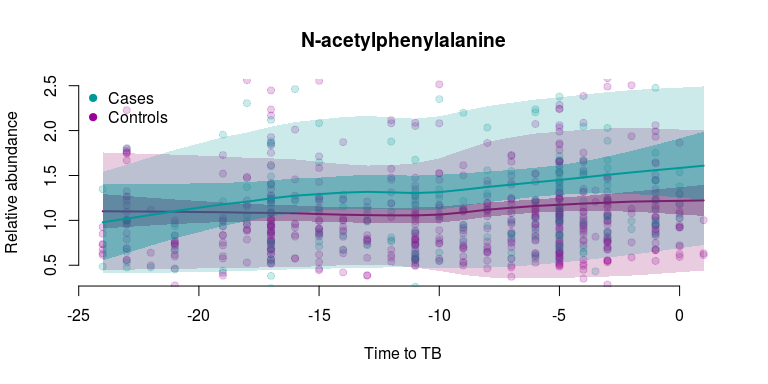
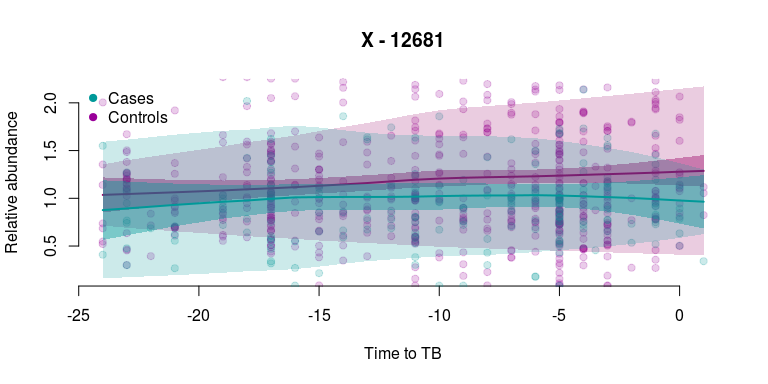
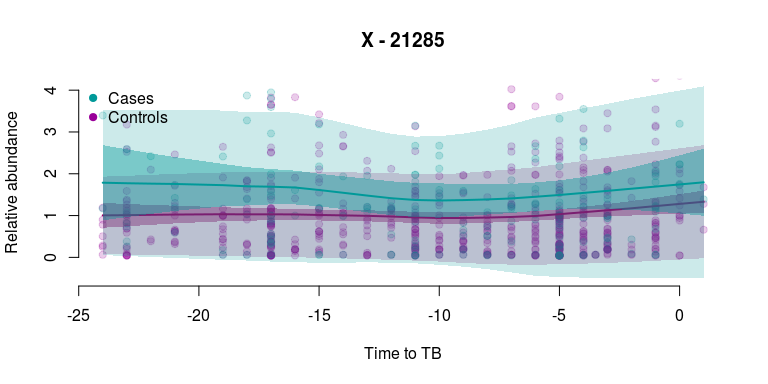
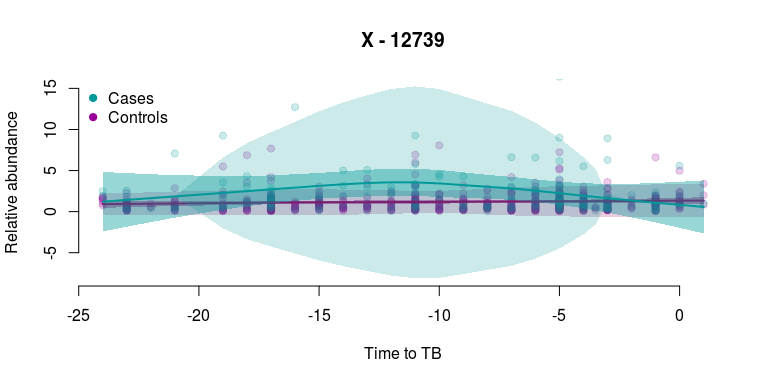
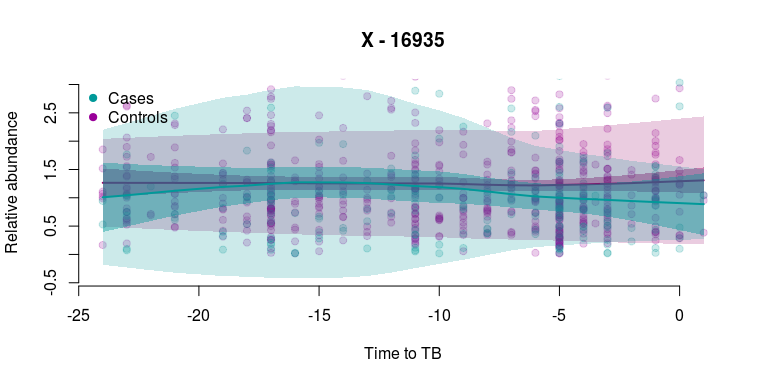
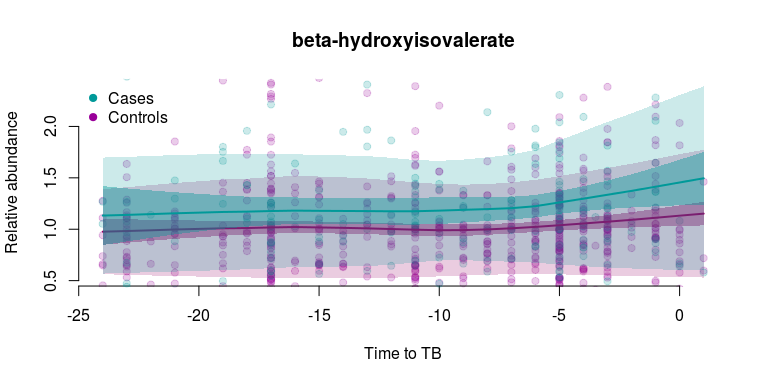
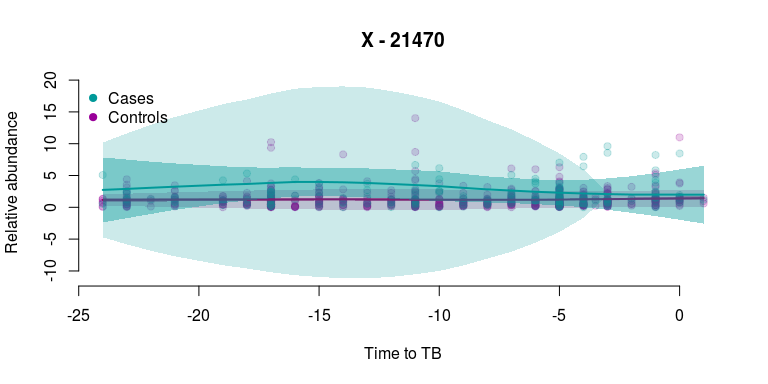
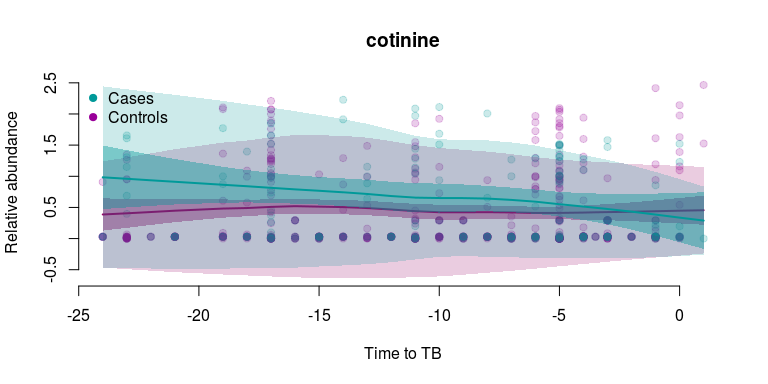
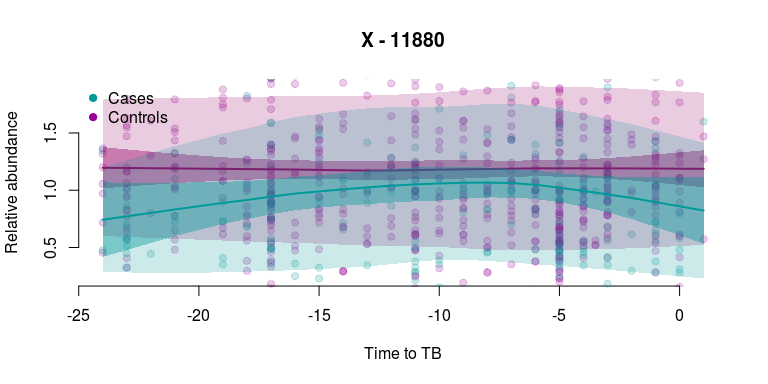
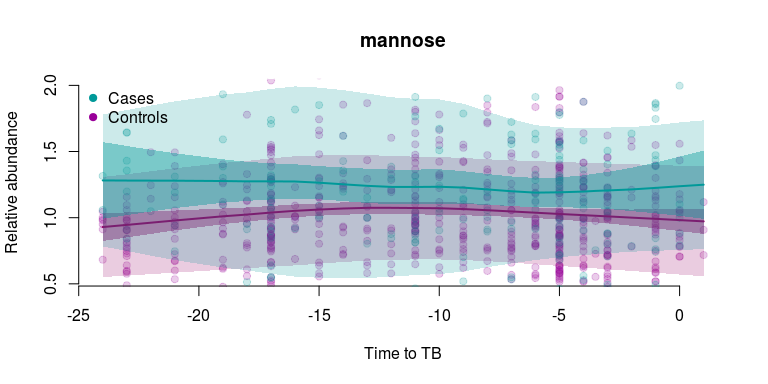
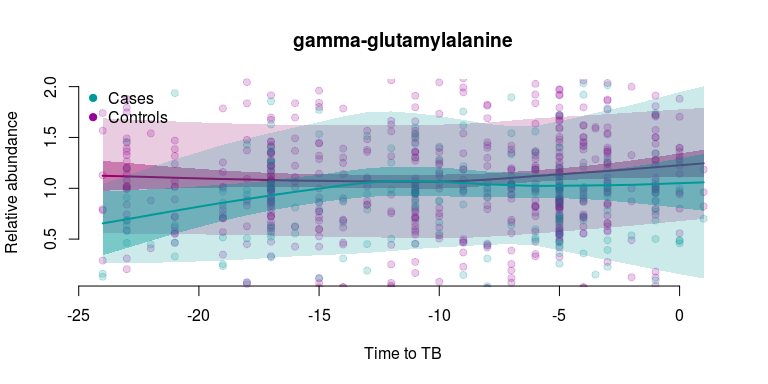
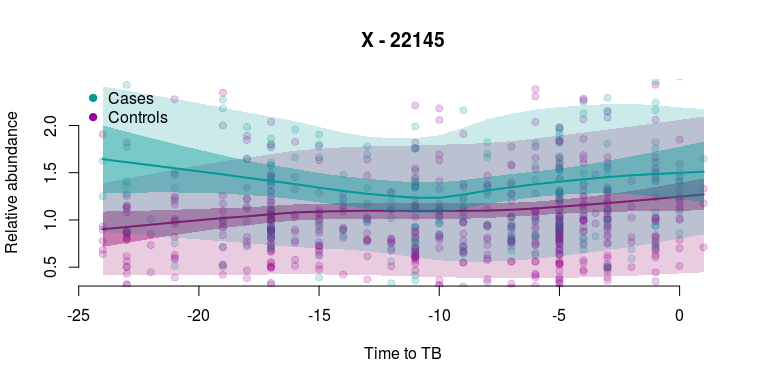
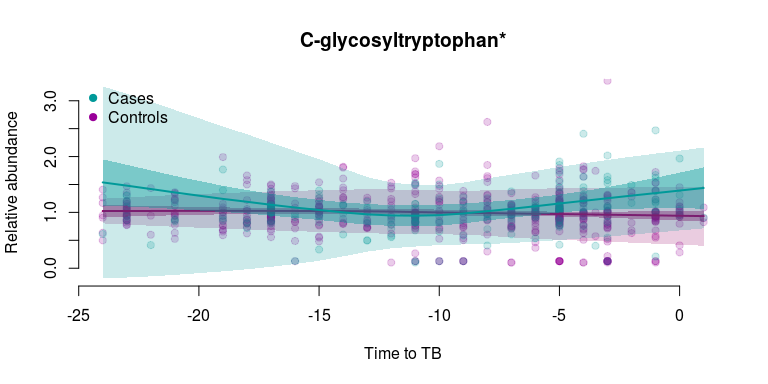
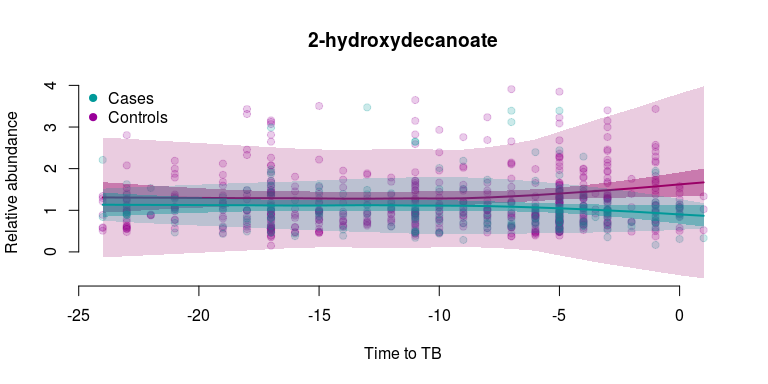
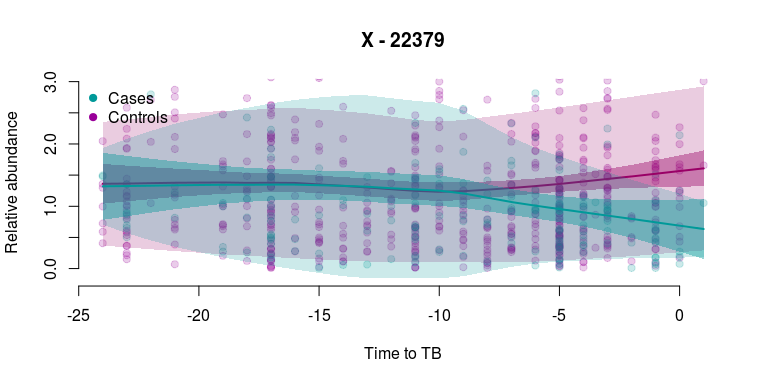
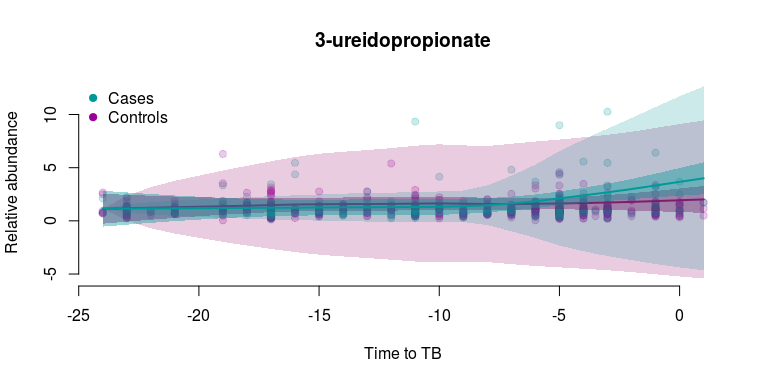
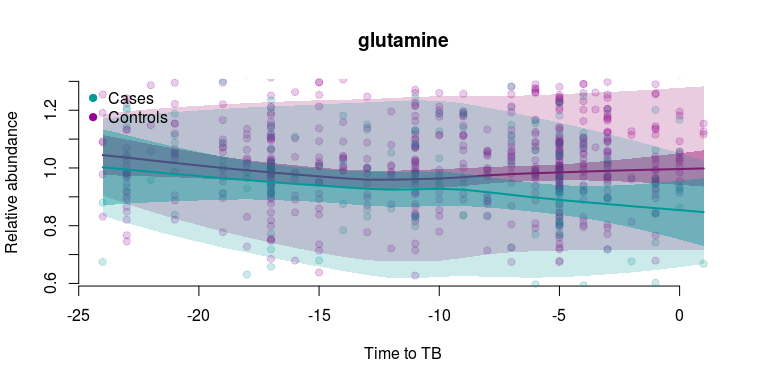
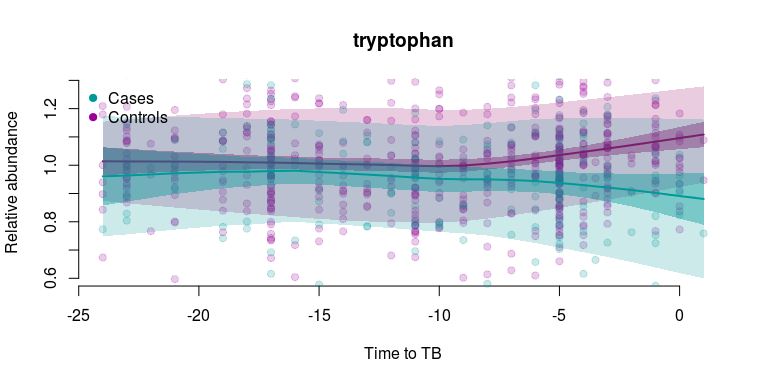
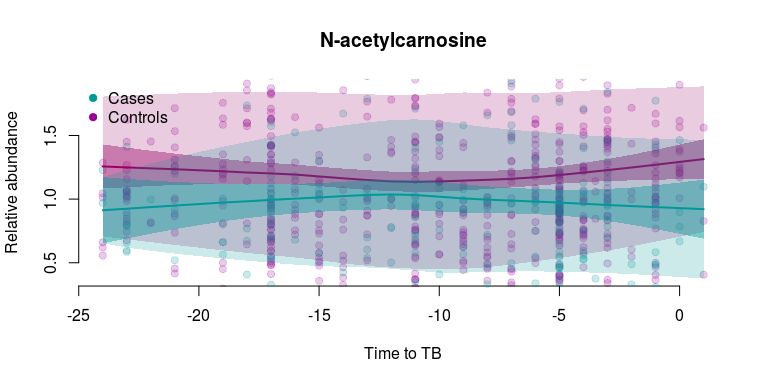
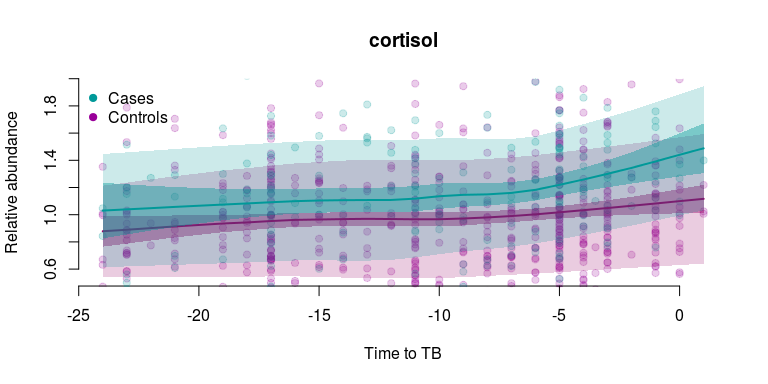
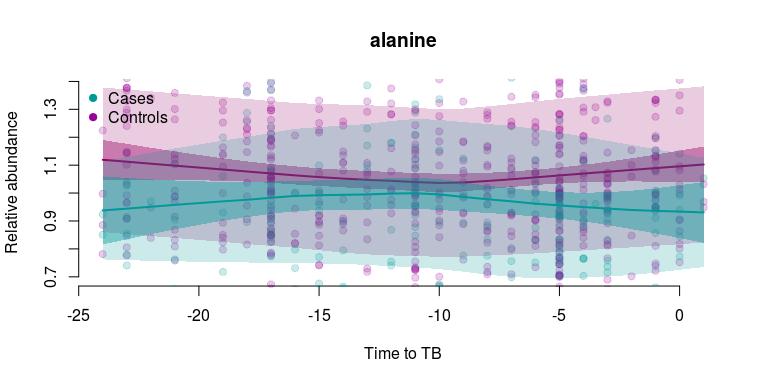
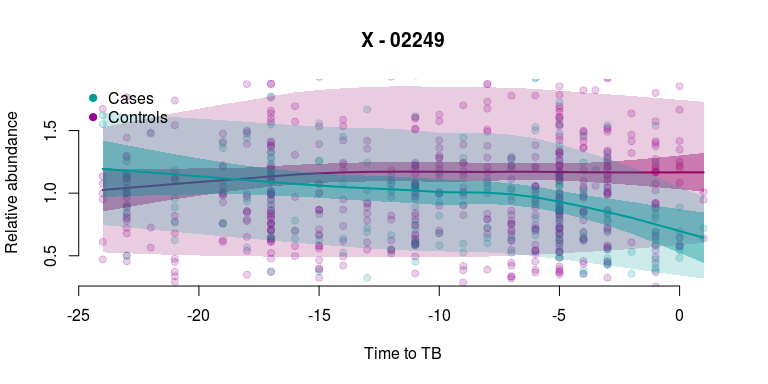
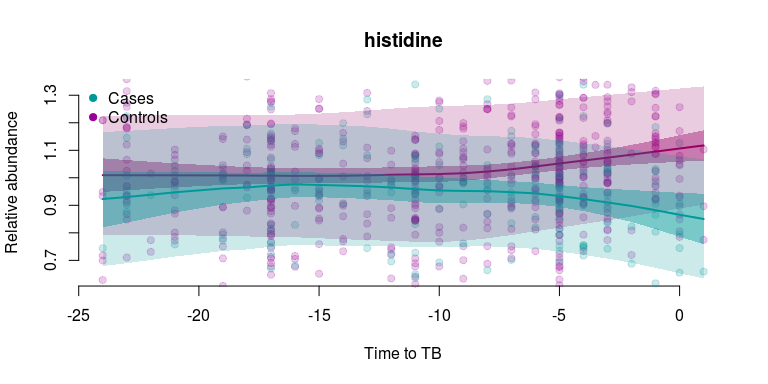
|  |  |  |
| --- | --- | --- |
|  | BIOCHEMICAL | adj.P.Val |
| **M.1126** | alanine | 1.3e-05 |
| **M.47656** | X - 12231 | 0.016 |
| **M.54** | tryptophan | 0.02 |
| **M.46726** | X - 21657 | 0.02 |
| **M.32735** | X - 01911 | 0.02 |
| **M.35625** | 1-myristoylglycerol (1-monomyristin) | 0.02 |
| **M.46600** | X - 11452 | 0.02 |
| **M.46728** | X - 21659 | 0.025 |
| **M.34393** | 1-linolenoylglycerol | 0.026 |
| **M.1592** | N-acetylneuraminate | 0.026 |
| **M.36803** | 3,7-Dihydroxy-5-cholestenoic acid | 0.028 |
| **M.53** | glutamine | 0.031 |
| **M.33935** | piperine | 0.045 |

**Supplementary Table 7.** Detailed information on the computing environment used to perform the analyses (sessionInfo() output).

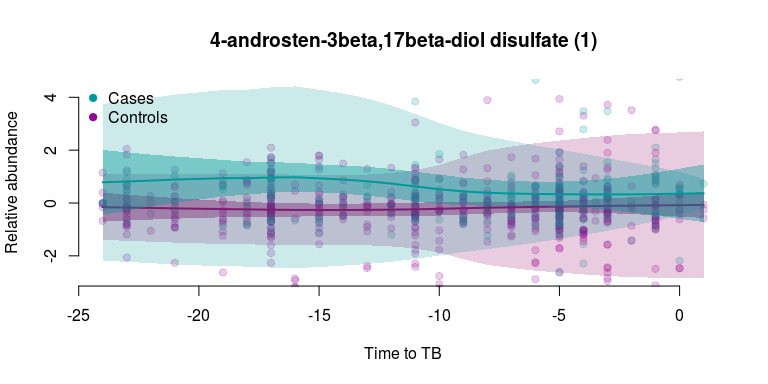
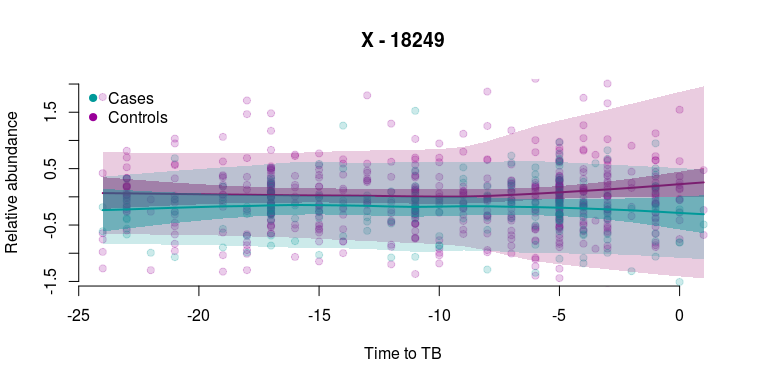
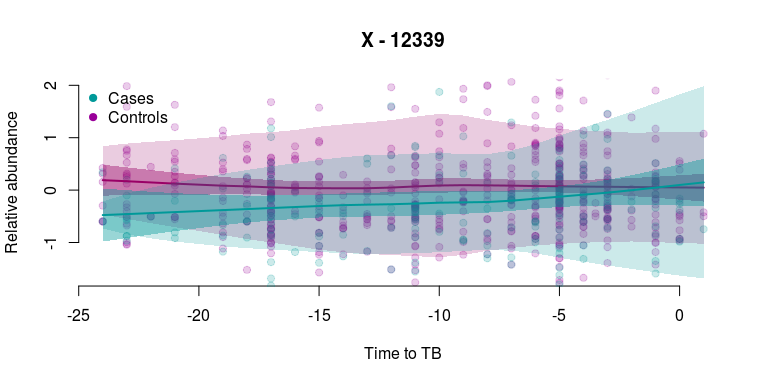
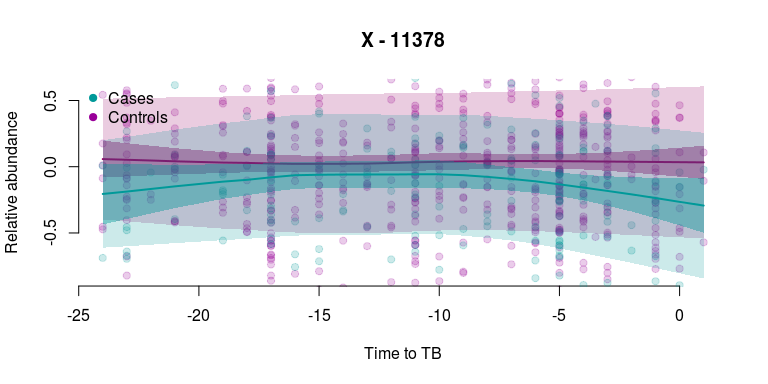
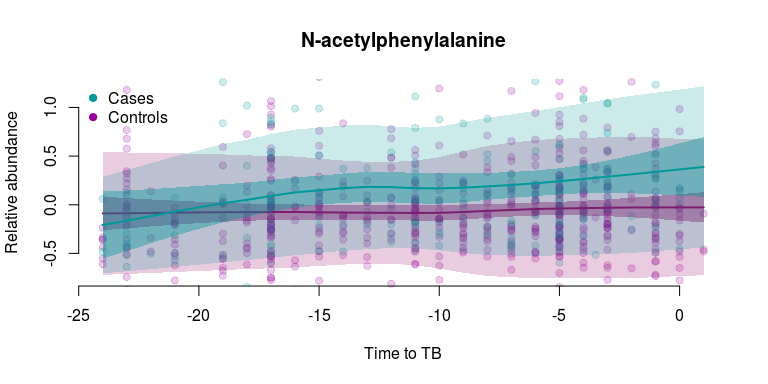
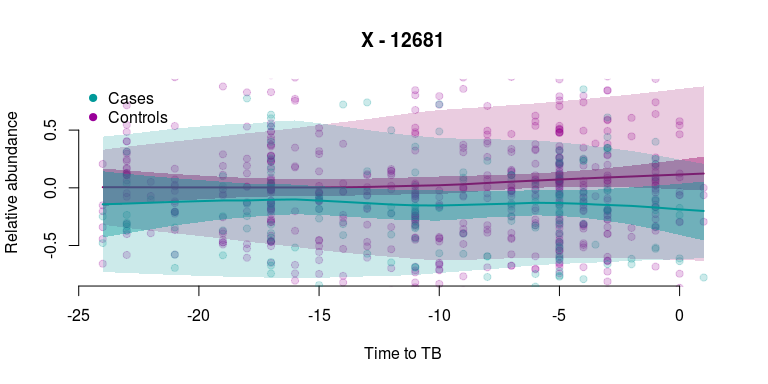
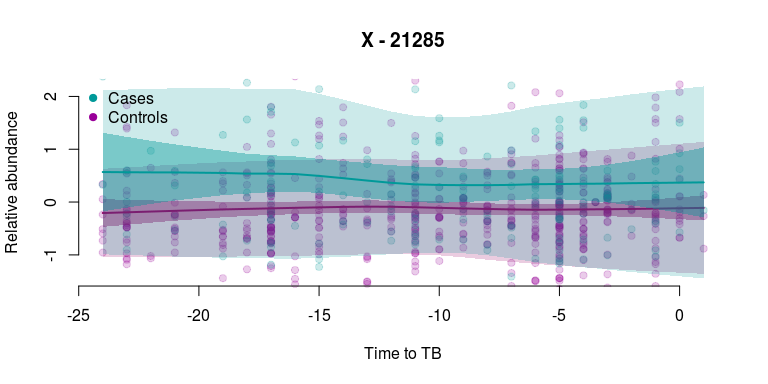
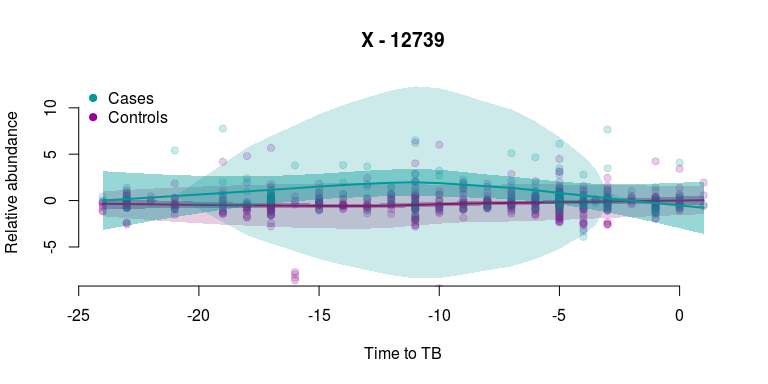
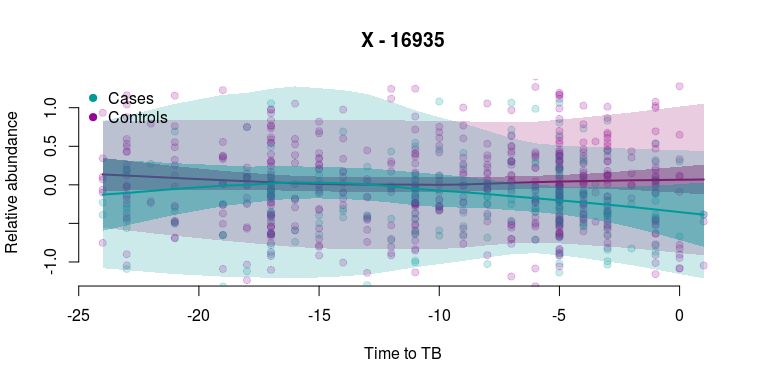
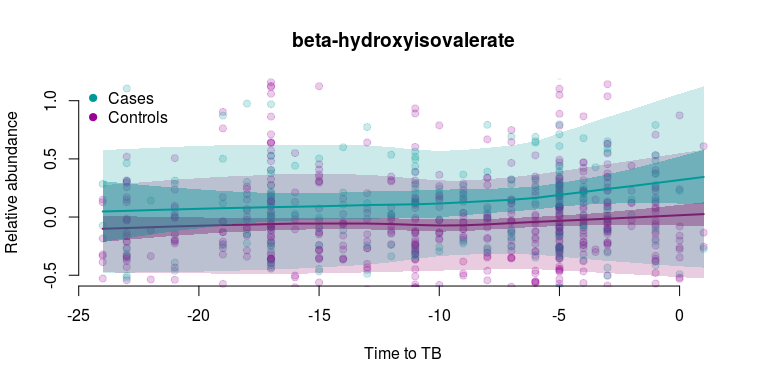
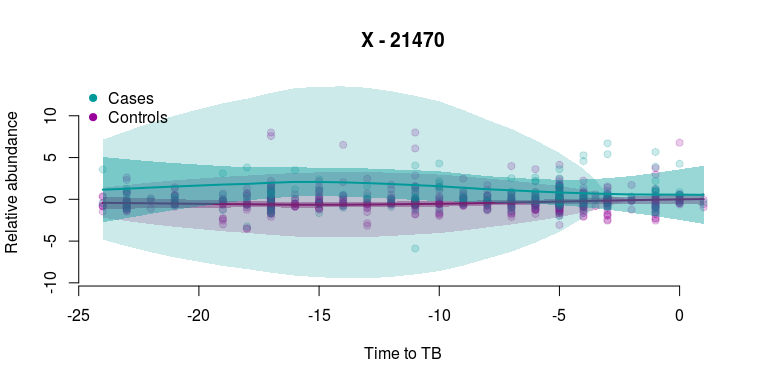
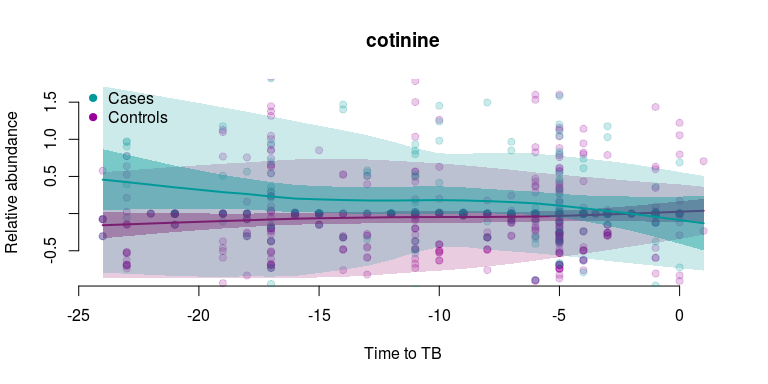
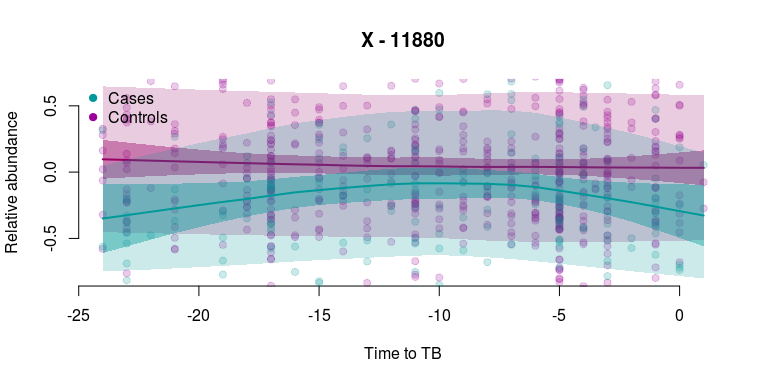
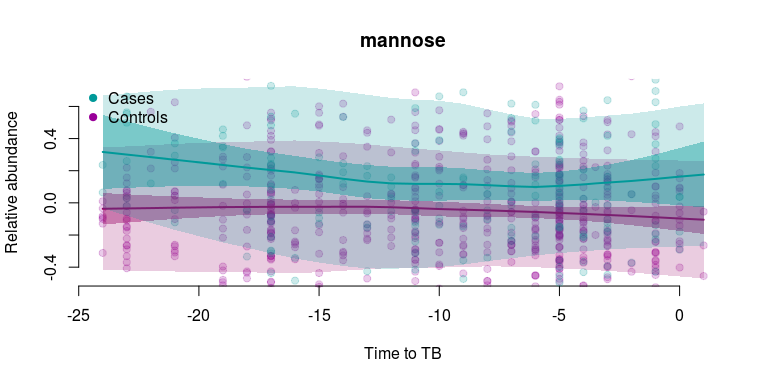
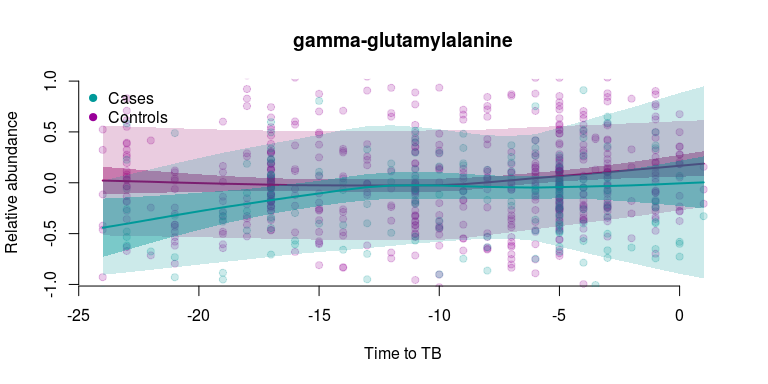
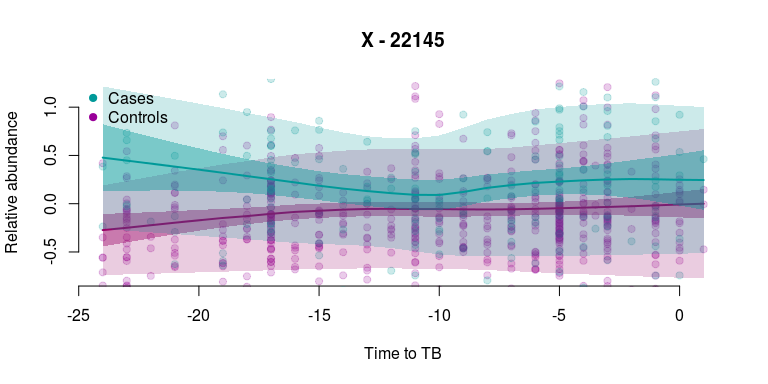
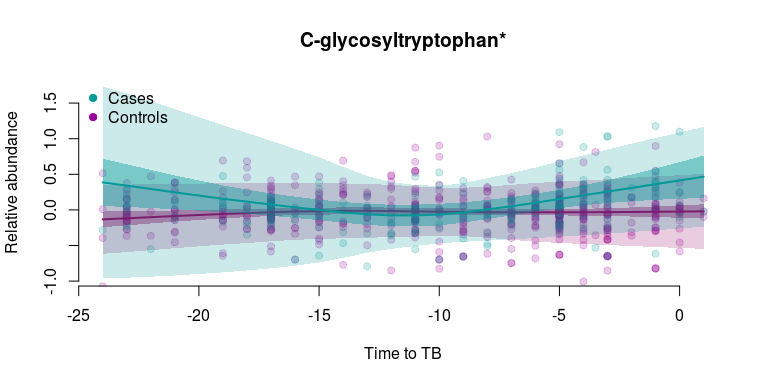
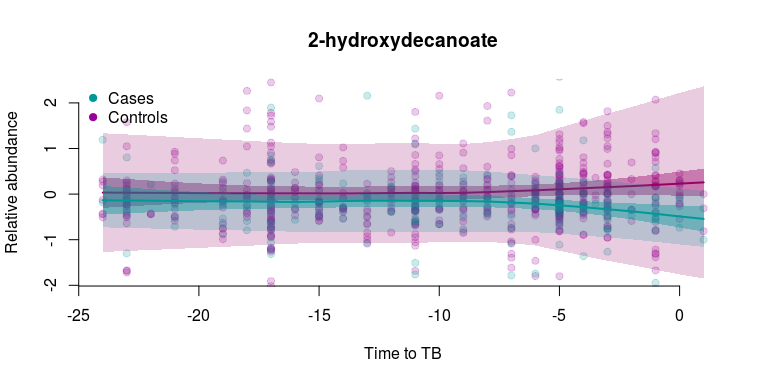
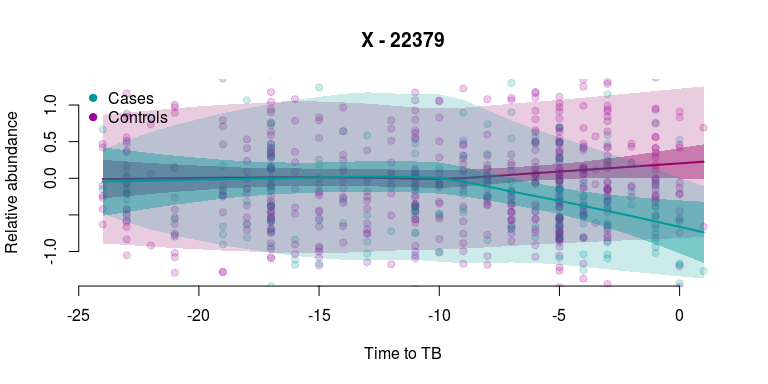
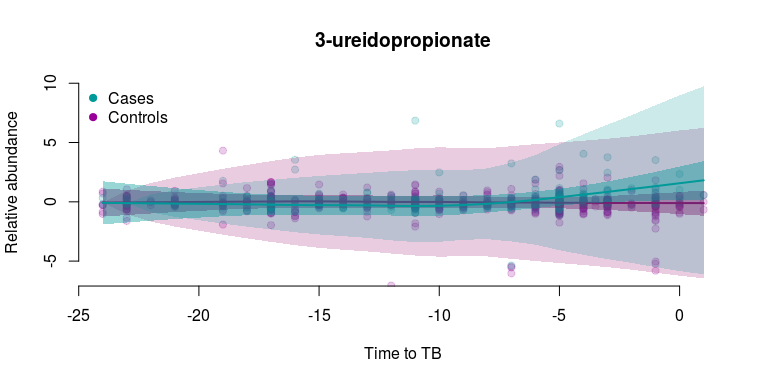
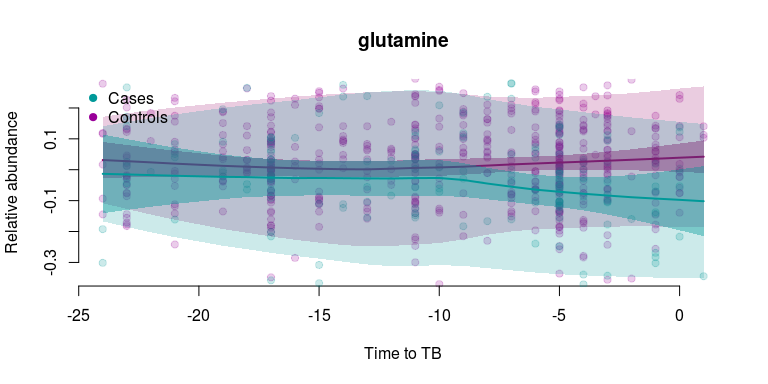
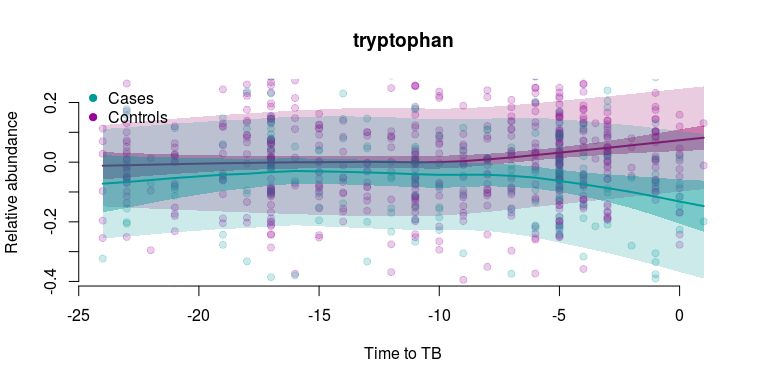
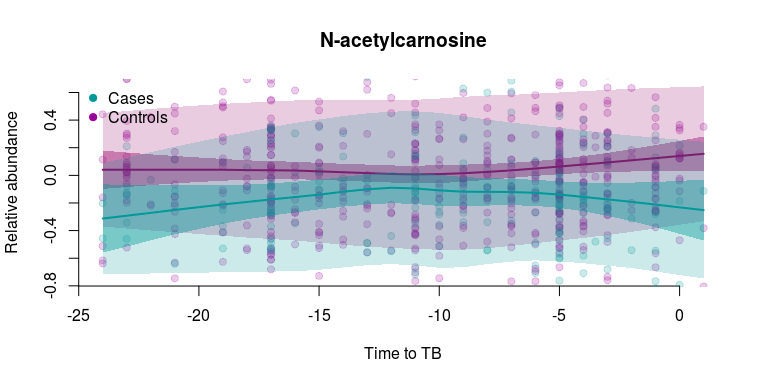
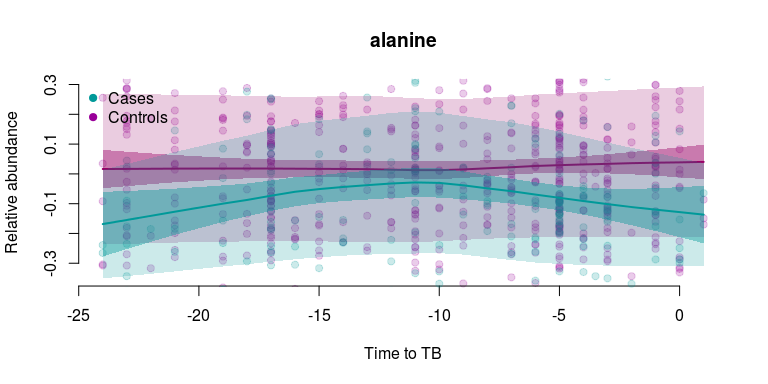
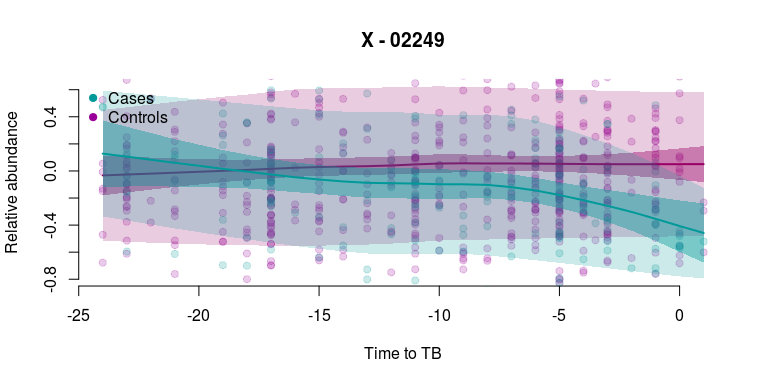
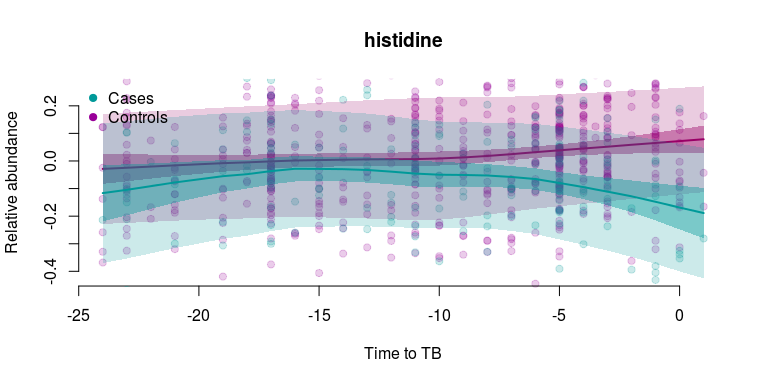
## R version 3.2.2 (2015-08-14)  
## Platform: x86\_64-pc-linux-gnu (64-bit)  
## Running under: Ubuntu 15.04  
##   
## locale:  
## [1] LC\_CTYPE=en\_US.UTF-8 LC\_NUMERIC=C   
## [3] LC\_TIME=de\_DE.UTF-8 LC\_COLLATE=en\_US.UTF-8   
## [5] LC\_MONETARY=de\_DE.UTF-8 LC\_MESSAGES=en\_US.UTF-8   
## [7] LC\_PAPER=de\_DE.UTF-8 LC\_NAME=C   
## [9] LC\_ADDRESS=C LC\_TELEPHONE=C   
## [11] LC\_MEASUREMENT=de\_DE.UTF-8 LC\_IDENTIFICATION=C   
##   
## attached base packages:  
## [1] methods stats graphics grDevices utils datasets base   
##   
## other attached packages:  
## [1] pander\_0.6.0 beeswarm\_0.2.0 tmod\_0.24   
## [4] pca3d\_0.8 randomForest\_4.6-10 myfuncs\_1.7   
## [7] limma\_3.24.13 pROC\_1.8   
##   
## loaded via a namespace (and not attached):  
## [1] rgl\_0.95.1247 Rcpp\_0.11.6 codetools\_0.2-14   
## [4] XML\_3.98-1.3 digest\_0.6.8 plyr\_1.8.3   
## [7] formatR\_1.2 magrittr\_1.5 ellipse\_0.3-8   
## [10] evaluate\_0.7 stringi\_0.5-5 rmarkdown\_0.7   
## [13] RColorBrewer\_1.1-2 tools\_3.2.2 tagcloud\_0.6   
## [16] stringr\_1.0.0 yaml\_2.1.13 htmltools\_0.3   
## [19] knitr\_1.10.5

# Supplementary figures

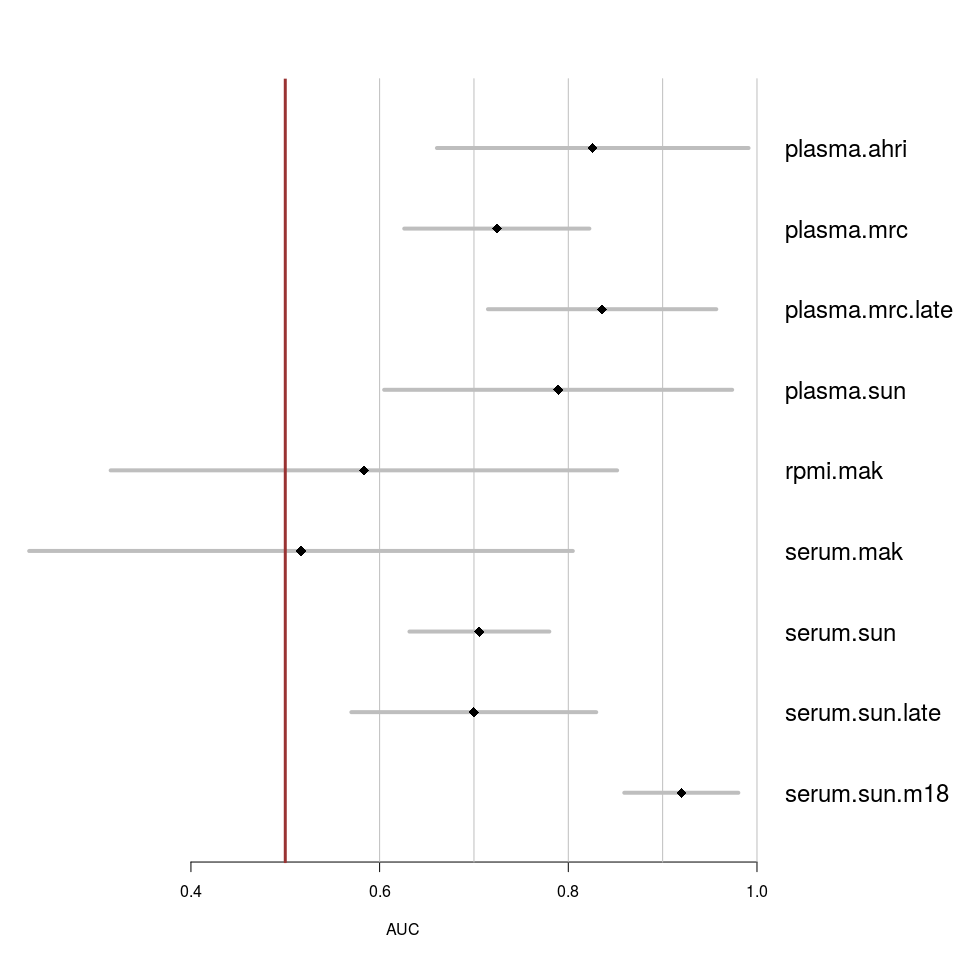
**Supplementary Figure 1.** Raw value profiles of selected compounds and loess fits showing changes in compounds in cases and controls. Purple, cases; green, controls. Narrow, darker shades indicate 95% estimate confidence intervals, while broader, lighter shades indicate 95% prediction confidence intervals.



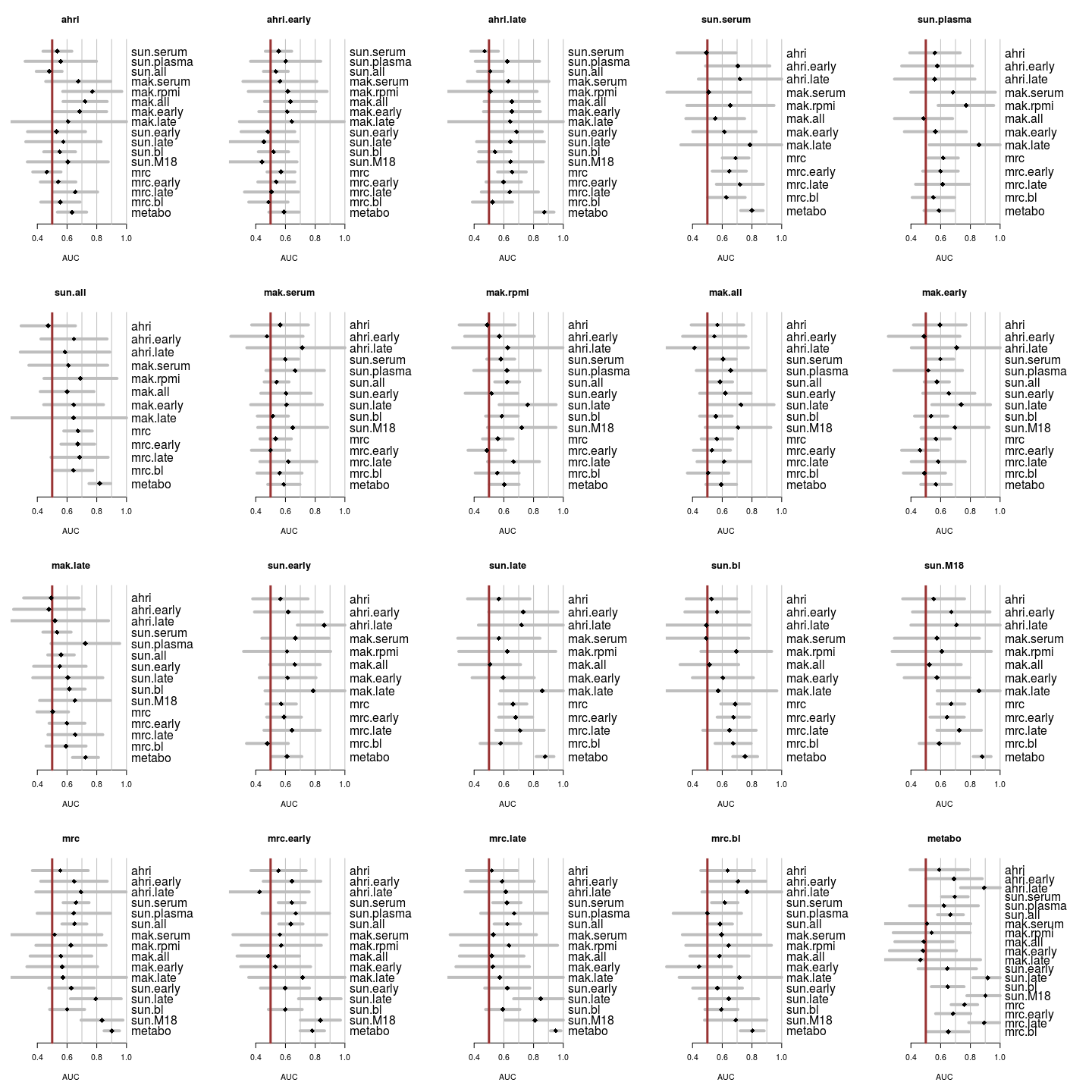
**Supplementary Figure 2.** Normalized residue profiles of selected compounds and loess fits showing changes in compounds in cases and controls. Purple, cases; green, controls. Narrow, darker shades indicate 95% estimate confidence intervals, while broader, lighter shades indicate 95% prediction confidence intervals.



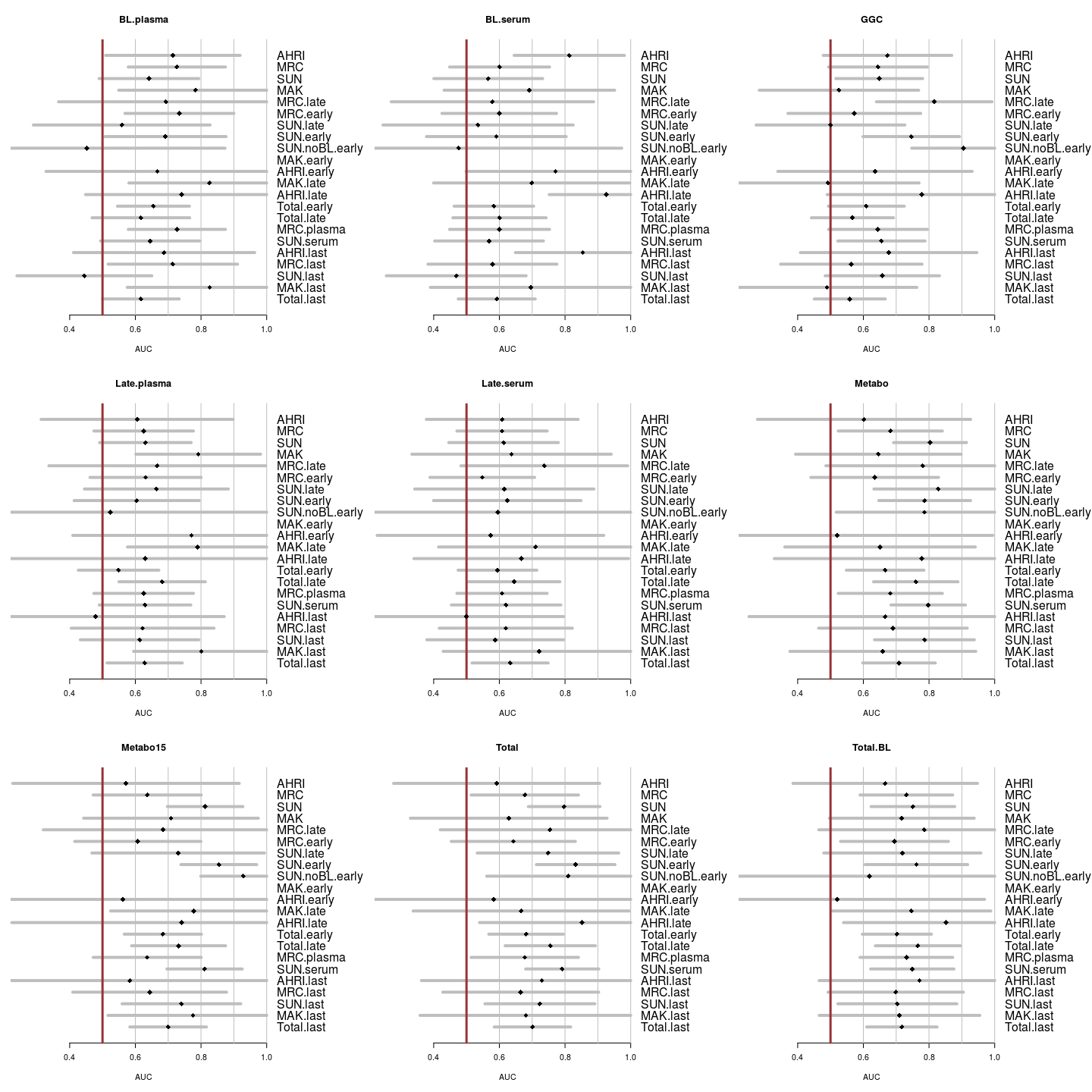
**Supplementary Figure 3.** Performance of several models cross-validated within the same training set. Dot and grey bar denote the estimated AUC and the 95% confidence interval, respectively.



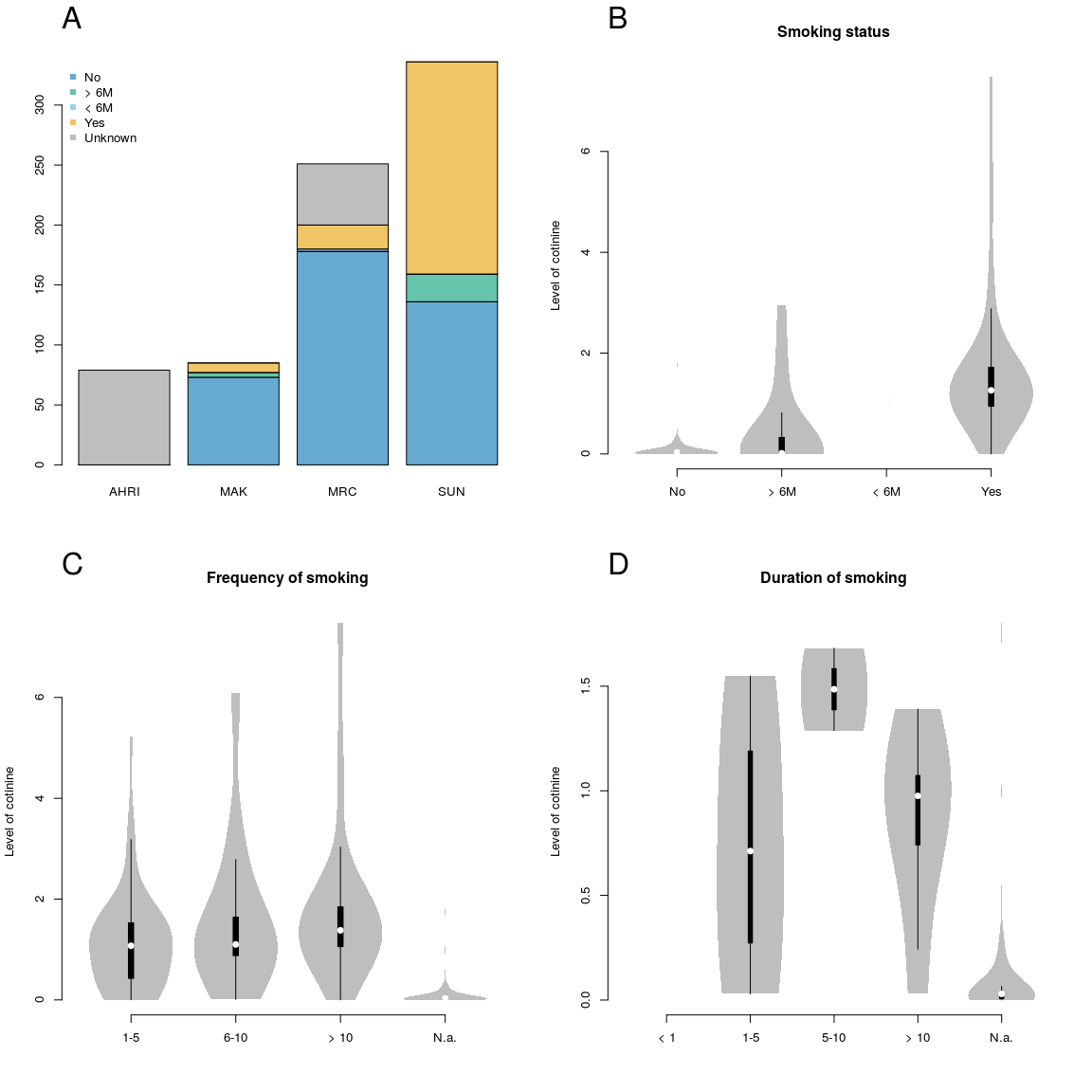
**Supplementary Figure 4.** Each panel shows the performance of a single model as applied to several training sets from other cohorts. Dot and grey bar denote the estimated AUC and the 95% confidence interval, respectively.



**Supplementary Figure 5.** Results of the blinded validation of selected models on the test data set. Each panel shows the performance of a single model as applied to several subsets of the test set. Dot and grey bar denote the estimated AUC and the 95% confidence interval, respectively.



**Supplementary Figure 6.** Smoking status and cotinine levels. *A*, number of smokers by study site; *B*, cotinine levels and smoking status (> 6M, more than six months since quitting smoking; < 6M, less than six months since quitting); *C*, cotinine levels and smoking frequency (in cigarettes per day; N.a. -- not applicable to non-smokers); *D*, cotinine levels and smoking duration (in years; N.a. -- not applicable to non-smokers).



# Bibliography

Amaral, J.J., Antunes, L.C.M., Macedo, C.S. de, Mattos, K.A., Han, J., Pan, J., Candéa, A.L., Maria das Graças, M., Ribeiro-Alves, M., Borchers, C.H., others, 2013. Metabonomics reveals drastic changes in anti-inflammatory/pro-resolving polyunsaturated fatty acids-derived lipid mediators in leprosy disease. PLoS Negl Trop Dis 7, e2381.

Feng, S., Du, Y.-Q., Zhang, L., Zhang, L., Feng, R.-R., Liu, S.-Y., others, 2015. Analysis of serum metabolic profile by ultra-performance liquid chromatography-mass spectrometry for biomarkers discovery: Application in a pilot study to discriminate patients with tuberculosis. Chinese medical journal 128, 159.

Frediani, J.K., Jones, D.P., Tukvadze, N., Uppal, K., Sanikidze, E., Kipiani, M., Tran, V.T., Hebbar, G., Walker, D.I., Kempker, R.R., others, 2014. Plasma metabolomics in human pulmonary tuberculosis disease: A pilot study. PloS one 9, e108854.

Langley, R.J., Tipper, J.L., Bruse, S., Baron, R.M., Tsalik, E.L., Huntley, J., Rogers, A.J., Jaramillo, R.J., O’Donnell, D., Mega, W.M., others, 2014. Integrative “omic” analysis of experimental bacteremia identifies a metabolic signature that distinguishes human sepsis from systemic inflammatory response syndromes. American journal of respiratory and critical care medicine 190, 445–455.

Langley, R.J., Tsalik, E.L., Van Velkinburgh, J.C., Glickman, S.W., Rice, B.J., Wang, C., Chen, B., Carin, L., Suarez, A., Mohney, R.P., others, 2013. An integrated clinico-metabolomic model improves prediction of death in sepsis. Science translational medicine 5, 195ra95–195ra95.

Li, S., Todor, A., Luo, R., 2016. Blood transcriptomics and metabolomics for personalized medicine. Computational and structural biotechnology journal 14, 1–7.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18–22.

Marchini, J.L., Heaton, C., Ripley, B.D., 2013. FastICA: FastICA algorithms to perform iCA and projection pursuit.

Organization, W.H., others, 2015. Global tuberculosis report 2015.

Weiner 3rd, J., Parida, S.K., Maertzdorf, J., Black, G.F., Repsilber, D., Telaar, A., Mohney, R.P., Arndt-Sullivan, C., Ganoza, C.A., Faé, K.C., others, 2012. Biomarkers of inflammation, immunosuppression and stress are revealed by metabolomic profiling of tuberculosis patients. PloS one 7, e40221.

Xie, Y., 2015. Dynamic documents with r and knitr. CRC Press.

Zak, D.E., Penn-Nicholson, A., Scriba, T.J., Thompson, E., Suliman, S., Amon, L.M., Mahomed, H., Erasmus, M., Whatney, W., Hussey, G.D., others, 2016. A blood rNA signature for tuberculosis disease risk: A prospective cohort study. The Lancet.

1. What about other diseases? [↑](#footnote-ref-1)
2. I am equating here biosignature with ML model to stress the fact that a biosignature is not a mere collection of variable names [↑](#footnote-ref-2)
3. More to come [↑](#footnote-ref-3)
4. Sample collection, database: we ask our colleagues from SUN to write the paragraph on this. [↑](#footnote-ref-4)