Metabolomic signatures in Elite cyclists: differential characterization of a seeming normal endocrine status regarding three serum hormones

Alain Paris[1], Boris Labrador[2], François-Xavier Lejeune[2], Cécile Canlet[3], Jérôme Molina[3,4], Michel Guinot[5], Armand Mégret[6], Michel Rieu[7], Jean-Christophe Thalabard[8], Yves Le Bouc[9]

Corresponding author: alain.paris@mnhn.fr

[1] Unité Molécules de Communication et Adaptation des Microorganismes (MCAM), Muséum national d'Histoire naturelle, CNRS, Paris, France

[2] Institut du Cerveau et de la Moelle épinière (ICM), Sorbonne Université, Inserm U 1127, CNRS UMR 7225, Hôpital Pitié Salpêtrière, Paris, France

[3] Axiom, Toxalim, INRAE, ENVT, INPT-EI Purpan, Université Paul Sabatier, Toulouse, France

[4] Dynamiques et écologie des paysages agriforestiers (DYNAFOR), INRAE, INPT-ENSAT, INPT-EI Purpan; Centre de recherche Occitanie-Toulouse, Auzeville, Castanet-Tolosan Cedex, France

[5] CHU Grenoble-Alpes, UM Sports et Pathologies, Grenoble, France. Université Grenoble-Alpes, INSERM U 1042, Hypoxia and Pathophysiology Unit, Grenoble, France. UM Sports et Pathologies, CHU Sud, Echirolles, France

[6] Fédération française de Cyclisme, 1 rue Laurent Fignon, France

[7] Agence Française de Lutte contre le Dopage (AFLD), Paris, France

[8] MAP5, Université de Paris, UMR CNRS 8145, Paris, France

[9] Sorbonne Université, INSERM, UMR S 938, Centre de Recherche Saint-Antoine (CRSA), Paris, France

Supplementary information

1. The cohort of elite cyclists

Blood samples were collected by puncture of the antecubital vein in fasting conditions between 7:30 a.m. and 9:30 a.m. They were centrifuged and the serum divided into aliquots and stored at -20°C for biochemical analysis done in a single laboratory (centralized testing) or at -80°C for achieving hormonal analyses (centralized testing) and delayed metabolomic studies. Serum cortisol was determined with the chemiluminescent assay Ready Pack Advia Centaur Cortisol (Bayer Diagnostic® France). Serum testosterone was determined with the chemiluminescent assay Ready Pack Advia Centaur Testosterone (Bayer Diagnostic® France). Serum IGF1 was determined with the RIA IBC-IGF-1 assay comprising an ethanol extraction step (Immunotech, Marseille, France).

For cortisol, normal serum concentrations were comprised between 90 and 220 ng/mL, values which corresponded in the present cohort to the respective quantiles: 0.105 and 0.561. For IGF1, the normal serum concentrations were comprised between 200 and 420 ng/mL, values which corresponded in the present cohort to the respective quantiles: 0.116 and 0.832. For testosterone, the normal serum concentrations were comprised between 3.0 and 8.3 ng/mL, values which corresponded in the present cohort to the respective quantiles: 0.123 and 0.867.

Minimal, median and maximal ages of subjects were 19.9, 24.3 and 41.0 yrs. (Table S1). For the subgroup of subjects who were submitted to a longitudinal follow-up and were repeatedly collected, their median minimal and maximal ages at blood collection time were estimated to 24.4 and 27.6 yrs., respectively, when their median age was calculated to 24.8 yrs. (Table S1).

For the cortisol phenotype, the *Low* class contained only 66 samples, when the *Normal* and the *High* classes contained 303 and 277 samples, respectively (Table S2). For IGF1 and testosterone, the size of the *Normal* class was much higher (respectively, 491 and 500 samples) than sizes for *Low* and *High* classes of these two endocrine phenotypes. Nevertheless, these two latter classes were more or less balanced with for IGF1 respectively 71 and 84 samples in these two abnormal classes and for testosterone 78 and 57 samples in these same classes (Table S2). Few missing concentration values were found in these endocrine phenotypes: 9 for cortisol, 9 for IGF1 and 20 for testosterone (Table S2). Non-hormonally phenotyped (*NP*) samples were discarded from the cohort for the given endocrine trait.

A special attention was given to the constitution of the control classes for cortisol, IGF1 and testosterone for which the normality of endocrine and haematological profiles of subjects has been detected simultaneously for the three hormones assayed one time for 67 subjects on a total of 90 fully hormonally normal subjects and twice for 17 additional subjects (Table S3). These sets of fully hormonally normal samples represented 18.6% of the total biobank of the study. Samples of some numerous disrupted endocrine phenotypes such as the *High* cortisol class or the *High* testosterone one were randomly selected from the FFC biobank. By contrast, all samples belonging to rare disrupted endocrine phenotypes were included in the study.

In a first exploration of the variation of hormonal concentrations with age in *Normal* classes, no significant dependency was detected between 20 and 34 yrs., excepted for IGF1 where a significant decrease was revealed (Figure S1). This observation was in agreement with prior studies (Brabant et al. 2003; Bidlingmaier et al. 2014; Elmlinger et al. 2004; Landin-Wilhelmsen et al. 2004; Rosario 2010). However in practice, in this latter case, no need for a prior correction of hormonal range to

define an age-corrected normality of IGF1 values was necessary given the large range of normal IGF1 values when assayed in such a restricted age range [20-34 yrs.].

2. Materials and Methods

2.1. ^{1}H NMR fingerprinting

Fingerprinting was performed by ^{1}H NMR spectroscopy after a rapid sample preparation performed as following. D_2O (500 µl) was added to serum (100 µl). After mixing the sample, it was transferred to a 5-mm NMR vial for ^{1}H NMR analysis.

Because serum contains heavy molecules, which display a large signal and hide narrow metabolite signals, the Carr-Purcell-Meiboom-Gill (CPMG) NMR sequence was used to discard this type of interfering signal.

The ^{1}H signal was acquired by accumulating 256 transients with a 12-ppm spectral width typically collecting 32 K data points. A spin-spin relaxation total delay of 43.9×10^{-3} s and a 3-s relaxation delay were applied. The Fourier transform (FT) was calculated on 64 K points. The total experiment was run for 23 min by sample. All ^{1}H NMR spectra were phased and the baseline corrected. A representative spectrum is given in Figure S2. The ^{1}H chemical shifts were calibrated on the resonance of lactate at 1.33 ppm. Then, serum spectra were data-reduced prior to statistical analysis using AMIX software (Analysis of Mixtures v 3.6.4) from Bruker Analytische Messtechnik (Karlsruhe, Germany). The spectral region δ 0.5-10.0 ppm was segmented into consecutive non-overlapping regions (buckets) of 0.01 ppm and normalized against the total signal intensity. The region around δ 4.8 ppm corresponding to the water resonance was excluded from the pattern recognition analysis to eliminate residual water artifacts.

2.2. Statistical analyses

2.2.1. Pre-processing and non-supervised data explorations

MAGIC (Markov Affinity-based Graph Imputation of Cells) processing was used to prior remove noisy information while highlighting relevant data structures, more particularly complex and non-linear interactions between sets of variables (van Dijk et al. 2018). This has resulted in a more contrasted network of variable correlations without requiring any dummy variables. PHATE can generate a low-dimensional embedding easily visualized on a 2D factorial map, and provides an accurate, denoised information which represents both the local and global extracted informational structures of a dataset (Moon et al. 2019). In most cases, parameters used for processing with MAGIC and PHATE were left as their default values.

2.2.2. Discrimination procedures

Metabolites were given in a 655 × 419 spectral matrix *X* for the whole cohort, except the NP subjects for their respective endocrine phenotype. For the three classes, a numeric coding factor (either 1, 2 or 3) represented the class of concentration of either cortisol, IGF1 or testosterone assayed respectively in the *Low*, *Normal* or *High* class. These coding values constituted the *Y* response class vector.

Basically, for any endocrine phenotype, the classification of any unknown subject from his metabolomic profile into one of the three classes could be computed according to a prior discrimination rule established from a learning (training) set of documented samples. In this aim, we

compared the following two approaches for their ability to handle the problem of multicollinearity, which typically arose when the explanatory variables are numerous.

First, partial least squares-discriminant analysis (PLS-DA), a classification method based on partial least squares regression (PLSR) (Barker & Rayens 2003), was used. Besides, this method was used here either with or without a prior Orthogonal Signal Correction (OSC) to remove on *X* the variation on a component orthogonal to the *Y* response vector. This unrelated information was filtered out using PLSR. OSC of the matrix *X* combined the nonlinear iterative partial least squares (NIPALS) algorithm (Wold et al. 1998) with that of Wise and Gallagher published by Westerhuis et al. (2001) (Matlab code available at https://eigenvector.com/resources/matlab-user-area/orthogonal-signal-correction/).

Second, a shrinkage discriminant analysis (SDA) (Ahdesmäki & Strimmer 2010) was used to get a discrimination rule from the training set of samples in order to apply it to samples from unknown classes. SDA trains a linear discriminant analysis classifier using James–Stein shrinkage estimates of correlations, variances and priors while predictor variables are ranked using correlation-adjusted t-scores (cat scores) (Zuber & Strimmer 2009), which also includes a particular feature (variable) selection to get a more homogenous variance between all variables. Every endocrine phenotype regarding the cortisol, IGF1 and testosterone traits was modelled independently from the two other phenotypes. All these statistical procedures were performed using R 3.6.0, the R package sda (Ahdesmäki et al. 2015) and home developed code for PLSR and OSC. R code is available on request to authors. The final algorithm was then defined from a preliminary comparison of PLS-DA and SDA classification performances.

2.2.3. Validation procedure

Further validation of the classification process was done considering i) the number NV of ordered and selected variables according to their cat score and ii) the number LS of statistical individuals (samples) randomly selected in the training dataset.

Classification performance according to the cohort size and the number of variables

To assess the performance of the classifier according to parameters LS and NV, the following procedure was repeated separately for cortisol, IGF1 and testosterone. In this aim, the classifier was evaluated on a total of 104 combinations (13 × 8) across the two sets of candidate values ranging respectively from 50 to 650 with a 50-step for statistical individuals, and ranging from 50 to 400 with a 50-step for variables (Figure 1, Block 3). Therefore, for every dimension combination, the discrimination rules were tested using a bootstrap approach involving 1000 replicates of the initial dataset. Variables were added 50 by 50 according to a prior ranking of the median cat score calculated thanks to a bootstrap (n = 1000) of the function *sda.ranking()* of the package sda applied for every endocrine phenotype to a set of 500 statistical individuals.

For every combination of LS and NV, a mean of the classification performances was calculated for every individual from bootstrapped results; this was a way to provide, at the sample level, global classification statistics which will be then submitted to PCA (Figure 1, Block 3).

Validation of class assignment predictions by a permutation analysis

For any endocrine phenotype, the validation of the class prediction was confirmed by using a repeated permutation analysis of the dataset (n = 1000). We have used datasets built from the 200 most informative variables with 400 statistical individuals in the training set. Class information of statistical individuals of the training set was randomly assigned and then prediction of the remaining

statistical individuals present in the test set was compared to true one. More precisely, a multivariate analysis of variance (MANOVA) test was done on the two resulting SDA components in function of the true class assignment. A Fisher value using the Hotelling-Lawley criterion was obtained at each iteration. A log10 transformation of the F-values was achieved to approximate their distribution to a seemingly Student one. The F-values of the true model obtained by bootstrap were compared to the random distribution.

Prediction of minimal cohort sizes to get a fixed GPR value

Fitting of sigmoidal curves used the generic following equation (Figure 1, Block 3):

$$y = \frac{b_0 + b_1 x}{c_0 + e^{-\sum_i^p a_i x^i}}$$

with for model (1) $p = 1$ and $b_1 = 0$, for model (2) $p = 2$ and $b_1 = 0$, for model (3) $p = 2$, for model (4) $p = 1$ and for model (5) $p = 7$, $b_0 = 1$, $b_1 = 0$ and $c_0 = 1$.

Initial values describing Q5 values were obtained from the bootstrap procedure described above and were used to extend these primary values with an increment of one in x, using the robust locally weighted regression function *loess()* in R. Such predicted values ranging from the inflection point of the sigmoidal curve to the maximal size of the training cohort were then used to fit a sigmoid model. To predict x above the maximal size of the training cohort (n = 600), maximal acceptable values for x were most of the time extended to 3000. Models (1) to (4) were adjusted directly using the function *nls()* in R. To adjust the model (5), coefficients a_i were estimated by multiple regression between *logit*(y) and the different x^i of the polynomials in (5), with $i \leq 7$. Prediction was assessed graphically and divergent models were discarded. In addition, the true residual variance between the values predicted by fitted models and the initial Q5-values ranging from the inflection point of the curve to the maximal size of the training cohort was calculated.

3. Results

3.1. Unsupervised multivariate assessment of the ^{1}H NMR data structure

Due to the fact that, for numerous cyclists of the cohort, several samples were collected in a longitudinal follow-up frame, these samples may display close fingerprints. If not, samples would appear as unrelated and could be considered in a first approximation as independent. To assess this point, unsupervised multivariate analyses of ^{1}H NMR fingerprints were done (Figure 1, Block 1).

Search for an explicit information structure present inside the metabolomic dataset, which can be unequivocally related to the available endocrine meta-information, could not be obtained neither by PCA (Figure S3), MDS (not shown) nor by PHATE analysis (Figure S4, Insert). Only a projection of samples phenotyped on the second ^{1}H NMR session is distinctly established, mostly on the 2nd PC of PCA or on the 2nd PHATE component, from those analyzed on the first session. Neither cluster displaying a seasonal effect, nor a discipline related effect was detected by PCA (Figure S3). No cluster of the 21 endocrine phenotypes accessible to this analysis which were resulting from combined unitary endocrine phenotypes, including normal subjects for every of the three endocrine phenotypes considered, could really be detectable. In the score PHATE projection (Figure S4, Insert), only a Guttman effect, which corresponded to a non-linear relationship between these two first components, was seen similarly to what was obtained by PCA (Figure S3) or MDS (not shown).

The distribution of scores of samples projected on the PHATE factorial map built with the two first components was not explained by a subject-based clustering as detailed hereafter. Most of

barycenters calculated for cyclists from respective coordinates of samples inside a longitudinal follow-up series of a given subject (number of samplings > 1) were mostly projected in the center part of the factorial map. Lines connecting barycenters to their respective samples revealed a large intra-subject variance, probably explained by environmental effects, which superimposed to the significant *'NMR session'* effect for which a clear discrimination of barycenters was displayed, mostly on the 2nd component (Figure S4).

A mixed-model effect analysis of the random variable consisting into the Euclidean distance of samples included in a given longitudinal follow-up series to their respective barycenter, considering the factor *'subject'* as random factor, and, as fixed factors, the *'NMR session'* factor and the *'number of samples'* for a given subject in a longitudinal series, did not reveal any effect of the *'number of samples'* neither nor any *'NMR session'* effect on such distances, but only a difference in the session-dependent fingerprints resulting in mean values of distances calculated for subjects phenotyped on both sessions which were larger than those calculated for subjects phenotyped in a unique NMR session ($p < 0.001$). In addition, MANOVA applied to the two variables corresponding to the two first PHATE components which was performed according to the *'sample-to-barycenter distance'* factor, the *'NMR session'* factor, the *'number of samples'* enclosed in a longitudinal series and, last, the dummy matrix built from all subjects submitted to more than one blood sample collection, only the *'NMR session'* factor and the *'sample-to-barycenter distance'* variable were very significant ($p < 0.001$ and $p < 0.01$, respectively). Among the subjects repeatedly collected who appeared in the dummy matrix, only 6 over 157 were significantly involved in the explanation of the projection of these specific samples on the PHATE factorial map with $p < 0.01$ (Figure S4). Barycenters of these rare subjects are projected preferentially in regions close to the extreme parts of the 2nd component of the factorial map (Figure S4).

So, on a first approximation, every sample could be considered as being independently sampled from a large population of samples collected on cyclists, even though an appreciable proportion was collected in the longitudinal follow-up of numerous subjects as indicated in Table S3. In the following part of the study, the longitudinal follow-up of subjects was not taken into account and, therefore, the total size of the dataset comprised 655 samples considered as independent statistical individuals.

Since no evident clustering of homogenous endocrine subgroup could be detected by unsupervised analyses, in particular by the MAGIC-PHATE approach, and considering the independence between the distribution of hormonal values of the three hormones (see hereafter § 3.2), metabolomic detection of anomalous profiles could only be achieved by a supervised multivariate classification procedure and was performed considering the three endocrine profiles separately.

3.2. Robust algorithm design for discrimination of endocrine phenotype classes

To define and validate a multivariate statistical strategy minimizing both false positive and false negative rates in the classification of metabolomic data coming from ^{1}H NMR-derived metabolic profiles acquired in one of the three classes named *Low*, *Normal* or *High*, *i.e.* low, normal or high concentrations of every three hormones, a deep datamining exploration was performed. A rather generic procedure was built to finally predict the minimal size a training set of samples should have to improve satisfactorily the classification performance of a training metabolomic dataset to get a global classification error rate lower than 0.1% (Figure 1, Block 2), a parameter which is in line with the 99.9%-credibility value of target biomarkers of ABP modules (Robinson et al. 2017).

In preliminary classification tests, a partial-least squares-based discriminant analysis (PLS-DA) was used on the total available dataset (419 variables) to get a separation of the three groups

corresponding to either *Low*, *Normal* or *High* classes for every endocrine trait considered here, with or without a prior data correction using a PLS-based orthogonal signal correction (OSC) achieved with only one orthogonal component (1 OSC). Because the number of variables (p = 419) was rather high, compared to the total number of statistical individuals (n = 655), this PLS-DA should be more robust than a conventional linear discriminant analysis and therefore should be privileged. Yet, no satisfying separation of phenotypic classes was observed without or with one OSC as shown for cortisol (Figures S5A & S5B). With a more robust method of classification based on the shrinkage discriminant analysis (SDA), we could separate the three classes (Figure S5C), the correction with one OSC giving a more satisfying separation between them (Figure S5D). Interestingly, Mahalanobis distances, used to predict the assignment of a given sample to every phenotypic class, gave a clear distinction between classes (Figure S5E), which is strikingly reinforced with a prior correction with one OSC (Figure S5F). So, this preliminary datamining assessment resulted in a simple classification algorithm which could be performed as following: i) orthogonal correction of the dataset according to the prior classification information in classes of a given endocrine phenotype considering only one orthogonal component using a PLS-based procedure, and then ii) SDA performed on the OSC-corrected dataset. But, SDA sorted variables according to a cat score based on a Student test (Ahdesmaki & Strimmer 2010; Zuber & Strimmer 2009). Therefore, it was necessary to validate such a classification algorithm by studying effects on discrimination performances of both the size of the training set (LS) and the number of variables (NV) being previously selected on their cat scores. These two parameters were systematically studied and results were summarized in abaci (Figure 1, Block 3).

For the four chemometric situations examined in the study, this first validation procedure was confirmed by a complementary permutation analysis. The F-values resulting from a MANOVA test performed on the two resulting SDA components calculated on permuted training datasets display a distribution of values which are far more lower than the true ones (Figure S6). True models were all highly significant ($p < 2 \times 10^{-16}$). Therefore, the probability that overfitting problems may arise when using an OSC-SDA approach as described here would be strictly limited.

4. Discussion

4.1. Parameters influencing GPR

Technically, the noisy information had a deep effect on performances of classification given by SDA or by PLS-DA, but it was only by SDA that high efficiency of PLS-based OSC was displayed. When this correction was applied, SDA resulted in a striking reduction of the intra-class variance thanks to the generalized shrinkage procedure used to calculate Mahalanobis distances (Ahdesmäki & Strimmer 2010), and then to calculate maximal distances between the three class barycenters, enabling therefore calculation of higher probability values to assign unknown samples in the most probable class, hence maximizing performance of the overall classification process. As exemplified here, this was well demonstrated in case of a deep imbalance in the distribution of subjects in three endocrine classes for every endocrine phenotype as often found in epidemiological studies, this constraint being explained here by the fact that abnormal serum concentrations were hopefully more rarely found than the normal ones. The James-Stein shrinkage of priors was very efficient in this case.

Probably, for these fitted cohort sizes from Q5 GPR predicted values, using a more balanced distribution of subjects across the three classes for every endocrine phenotype should give relatively

higher values of GPR than anticipated here through fitting studies. Yet, given the epidemiological constraint linked necessarily to the imbalanced distribution of subjects in three endocrine classes for every endocrine phenotype, GPR is a simple way to weigh overall classification performances. Therefore, study of GPR according to LS and NV was helpful to get global non-linear tendencies of variation.

4.2. Tracking heterogeneities in every *Normal* class of the three endocrine phenotypes and putative biomarkers supporting them

4.2.1. General considerations

Probably, only two main δ regions were involved in the construction of IC_λ characterising outliers of the testosterone *Normal* class, *i.e.* the δ regions [4.465-4.255 ppm] and [1.665-1.615 ppm], completed with δ 2.915 and δ 1.065 ppm; hence, they did not support really an efficient integration of the quantitative signal into a predefined list of known metabolites (see Figure S2) which relative amount was integrated thanks to the BATMAN processing (Table S4). Surprisingly, L-threonine was found as a BATMAN-integrated metabolite which was significantly correlated to IC_λ, but at a very low frequency (14‰), even this analyte seemed to be highly correlated to IC_λ when considering separately the four different δ variables manually assigned to L-threonine, which were all correlated with a very high score frequency (equal to or higher than 963 over 1000) to the different IC_λ repeatedly calculated (Table 2). This point can be easily explained by the way every IC_λ was built and which took into account at the same time the different bucketed δ variables, *i.e.* those presenting a very high score frequency. It was only inside this set of so-selected bucketed δ variables displaying a structured information as shown in Figure S9 that an expert-based assignment could be obtained. But this information structure was not encountered in the BATMAN-integrated dataset and therefore the prior assigned metabolites in the sole δ regions [4.465-4.255 ppm] and [1.665-1.615 ppm] were not in a sufficient number to get usable correlations.

For the cortisol *Normal* class, on the contrary, the BATMAN-integrated metabolite dataset displayed for glycine and 7-methylxanthine found as candidate biomarkers a significant correlation to IC_λ built from bucketed δ variables (Table 2). The five main δ regions found in the following intervals: [3.435-3.425 ppm], [2.745-2.695 ppm], [2.445-2.435 ppm], [2.285-2.265 ppm] and [1.165-1.155 ppm] authorized such a positive correlation (Figure S10B).

Table S1. Synthetic review of age parameters regarding blood collection of subjects repeatedly collected or not.

Parameters	Minimal	Median	Maximal
Age of subjects	19.9	24.3	41.0
Median values of age of subjects repeatedly collected[a]	24.4	24.8	27.6

[a] Only subjects submitted to more than one blood sample collection were considered (n = 135).

Table S2. Distribution of samples in every endocrine phenotype between the class displaying concentrations lower than the lowest limit of normal values (class *Low*), the class displaying seemingly normal concentrations (class *Normal*) and the class displaying concentrations higher than the highest limit of normal values (class *High*). The number of non-phenotyped samples (*NP*) for a given endocrine trait is also given.

Endocrine traits	NMR sessions	Classes										
		Low		*Normal*		*High*		*NP*		Total without *NP*		Total
cortisol	1	30	66	222	303	115	277	1	9	367	646	655
	2	36		81		162		8		279		
IGF1	1	67	71	224	491	77	84	0	9	368	646	
	2	4		267		7		9		278		
testosterone	1	53	78	267	500	42	57	6	20	362	635	
	2	25		233		15		14		273		

Table S3. Distribution of samplings collected one time or repeatedly done among fully hormonally normal (non-disrupted for any endocrine trait) or hormonally disrupted subjects.

Number of samplings by subject	Number of subjects with a fully normal endocrine profile	Number of subjects with a disrupted endocrine profile
1	67	91
2	17	63
3	5	21
4	1	24
5		17
6		7
7		2
9		1
Total number of samples without *NP*	120	526

Table S4. List of candidate metabolites which relative quantification is obtained by a BATMAN-based procedure.

Selected metabolites				
D-glucose	capryloylglycine	D-arginine	glycine	D-xylose
L-proline	D-alanine	N-acetylserine	L-lactic acid	7-methylxanthine
α-hydroxy-isobutyric acid	2-hydroxy-3-methylbutyric acid	aminoadipic acid	N-acetyllactosamine	hydroxyoctanoic acid
cysteine-S-sulfate	2-ethyl-2-hydroxybutyric acid	putrescine	succinyl-acetone	propyl alcohol
L-homoserine	3α,6β,7β-trihydroxy-5β-cholanoic acid	L-leucine	2-hydroxybutyric acid	dehydroepiandrosterone
γ-caprolactone	oxalacetic acid	3α,6α,7β-trihydroxy-5β-cholanoic acid	threonic acid	Taurine
β-N-acetyl-glucosamine	guanidoacetic.acid	glycerol-3-phosphate	L-cystathionine	Rhamnose
4,5-dihydroorotic acid	methyl-isobutyl ketone	α-D-glucose	ethylmalonic acid	malic acid
canavanine	epi-coprostanol	L-α-amino-butyric acid	benzene-acetic acid	glycolic acid
maltotetraose	deoxycholic acid	O-phospho-ethanolamine	scyllo-inositol	methionine sulfoxide
sphingosine	L-threonine	homo-L-arginine	2-hydroxy-2-methylbutyric acid	D-mannose
citrulline	L-alanine	L-valine		

Figure captions

Figure 1. Flow-chart of the different statistical analyses sequentially used to mine the ^{1}H NMR dataset and data subsets used to endocrine phenotypes of cyclist sportsmen. It was divided in four block corresponding to parts of results indicated in brackets as following: [3.2.1 & 4.1], [4.2], [4.3 & 4.4] and [4.7]. Objectives of these specific parts are indicated in conclusive items colored in light violet. Specific statistical tools are indicated in black bold on background colored in green.

Figure 2. Abacuses describing the variation of the median GPR and the confidence interval at 90% according to the training set size (50-650) and the number of prior selected informative variables (50, 100, 150, 200 and 400) thanks to the cat scores. The confidence interval limits (in dotted lines) were colored as their respective median curve. Only one orthogonal component was used to correct the dataset for cortisol (A), IGF1 (B) and testosterone (C). Four orthogonal components were used to significantly increase the GPR performances for IGF1 (D).

Figure 3. Score plots and frequency distributions along the two first PCs for cortisol (A), IGF1 (B) and testosterone (C) with one orthogonal component (1 OSC), and IGF1 with 4 orthogonal components (4 OSC) (D). The 3 classes are given in green, violet and red for classes *Normal*, *Low* and *High* of every endocrine phenotype. Light grey squares marked with a green line focus on the apparent outliers of the *Normal* class excluded from the ellipses drawn at a confidence limit above 97.5%.

Figure 4. Metabolic modules containing the putative biomarkers characterizing the outliers of the testosterone (A) and cortisol (B) *Normal* classes and connected metabolites as summarized in the respective sub-networks defined from a significant enrichment according to a Benjamini-Hochberg correction of the prior selected biomarkers using the MetExplore web tool. 1-Methylhistidine (meat consumption) also detected as marker for cortisol *Normal* class was not enclosed in the prior selection and enrichment analysis of targeted metabolic biomarkers.

Supplementary figure captions

Figure S1. Distribution of cortisol (A), IGF1 (B) and testosterone (C) concentrations of the cyclists' cohort as a function of age. A distinction was done between the groups they belonged to, *i.e.* the hormonally normal group in green, the hormonally higher than normal group (group *High*) in red and the hormonally lower than normal group (group *Low*) in violet. For every subgroup inside a hormonal trait, a linear regression between serum concentration and age [20-34 yrs.] was performed. Although a very significant variation of IGF1 concentration with age was observed in the normal IGF1 group studied here, the slope value was not sufficient to conclude that age would be considered as a possible confounding factor when older cycling athletes with very low IGF1 concentration in serum would be still considered as normal ones. For ease of representation, the sample collected on the subject 41.0 yrs. old was discarded.

Figure S2. Example of a ^{1}H NMR metabolic fingerprint obtained at 600.13 MHz on a serum collected on an anonymous subject. Some assignments to known chemical shifts were done from those usually recorded for analytes detected in serum or plasma. Glu, Gln and Val corresponds to L-glutamate, L-glutamine and L-valine, respectively.

Figure S3. Score plot of the different statistical individuals from a PCA analysis of the raw dataset comprising bucketed ^{1}H NMR fingerprints (p = 419 variables) acquired on two different NMR sessions (1 and 2). Three types of factors were encoded differently according to the color of plot symbols for cycling discipline, their shape for season where blood samples were collected, and their size for the NMR session of samples fingerprinting.

Figure S4. Score plot of the different statistical individuals from a PHATE analysis of raw dataset comprising bucketed ^{1}H NMR fingerprints (p = 419 variables) acquired on two different NMR sessions (1 and 2) appearing in filled circles and triangles, respectively. PHATE analysis was performed on a data subset (n = 633) for which all samples were completely phenotyped for the three endocrine traits with no non-phenotyped (NP) data. Circled numbers give the number of samples collected in a longitudinal follow-up for every subject and which are projected on a barycenter corresponding to the mean score point calculated from these different points acquired respectively on the 1st (red circles), 2nd (green circles) or both NMR sessions (blue circles). Most of mean points with a number of samplings per cyclist above one are projected in the center part of the factorial map. Lines connect respective barycenters to the different points corresponding to samples collected for every cyclist and are colored according to the different subjects. Using MANOVA between the two coordinates of the factorial map on one part and, on another part, the set of following factors, *i.e.* the *'NMR session'* factor, the number of samples in a longitudinal series, the Euclidean distance calculated from barycenters to samples inside a longitudinal series, all these factors being joined to the dummy matrix built from subjects phenotyped more than one time, shows in addition to a highly significant *'NMR session'* effect and a *'barycenter-to-sample distance'* effect, an effect of some rare subjects whose barycenters are projected close to the extreme parts of the 2nd component of the factorial map and are indicated by the following labels: ● when p < 0.10, * when p < 0.05, ** when p < 0.01 and * when p < 0.05. The 21 specific endocrine phenotypes appear as differently colored from dark blue to light yellow for disrupted samples and in grey for controls (or equivalent to the combined normal endocrine phenotype 'Cortisol N / IGF1 N / Testo N'). Insert gives only the PHATE scores of the different samples on the plan 1 × 2 with circles corresponding to samples analyzed on the 1st NMR session and signs + for samples analyzed on the 2nd one.**

Figure S5. Progressive building of a generic classification algorithm. Partial least squares-discriminant analysis (PLS-DA) without (A) or with a prior PLS-based orthogonal signal correction with one orthogonal component (1 OSC) (B) considering the cortisol grouping factor. Shrinkage discriminant analysis (SDA) without (C) or with a prior OSC with 1 component (D). Mahalanobis distance plots used to calculate probabilities to predict belonging of one individual to a given phenotypic subgroup by SDA without (E) or with a prior OSC with 1 component (F). Legend: the different phenotypic subgroups are given with normal cortisol concentrations in serum in green, lower cortisol concentrations than normal in violet, and higher cortisol concentrations than normal in red.

Figure S6. Density plots of the distribution log10-transformed F-values coming from a MANOVA test (Hotelling-Lawley criterion) applied to SDA components repeatedly calculated in a bootstrap procedure using at each iteration a permutation test (line curves in red). True values for cortisol (A), IGF1 (B) and testosterone (C) datasets, all corrected by OSC with one orthogonal component, and for IGF1 dataset prior submitted to 4 OSC components (D), were indicated by a narrow distribution obtained by bootstrap (n = 1000, line curves in black). All these specific distributions were significantly different from those resulting from random permutations ($p < 2 \times 10^{-16}$).

Figure S7. Variation of the prediction rates of classification of individuals in one of the different phenotypic subgroups for either cortisol (A), IGF1 (B and D) or testosterone (C) assayed in serum according to the size of the training set, which varied from 50 to 650. One PLS-based correction component (1 OSC) was used to get a prior correction of the dataset for the different hormonal traits. Influence of a correction with 3 supplementary orthogonal components (4 OSC) was also tested on the prediction rates of the IGF1 status. For every hormonal trait, prediction rates were presented for the different subgroups. A global prediction rate (GPR) corresponding to the multiplication of the three specific ones for a given hormonal trait was done at every training set size. Thanks to the resampling procedure used (n = 1000), a confidence interval (in light grey) was calculated in the 5%-95% quantile range (dotted lines).

Figure S8. PHATE analysis of the distribution of individuals belonging to the class *Normal* of the following endocrine phenotypes: cortisol (A), IGF1 (B) and testosterone (C) through 4 groups. This number of groups in the *Normal* class corresponded to the maximal one for which all outliers of the cortisol *Normal* class were assigned to a unique subgroup. The same number of subgroups was used to display the metabolomic heterogeneity inside IGF1 and testosterone *Normal* classes and then to check the distribution of outliers inside the respective different subgroups. Outliers prior detected by a PCA analysis of the bootstrap-supported cross-validation data are marked with larger size points (circle or triangle). Only for the normal cortisol trait, outliers belonged to a unique subgroup.

Figure S9. Clustering of bucketed δ variables involved in the building of the most important independent component discriminating outliers from true normal individuals for the testosterone endocrine phenotype. All selected variables displayed a score above 700. A: Heatmap focused on selected δ variables, B: spectrum plotting of selected δ variables, and C: expert-based putative assignment of chemical shifts.

Figure S10. Clustering of bucketed δ variables involved in the building of the most important independent component discriminating outliers from true normal individuals for the cortisol endocrine phenotype. All selected variables displayed a score above 899. A: Heatmap focused on selected δ variables, B: spectrum plotting of selected δ variables, and C: expert-based putative assignment of chemical shifts.

Figures

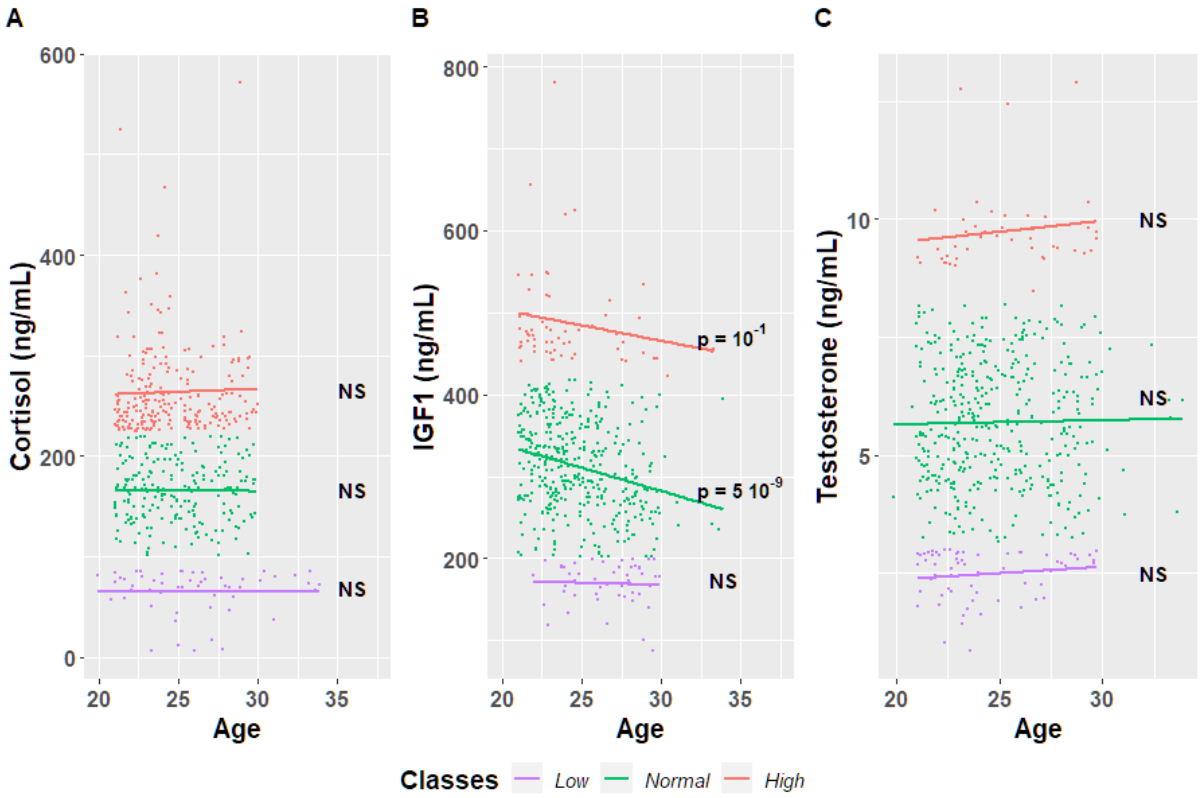

A
B
C
Cortisol (ng/mL)
IGF1 (ng/mL)
Testosterone (ng/mL)
Age
NS
NS
NS
p = 10-1
p = 5 10-9
NS
NS
NS
NS
Classes
Low
Normal
High

Figure S1.

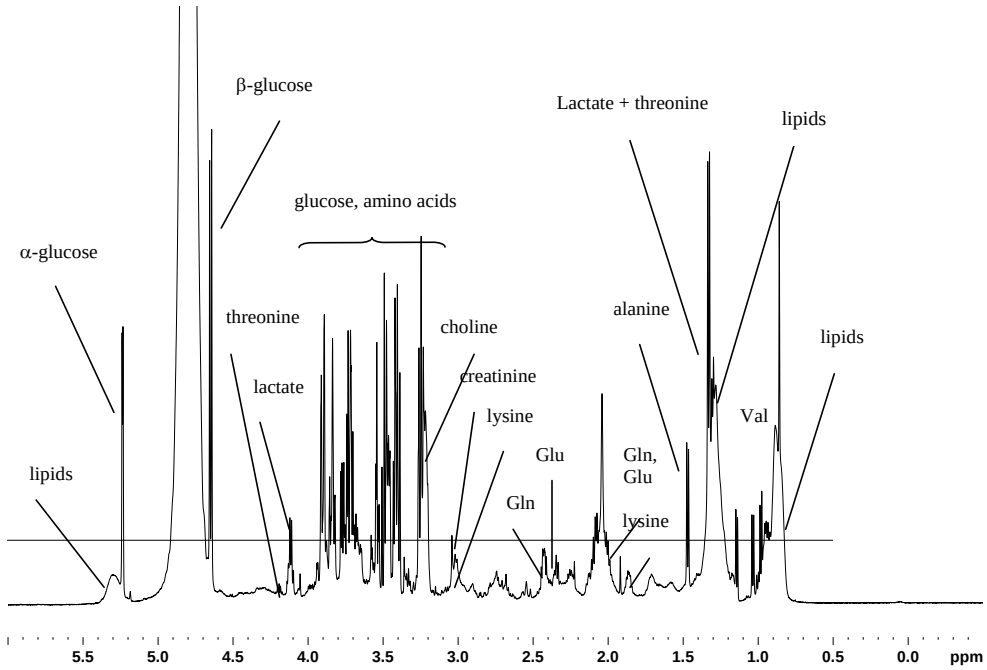

β-glucose
Lactate + threonine
lipids
glucose, amino acids
α-glucose
threonine
choline
alanine
lipids
lactate
creatinine
lysine
Val
lipids
Glu
Gln,
Glu
Gln
lysine
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
ppm

Figure S2.

.

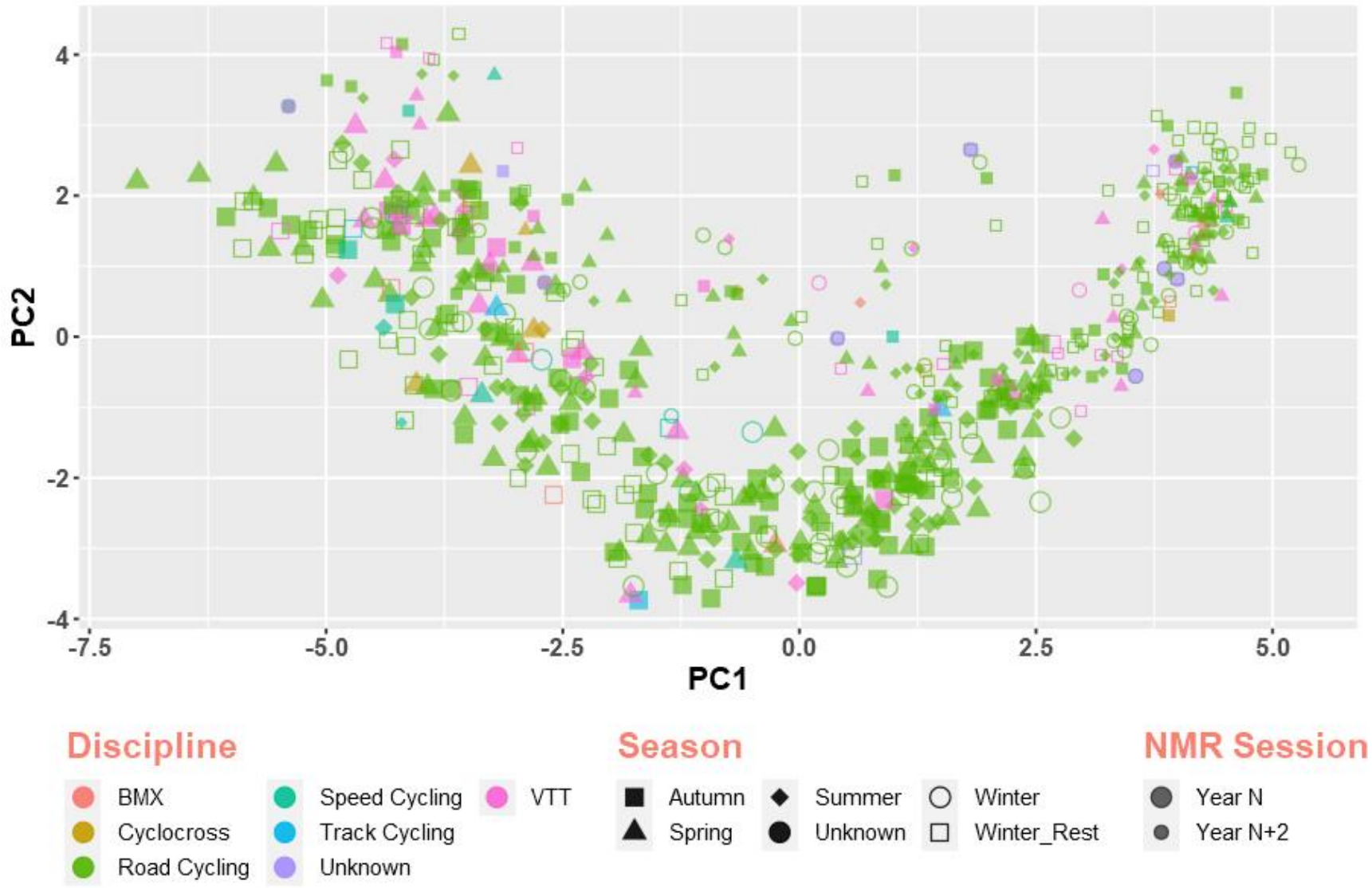

PC2
4
2
0
-2
-4
-7.5
-5.0
-2.5
0.0
2.5
5.0
PC1
Discipline
BMX
Cyclocross
Road Cycling
Speed Cycling
Track Cycling
Unknown
VTT
Season
Autumn
Spring
Summer
Unknown
Winter
Winter_Rest
NMR Session
Year N
Year N+2

Figure S3.

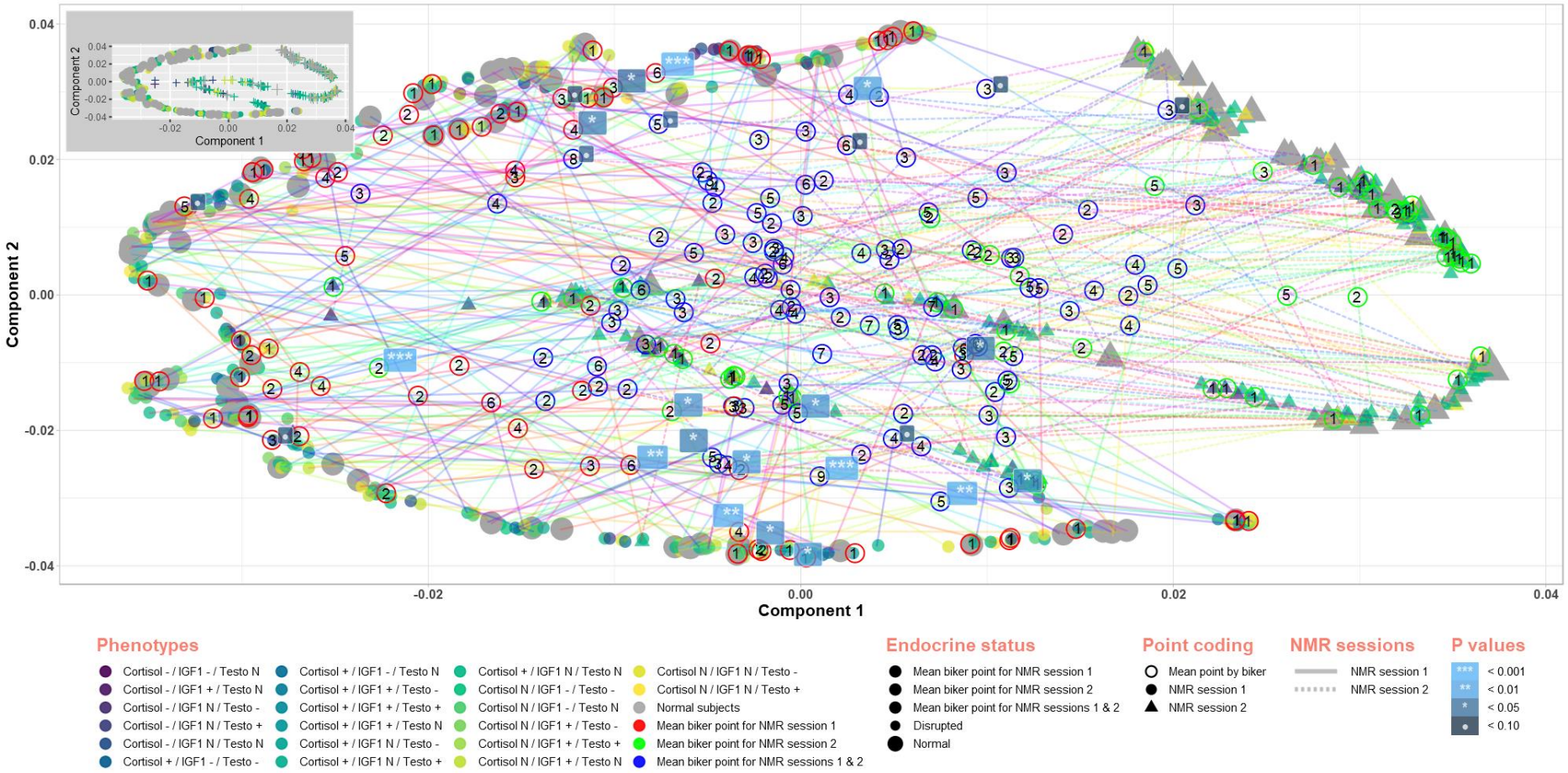

Component 2
Component 1
0.04
0.02
0.00
-0.02
-0.04
Phenotypes
Cortisol - / IGF1 - / Testo N
Cortisol - / IGF1 + / Testo N
Cortisol - / IGF1 N / Testo -
Cortisol - / IGF1 N / Testo +
Cortisol - / IGF1 N / Testo N
Cortisol + / IGF1 - / Testo -
Cortisol + / IGF1 - / Testo N
Cortisol + / IGF1 + / Testo -
Cortisol + / IGF1 + / Testo +
Cortisol + / IGF1 + / Testo N
Cortisol + / IGF1 N / Testo -
Cortisol + / IGF1 N / Testo +
Cortisol + / IGF1 N / Testo N
Cortisol N / IGF1 - / Testo -
Cortisol N / IGF1 - / Testo N
Cortisol N / IGF1 + / Testo -
Cortisol N / IGF1 + / Testo +
Cortisol N / IGF1 + / Testo N
Cortisol N / IGF1 N / Testo -
Cortisol N / IGF1 N / Testo +
Normal subjects
Mean biker point for NMR session 1
Mean biker point for NMR session 2
Mean biker point for NMR sessions 1 & 2
Endocrine status
Mean biker point for NMR session 1
Mean biker point for NMR session 2
Mean biker point for NMR sessions 1 & 2
Disrupted
Normal
Point coding
Mean point by biker
NMR session 1
NMR session 2
NMR sessions
NMR session 1
NMR session 2
P values
*** < 0.001
** < 0.01
* < 0.05
• < 0.10

Figure S4.

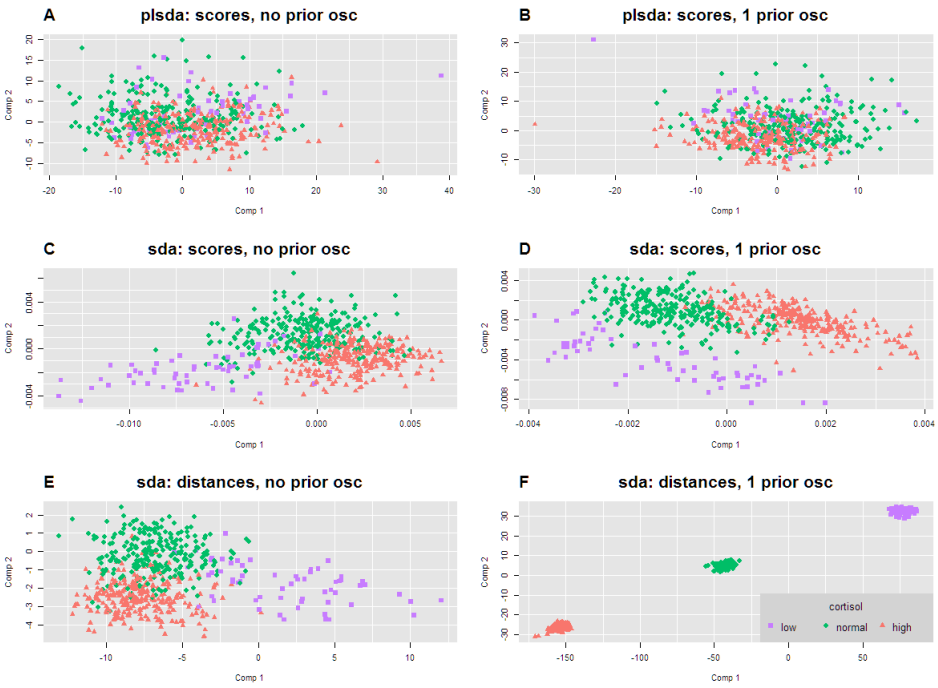

A
plsda: scores, no prior osc
B
plsda: scores, 1 prior osc
C
sda: scores, no prior osc
D
sda: scores, 1 prior osc
E
sda: distances, no prior osc
F
sda: distances, 1 prior osc
Comp 1
Comp 2
cortisol
low
normal
high

Figure S5.

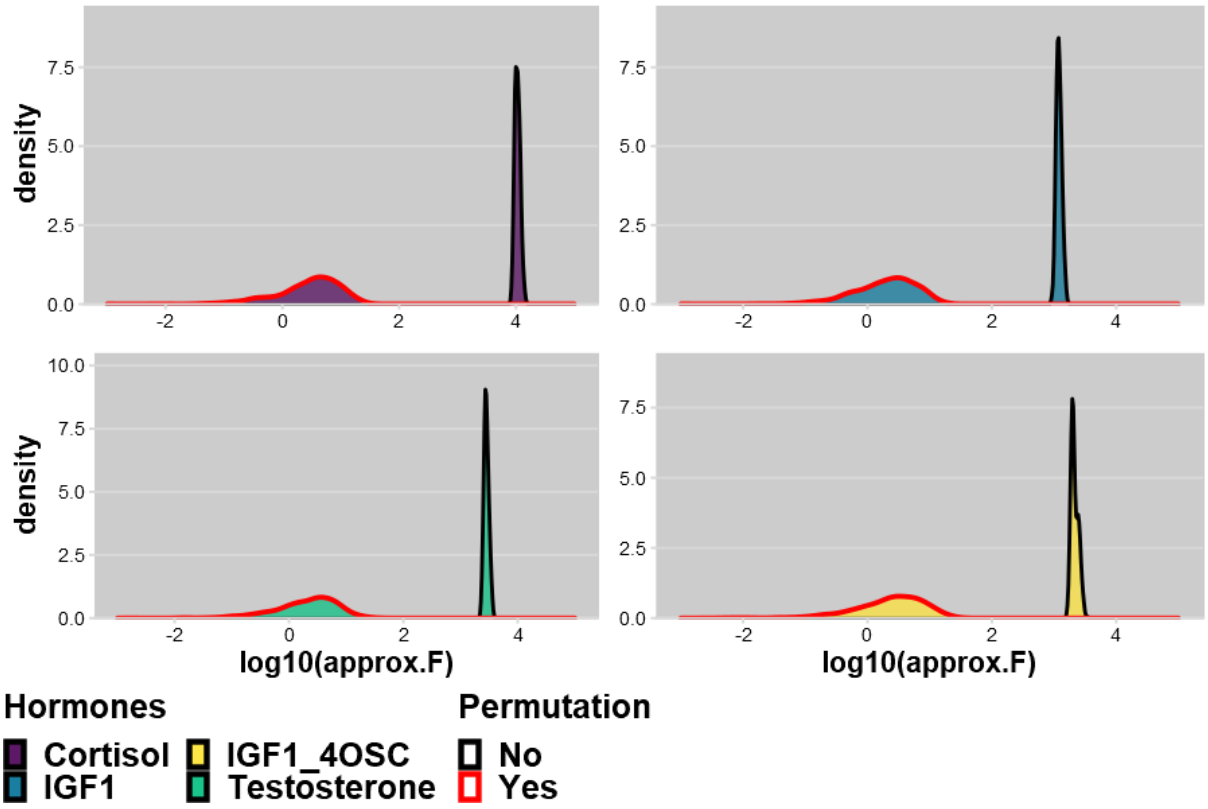

density
7.5
5.0
2.5
0.0
-2
0
2
4
10.0
log10(approx.F)
Hormones
Cortisol
IGF1
IGF1_4OSC
Testosterone
Permutation
No
Yes

Figure S6.

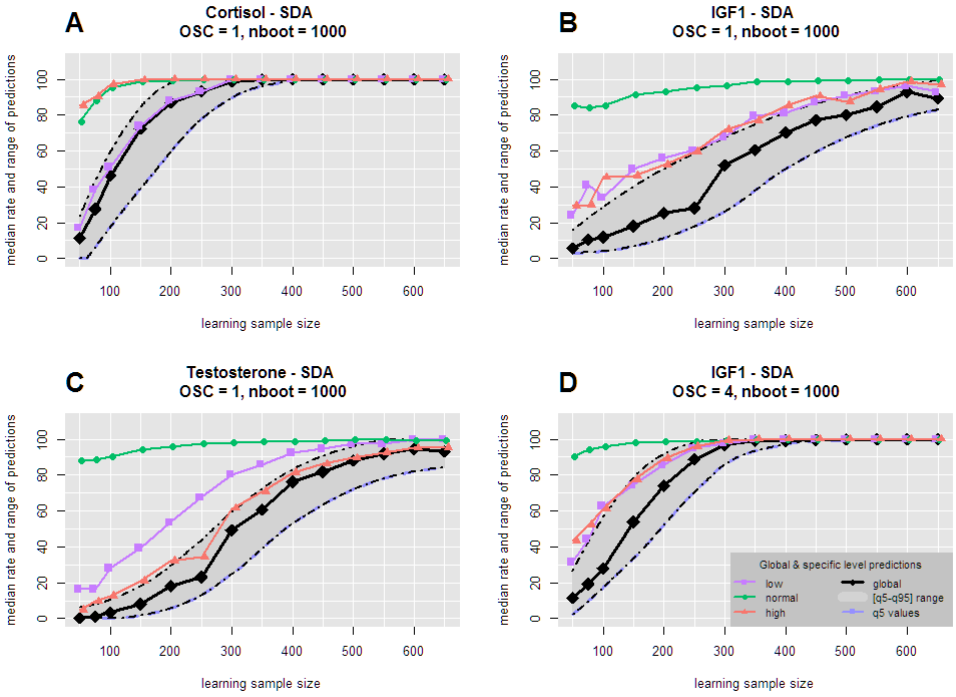

A
Cortisol - SDA
OSC = 1, nboot = 1000
B
IGF1 - SDA
OSC = 1, nboot = 1000
C
Testosterone - SDA
OSC = 1, nboot = 1000
D
IGF1 - SDA
OSC = 4, nboot = 1000
median rate and range of predictions
learning sample size
Global & specific level predictions
low
normal
high
global
[q5-q95] range
q5 values

Figure S7.

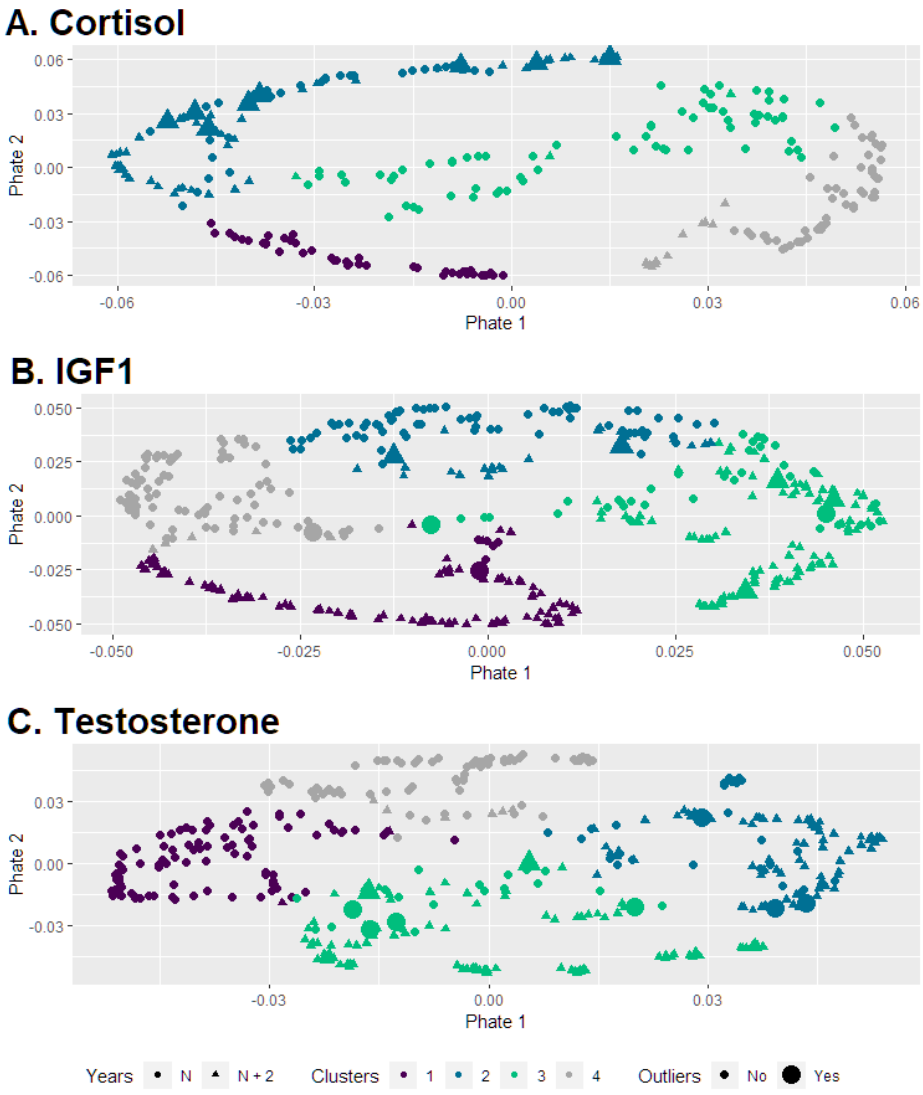
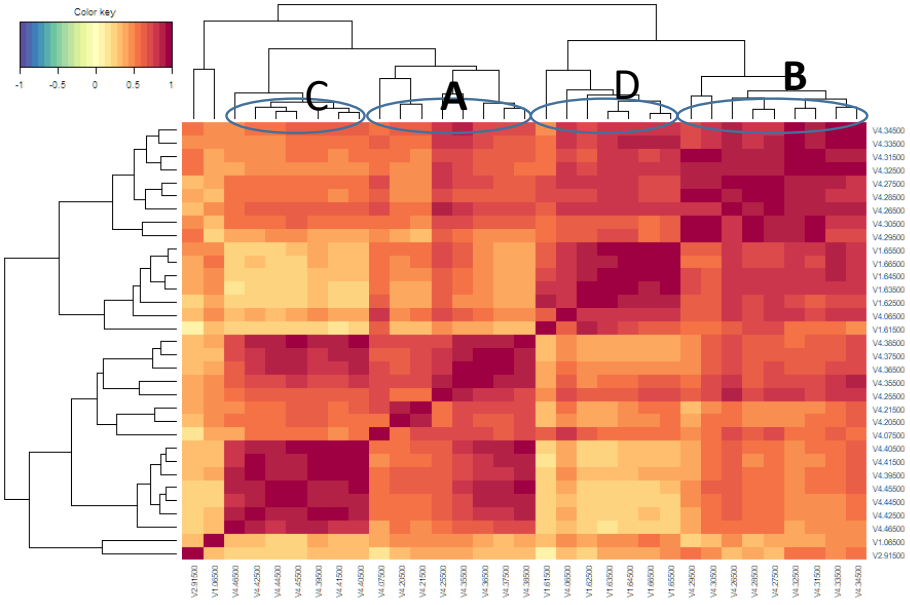

A. Cortisol
Phate 2
Phate 1
B. IGF1
Phate 2
Phate 1
C. Testosterone
Phate 2
Phate 1
Years N N + 2
Clusters 1 2 3 4
Outliers No Yes

Figure S8.

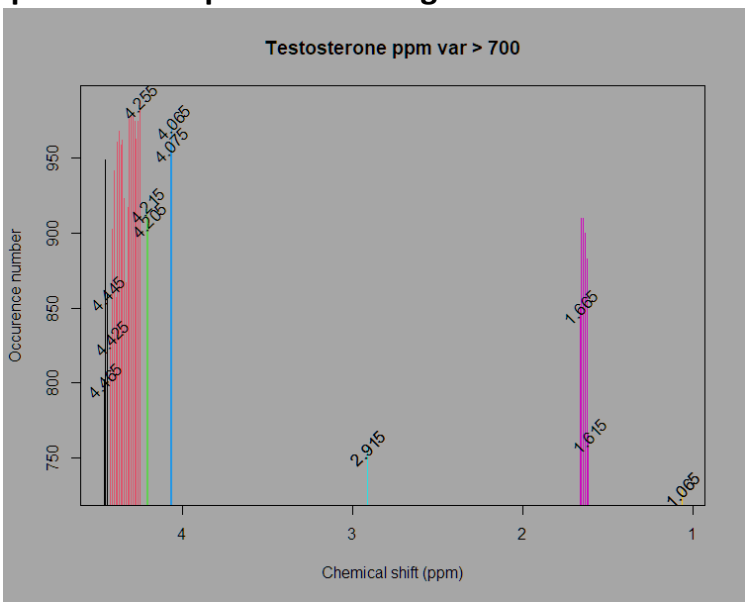
A

Heatmap of candidate δ variables involved in independent component building with a score above 700



B

Spectrum plotting of candidate δ variables involved in independent component building with a score above 700



C

Expert-based assignment of 4 groups of variables named according to δ (ppm)

Priority and decreasing relative importance sorted from cluster A to D

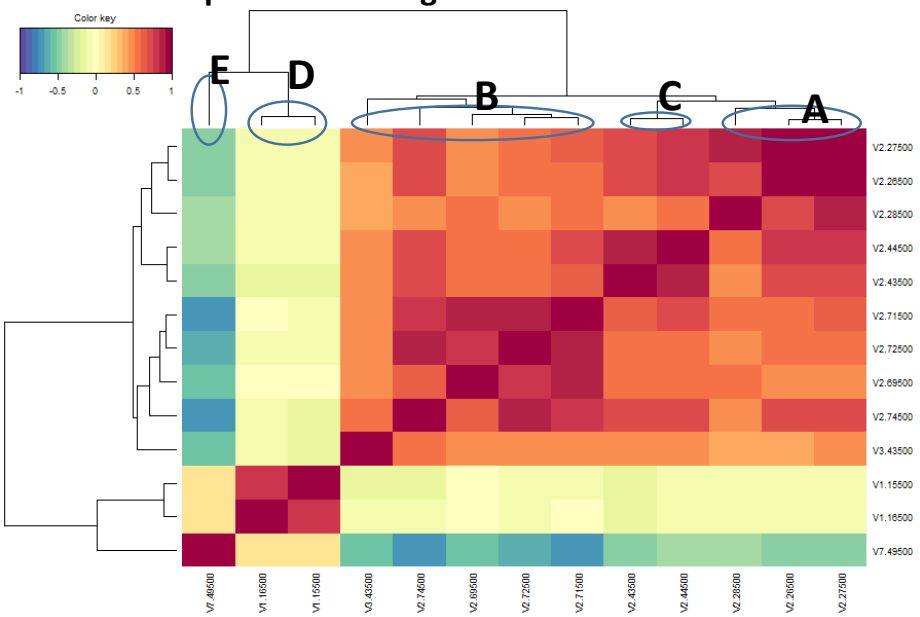
(): score value in bootstrap procedure (from 765 to 988 over 1000)

- # **B** # 1. 4.295 (978) 4.305 (978): **unknown**; 4.265 (965) 4.285 (975) 4.275 (963): **threonine**; 4.325 (917) 4.315 (976) 4.335 (867) 4.345 (923): **α-glycerylphosphorylcholine**
- # D # 2. 4.065 (974): **choline**; 1.615 (765) 1.625 (883) 1.635 (900) 1.645 (910) 1.665 (851) 1.655 (910): **arginine**
- # **A** # 3. 4.075 (959): **choline** (4.075); 4.205 (909) 4.215 (919): **unknown** (4.205, 4.215); 4.255 (988): **threonine** (4.255); 4.355 (962) 4.365 (959) 4.375 (968) 4.385 (961): **noise**
- # C # 4. 4.405 (942) 4.415 (903) 4.395 (857) 4.455 (949) 4.445 (860) 4.425 (830) 4.465 (802): **unknown** (weak and large signals)

Figure S9.

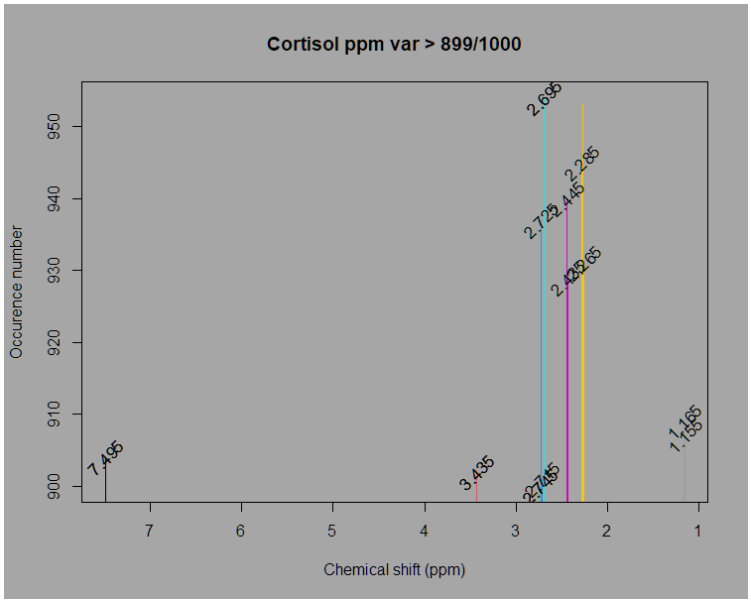
A

Heatmap of candidate δ variables involved in independent component building with a score above 899



B

Spectrum plotting of candidate δ variables involved in independent component building with a score above 899



C

Expert-based assignment of 5 groups of variables named according to δ (ppm)

Priority and decreasing relative importance sorted from cluster A to E

(): score value in bootstrap procedure (from 900 to 953 over 1000)

- # E # 1. 7.495 (904): **unknown**
- # D # 2. 1.165 (909) 1.155 (907): **lipids**
- # **B** # 3. 3.435 (902): **glucose**; 2.745 (900) 2.725 (937): **lipids**; 2.695 (954) 2.715 (901): **citrate**
- # **C** # 4. 2.435 (929) 2.445 (940): **glutamine**
- # **A** # 5. 2.285 (945) 2.265 (931) 2.275 (953): **valine**

Figure S10.

References

Ahdesmäki, M., Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Annals of Applied Statistics* 4:503-519.

Ahdesmäki, M., Zuber, V, Gibb, S., Strimmer, K. (2015). sda: Shrinkage Discriminant Analysis and CAT score variable selection. R package version 1.3.7. https://CRAN.R-project.org/package=sda.

Barker, M.L., Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics* 17:166-173. doi: 10.1002/cem.785

Bidlingmaier, M., Friedrich, N., Emeny, R.T., Spranger, J., Wolthers, O.D., Roswall, J., Körner, A., Obermayer-Pietsch, B., Hübener, C., Dahlgren, J., Frystyk, J., Pfeiffer, A.F., Doering, A., Bielohuby, M., Wallaschofski, H., Arafat, A.M. (2014). Reference intervals for insulin-like growth factor-1 (igf-1) from birth to senescence: results from a multicenter study using a new automated chemiluminescence IGF-I immunoassay conforming to recent international recommendations. *Journal of Clinical Endocrinology and Metabolism* 99(5):1712-1721. doi: 10.1210/jc.2013-3059.

Brabant, G., von zur Mühlen, A., Wüster, C., Ranke, M.B., Kratzsch, J., Kiess, W., Ketelslegers, J.M., Wilhelmsen, L., Hulthén, L., Saller, B., Mattsson, A., Wilde, J., Schemer, R., Kann, P., German KIMS Board. (2003). Serum insulin-like growth factor I reference values for an automated chemiluminescence immunoassay system: results from a multicenter study. *Hormone Research* 60(2):53-60. doi: 10.1159/000071871.

Elmlinger, M.W., Kühnel, W., Weber, M.M., Ranke, M.B., Elmlinger, M.W., et al. (2004). Reference ranges for two automated chemiluminescent assays for serum insulin-like growth factor I (IGF-I) and IGF-binding protein 3 (IGFBP-3). *Clinical Chemistry and Laboratory Medicine* 42(6):654-664. doi: 10.1515/CCLM.2004.112.

Landin-Wilhelmsen, K., Lundberg, P.A., Lappas, G., Wilhelmsen, L. (2004). Insulin-like growth factor I levels in healthy adults. *Hormone Research* 62 Suppl 1:8-16. doi: 10.1159/000080753.

Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A.V.D., Hirn, M.J., Coifman, R.R., Ivanova, N.B., Wolf, G., Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology* 37(12):1482-1492. doi: 10.1038/s41587-019-0336-3. Erratum in: *Nature Biotechnology* (2020) 38(1):108.

Robinson, N., Sottas, P.E., Schumacher, Y.O. (2017). The athlete biological passport: how to personalize anti-doping testing across an athlete's career? *Medicine and Sport Science* 62:107-118. doi: 10.1159/000460722.

Rosario, P.W. Normal values of serum IGF-1 in adults: results from a Brazilian population. (2010). *Arquivos Brasileiros de Endocrinologia & Metabologia* 54(5):477-481. doi: 10.1590/s0004-27302010000500008.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., Pe'er, D. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* 174(3):716–729.e27. doi:10.1016/j.cell.2018.05.061.

Westerhuis, J.A., de Jong, S., Smilde, A.K. (2001). Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems* 56:13-25. doi: 10.1016/S0169-7439(01)00102-2.

Wold, S., Antti, H., Lindgren, F., Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems* 44:175-185. doi: 10.1016/s0169-7439(98)00109-9.

Zuber, V., Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics* 25:2700-2707. doi: 10.1093/bioinformatics/btp460.