URINE METABOLOMICS

Report on data description and quality control

Dr. Chung-Ho E. Lau
Dr. Alexandros Siskos
Dr. Hector Keun
Dr. Muireann Coen

Imperial College London

Version 2.0 (updated in January 2017)

For more information please contact:

Dr. Chung-Ho E. Lau
Research Associate
Imperial College London
esmond.lau06@imperial.ac.uk

Dr. Muireann Coen
Lecturer in Metabonomics & Biochemical Toxicology
Imperial College London
m.coen@imperial.ac.uk

Report contents

Tables and figures

Part 1 Sample Run description

Urinary metabolic profiles were analysed on a 14.1 Tesla (600MHz ^{1}H) NMR spectrometer at Imperial College London (ICL) in the final quarter of 2015. In total, 1366 urine samples from 1212 children were analysed - please refer to Table 1 for a full breakdown by cohort. All samples and batching were fully randomised to prevent potential bias in the analytical run from impacting on subsequent data processing/analysis.

We and others have previously shown that combined urine pool, from the morning and the night before, is better at capturing the temporal variability within an individual than a single spot urine sample, and that pooled urine samples result in higher intraclass correlation coefficients (ICC). Thus we have analysed the combined urine samples whenever possible. The morning and night samples were only analysed if we could not create a combine pool due to missing samples, and 48 of the 1273 combined urine samples were pooled at ICL rather than at the source laboratory

Table 1 shows the full breakdown of the urine sample set by cohort and by urine sampling type

	BIB	**EDEN**	**KANC**	**MOBA**	**RHEA**	**INMA**	**Overall**
Number of urine samples	230	184	233	230	230	259	1366
Number of children	202	157	204	230	200	219	1212
By urine sampling							
combined pool	207	164	228	223	200	251	1273
morning	21	8	4	5	7	7	52
night	2	12	1	2	23	1	41

Part 2 Analytical Method

Proton Nuclear Magnetic Resonance Spectroscopy (^{1}H NMR) were used to profile urinary metabolites in the HELIX children cohort study. ^{1}H NMR is a robust and an untargeted (non-selective) profiling platform, and is particularly suited to measuring abundant compounds in urine samples.

Urine samples were analysed at 300 Kelvin on a BRUKER BioSpin AVANCE III 600 Mhz spectrometer equipped with an automated SampleJet using the noesygppr1d pulse sequence.

Data acquisition parameters:

Pulse program	noesygppr1d
Time domain	65536
Dummy scans	4
Scans	32
Sweep width	20 ppm
Acquisition time	2.726 s
Relaxation delay	4 s
Receiver gain	90.5
Dwell time	41.6 µs
Mixing time	0.01 s
Line broadening	0.3 Hz

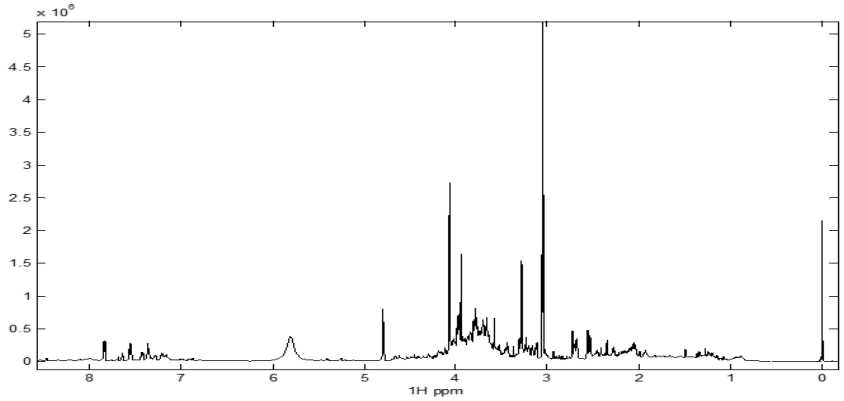

Figure 1. A representative NOESY urine spectrum

Part 3 NMR data processing

We have checked that all the spectra included pass the pre-defined quality control criteria based on reference compound linewidth and peakshape, and have processed the NMR spectral data by performing phasing, baseline correction, and chemical shift referencing. Whilst the data acquisitions were untargeted, our data processing workflow enables us to select the most abundant and prevalent metabolites for targeted data analysis (Table 2) and 44 metabolites have been identified and quantified using the method described below:

1D presat-NOESY spectra (Figure 1) were imported into MATLAB (MathWorks, Massachusetts, US), and were aligned to the pooled quality control sample using recursive segment-wise peak alignment - an algorithm based on cross-autocorrelation with a defined reference sample (Veselkov et al., 2009). Individual metabolite peaks are shown in Figure 2, with the boundaries of each corresponding peak segment region represented by the red dotted lines.

For a subset of samples, poor peak alignment may result from multiple peak features occupying the same spectral region thus causing a lost in feature selectivity, and this is considered a major issue for NMR based metabolomics. Thus, we also assessed the peak alignment for individual sample metabolite signals and computed Pearson's correlation coefficients R^2 between the aligned peak segment and the reference peak segment. The distributions of the sample alignment scores for individual metabolites are shown in Figure 3. While a high correlation coefficient represents good alignment, a low correlation coefficient represents that the sample peak segment has little resemblance to the reference peak segment. Finally metabolite peak signals - area under the curve within the spectral segments as defined by the red dotted lines in Figure 2, were estimated using trapezoidal numerical integration.

To help validate our data processing workflow, we further compared our metabolite signal data to that processed using an alternative commercial software, "Chenomx" (Tredwell et al., 2011) - which contains its own target compound library for reference peak matching. We found good correlations (Pearson R > 0.8) between these two different data processing methods for a large number of metabolites (see Table 2) and for those metabolite signals whose identifies cannot be validated using Chenomx, we also seek to confirm their annotations using 2D NMR and/or STOCSY (Cloarec et al., 2005), an approach that takes advantage of the multicollinearity of the intensity variables in the spectra dataset (see Table 2).

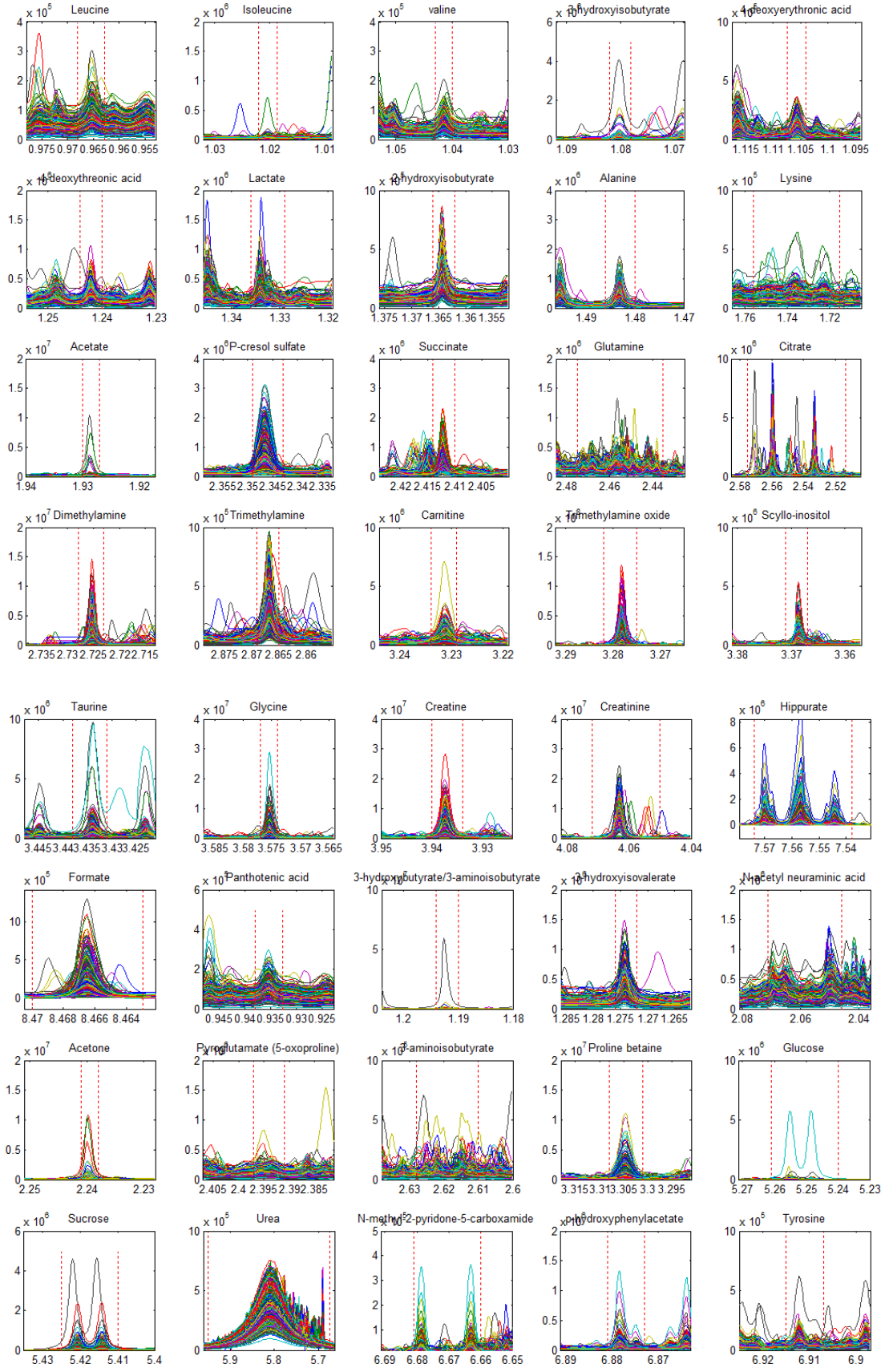

Figure 2. NMR peak resonance for each metabolite signal identified. Red dotted lines indicate the boundaries of the peak signal segment windows

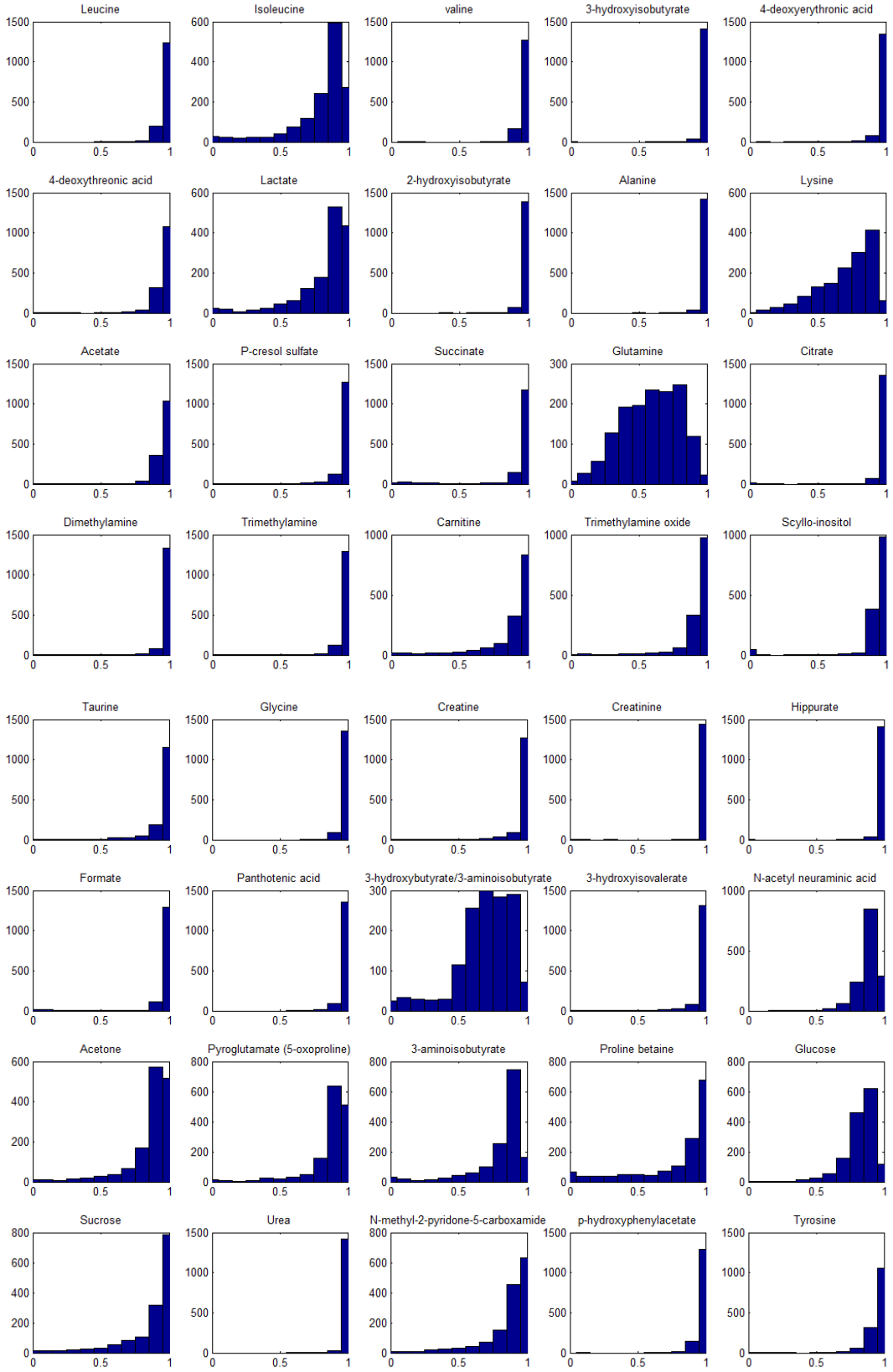

Figure 3. Distribution of peak alignment scores - R^2 of fit between individual samples and pooled QC after peak alignment.

Table 2. Summary of peak annotation and assignment

Signal	resonance (ppm)	Annotation/ assignment
Leucine	0.97	STOCSY
Isoleucine	1.02	STOCSY
Valine	1.04	Chenomx , STOCSY
3-hydroxyisobutyrate	1.08	Chenomx
4-deoxyerythronic acid	1.11	tentative
4-deoxythreonic acid	1.24	tentative
Lactate	1.33	STOCSY
2-hydroxyisobutyrate	1.36	Chenomx
Alanine	1.48	Chenomx, STOCSY
Lysine	1.74	STOCSY
Acetate	1.93	2D
P-cresol sulfate	2.35	STOCSY
Succinate	2.41	Chenomx
Glutamine	2.46	STOCSY
Citrate	2.54	STOCSY
Dimethylamine	2.73	Chenomx, STOCSY
Trimethylamine	2.87	2D
Carnitine	3.23	STOCSY
Trimethylamine oxide	3.28	Chenomx, STOCSY
Scyllo-inositol	3.37	tentative
Taurine	3.43	Chenomx, STOCSY
Glycine	3.58	Chenomx, STOCSY
Creatine	3.94	Chenomx, STOCSY
Creatinine	4.06	Chenomx, STOCSY
Hippurate	7.56	Chenomx, STOCSY
Formate	8.47	Chenomx, STOCSY
Pantothenic acid	0.94	tentative
3-hydroxybutyrate/3-aminoisobutyrate	1.19	tentative
3-hydroxyisovalerate	1.27	Chenomx, STOCSY
N-acetyl neuraminic acid	2.06	tentative
Acetone	2.24	2D
5-oxoproline	2.39	2D
3-aminoisobutyrate	2.62	tentative
Proline betaine	3.30	tentative
Glucose	5.25	STOCSY
Sucrose	5.42	Chenomx, STOCSY
Urea	5.81	Chenomx
N-methyl-2-pyridone-5-carboxamide	6.67	STOCSY
p-hydroxyphenylacetate	6.88	Chenomx, STOCSY
Tyrosine	6.91	2D
3-Indoxylsulfate	7.52	Chenomx, STOCSY
N-methylpicolinic acid	8.72	tentative
N-methylnicotinic acid	9.12	Chenomx, STOCSY
N1-methyl-nicotinamide	9.28	Chenomx, STOCSY

Note.
Chenomx: Indicates good agreement with data obtained using alterative data processing workflow ("Chenomx"). Pearson correlation coefficient $R > 0.8$;
STOCSY: Indicates correlations observed with another resonance in the spectra that are either structurally or biochemically related;
2D: Matching evidence (mainly HSQC spectrum) on a representative INMA child sample.

Method for estimating urinary metabolite concentrations

Finally, the sample concentration of a given metabolite can be estimated from the signal of the internal standard trimethylsilylpropanoic acid (TSP) using the following formula:

$$[M] = [Standard] \times \frac{I_m}{I_s} \times \frac{N_s}{N_m} \times C_{T1}$$

where [*M*] is the metabolite molar concentration, [*Standard*] is the known molar concentration of internal standard TSP, I_m is the metabolite intensity, I_S is the intensity of the TSP peak, N_m is the number of ^{1}H nuclei contributing to the metabolite peak, and N_s is the number of proton contributing to the spiked standard's resonance peak. And C_{T1} is the correction factor for incomplete longitudinal relaxation:

$$C_{T1} = \frac{1 - e^{-t/T1_s}}{1 - e^{-t/T1_m}}$$

where t is the sum of the recycling delay and acquisition time in the pulse sequence, $T1_m$ and $T1_s$ are respectively the T_1 (longitudinal relaxation time) of the metabolite and the TSP resonance as measured using a standard inversion recovery experiment. In total concentrations of 26 out of the 44 metabolites have been estimated using this procedure.

Part 4 Performance and quality control

The coefficients of variation (CV) were assessed based on repeated analysis of 60 identical pooled quality control urine samples and these were analysed at regular intervals during the sample run. A low coefficient of variation represents high analytical precision/stability from repeated measures, and we found that vast majority of metabolites/features achieved a coefficient of variation of < 10% (Figure 4), indicating that sample run was of good analytical reproducibility.

The coefficient of variation and the percentages of non-detected samples for each metabolite are provided in Table 3. We cannot detect the metabolite - N-methylpicolinic acid, in a high proportion of the samples; zero value is recorded when a metabolite is not detected in a sample.

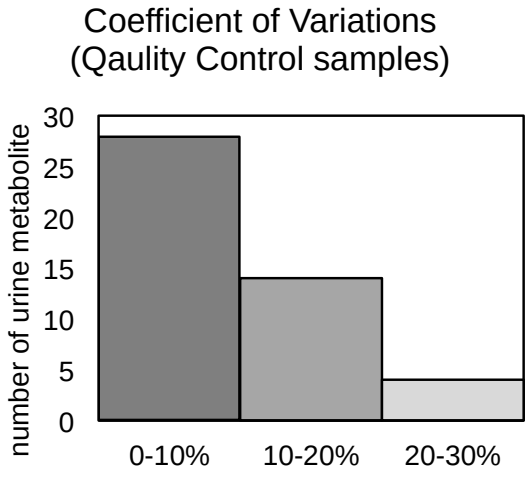

Figure 4. Coefficient of variation (%) for a selection of urinary metabolites in QC samples analysed by ^{1}H NMR spectroscopy.

Table 3. Percentages of non-detected samples for individual metabolites and coefficient of variations based on quality control run

Signals	% Non-detected	Coefficient of variation
Creatinine	0%	7%
Leucine	0%	7%
Isoleucine	1%	14%
Valine	0%	10%
3-hydroxyisobutyrate	0%	7%
4-deoxyerythronic acid	0%	8%
4-deoxythreonic acid	0%	6%
Lactate	0%	4%
2-hydroxyisobutyrate	0%	6%
Alanine	0%	6%
Lysine	0%	7%
Acetate	0%	22%
p-cresol sulfate	0%	8%
Succinate	7%	12%
Glutamine	1%	5%
Citrate	0%	7%
Dimethylamine	1%	14%
Trimethylamine	0%	7%
Carnitine	1%	12%
Trimethylamine oxide	2%	6%
Scyllo-inositol	4%	17%
Taurine	1%	5%
Glycine	0%	9%
Creatine	0%	7%
Hippurate	0%	6%
Formate	0%	9%
Pantothenic acid	0%	7%
3-hydroxybutyrate/3-aminoisobutyrate	3%	10%
3-hydroxyisovalerate	0%	6%
N-acetyl neuraminic acid	1%	13%
Acetone	1%	28%
5-oxoproline	1%	12%
3-aminoisobutyrate	2%	10%
Proline betaine	8%	20%
Glucose	0%	11%
Sucrose	1%	6%
Urea	0%	8%
N-methyl-2-pyridone-5-carboxamide	6%	12%
p-hydroxyphenylacetate	0%	7%
Tyrosine	0%	10%
3-Indoxylsulfate	2%	15%
N-methylpicolinic acid	46%	27%
N-methylnicotinic acid	0%	13%
N1-methyl nicotinamide	1%	20%

Part 5. Data transfer and variables structure

The updated version of the urine metabolomics data file (*Urine Metabolomics Children subcohort Data v2.xlsx*) has been transferred to the HELIX consortium, and this file contains information on 44 quantified metabolites in 1366 samples from 1212 individuals. As mentioned in Part 4, N-methylpicolinic acid cannot be detected in a high proportion of the samples and a zero value is recorded when a metabolite is not detected in a sample.

The file contains 4 worksheets:

1. *Variable description* – this gives info on sample related variables and metabolite measurements
2. *Concentration* – this dataset contains absolute concentrations of 26 metabolites and semi-quantified concentrations of a further 18 metabolites.
3. *Creatinine norm* – this dataset is derived from worksheet 2 with concentration data normalised to creatinine concentration in each sample (ie presented as concentration per unit of creatinine).
4. *Median fold change norm* – this dataset is derived from worksheet 2 with concentration data normalised for sample dilution. Dilution factors were estimated using the distribution of the 44 metabolite concentrations and by identifying a median quotient for each sample. This spreadsheet contains one variable column specifying the estimates of urine sample dilution which have been used to compute the normalised data.

Other sample-related variables included in the data file:

Urine Sample ID	Sample name containing information on Helix Child ID, urine sampling type, cohort and panel information
Centre Cohort	BIB, EDP, MOB, KAN, RHE, SAB
Child_ID	HELIX subject number
SamplingType	Cux (Morning-Night pool sample), Mux (Morning sample), Nux (Night Sample)
RunOrder	Order used for sample preparation and instrument run; this has been randomised to prevent potential analytical batch effect
Estimate of urine sample dilution	Calculated based on fold changes of 44 urine metabolite concentrations compared to the median sample; higher values represent samples are more diluted

The *Concentration* data table represents the absolute concentration information we can extract from the urine samples, whilst *Creatinine norm* and *Median fold change norm* tables are provided as two different means of normalising the effect of dilution in urine. Creatinine has long been used as a marker for normalisation, but creatinine in urine is also found to be associated with children's age thus potentially exacerbating any effects between HELIX

cohorts. Median fold change normalisation takes into account the distribution of relative levels from all 44 metabolites compared to the reference sample in determining the most probable dilution factor. We recommend to use the *median fold change norm* data for discovery analysis, and use either the *concentration* data or the *creatinine norm* data to validate any findings.

Table 4. Measurement units in the urine metabolomics data tables.

	measurement units		
metabolite variables	***concentration***	***creatinine norm***	***median fold change norm***
Creatinine	umol/L	NA	AU
Leucine	umol/L	umol/mmol of creatinine	AU
Isoleucine	umol/L	umol/mmol of creatinine	AU
Valine	umol/L	umol/mmol of creatinine	AU
3-hydroxyisobutyrate	umol/L	umol/mmol of creatinine	AU
4-deoxyerythronic acid	umol/L	umol/mmol of creatinine	AU
4-deoxythreonic acid	umol/L	umol/mmol of creatinine	AU
Lactate	umol/L	umol/mmol of creatinine	AU
2-hydroxyisobutyrate	umol/L	umol/mmol of creatinine	AU
Alanine	umol/L	umol/mmol of creatinine	AU
Lysine	umol/L	umol/mmol of creatinine	AU
Acetate	umol/L	umol/mmol of creatinine	AU
p-cresol sulfate	umol/L	umol/mmol of creatinine	AU
Succinate	umol/L	umol/mmol of creatinine	AU
Glutamine	umol/L	umol/mmol of creatinine	AU
Citrate	umol/L	umol/mmol of creatinine	AU
Dimethylamine	umol/L	umol/mmol of creatinine	AU
Trimethylamine	umol/L	umol/mmol of creatinine	AU
Carnitine	umol/L	umol/mmol of creatinine	AU
Trimethylamine oxide	umol/L	umol/mmol of creatinine	AU
Scyllo-inositol	umol/L	umol/mmol of creatinine	AU
Taurine	umol/L	umol/mmol of creatinine	AU
Glycine	umol/L	umol/mmol of creatinine	AU
Creatine	umol/L	umol/mmol of creatinine	AU
Hippurate	umol/L	umol/mmol of creatinine	AU
Formate	umol/L	umol/mmol of creatinine	AU
Pantothenic acid	AU (proportional to concentration)	AU	AU
3-hydroxybutyrate/3-aminoisobutyrate	AU (proportional to concentration)	AU	AU
3-hydroxyisovalerate	AU (proportional to concentration)	AU	AU
N-acetyl neuraminic acid	AU (proportional to concentration)	AU	AU
Acetone	AU (proportional to concentration)	AU	AU
5-oxoproline	AU (proportional to concentration)	AU	AU
3-aminoisobutyrate	AU (proportional to concentration)	AU	AU
Proline betaine	AU (proportional to concentration)	AU	AU
Glucose	AU (proportional to concentration)	AU	AU
Sucrose	AU (proportional to concentration)	AU	AU
Urea	AU (proportional to concentration)	AU	AU
N-methyl-2-pyridone-5-carboxamide	AU (proportional to concentration)	AU	AU
p-hydroxyphenylacetate	AU (proportional to concentration)	AU	AU
Tyrosine	AU (proportional to concentration)	AU	AU
3-Indoxylsulfate	AU (proportional to concentration)	AU	AU
N-methylpicolinic acid	AU (proportional to concentration)	AU	AU
N-methylnicotinic acid	AU (proportional to concentration)	AU	AU
N1-methyl nicotinamide	AU (proportional to concentration)	AU	AU

Note - AU denotes arbitrary unit

Part 6 Preliminary analysis: Cohort effect

Principal Component Analysis (PCA) was performed on the 1366 unique children urinary ^{1}H NMR spectra using the 44 metabolite peak signals to access the correlation structure in the dataset. In this analysis, urinary metabolite concentrations were normalised using probabilistic quotient normalisation/ median fold change (Dieterle et al., 2006) followed by mean centring and unit-variance scaling. The data indicate that approximately 16% of the variance can be represented by the first 2 principal components (Figure 5). Cohort effect was difficult to visualise using PCA analysis.

We also computed spearman correlation coefficients between each metabolite and each cohort by splitting cohort into 6 dummy binary variables. Bonferroni corrections were used to account for multiple testing and significant associations between metabolite and cohort are shown in Figure 6.

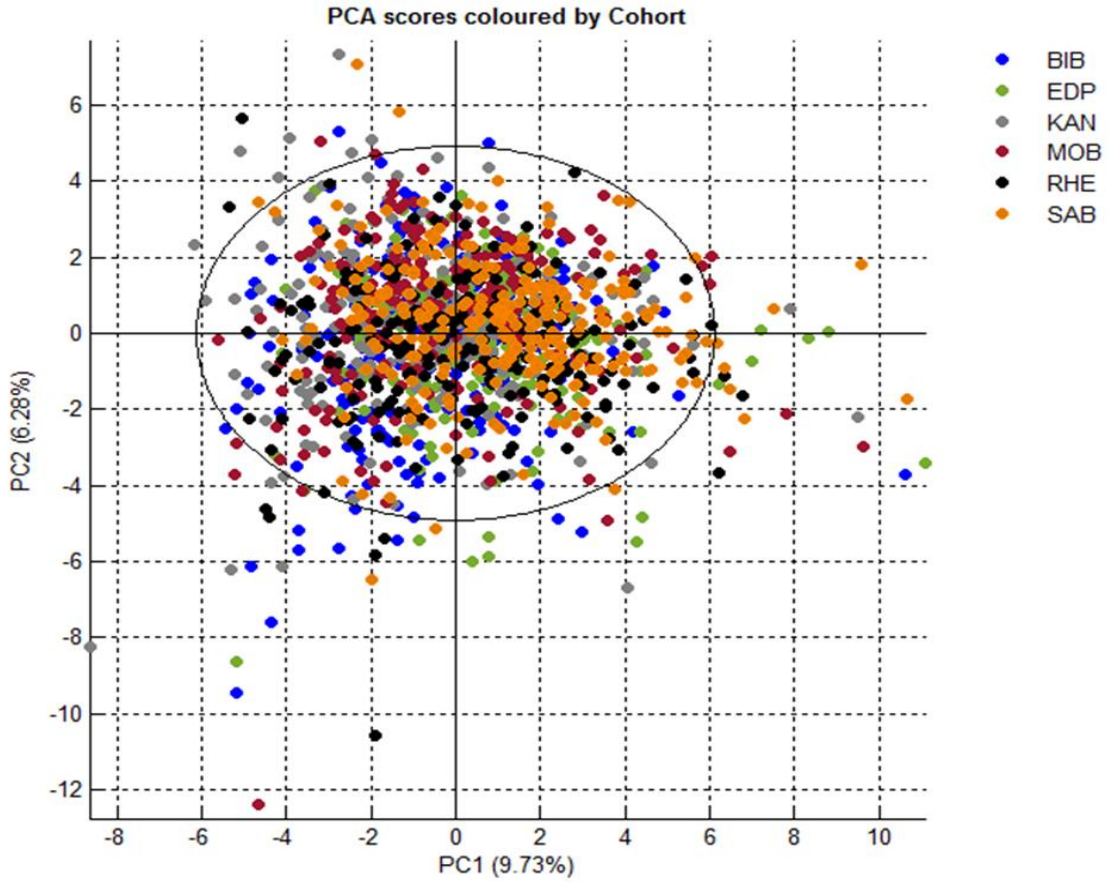

Figure 5. Score plot from preliminary Principal Component Analysis coloured by cohort

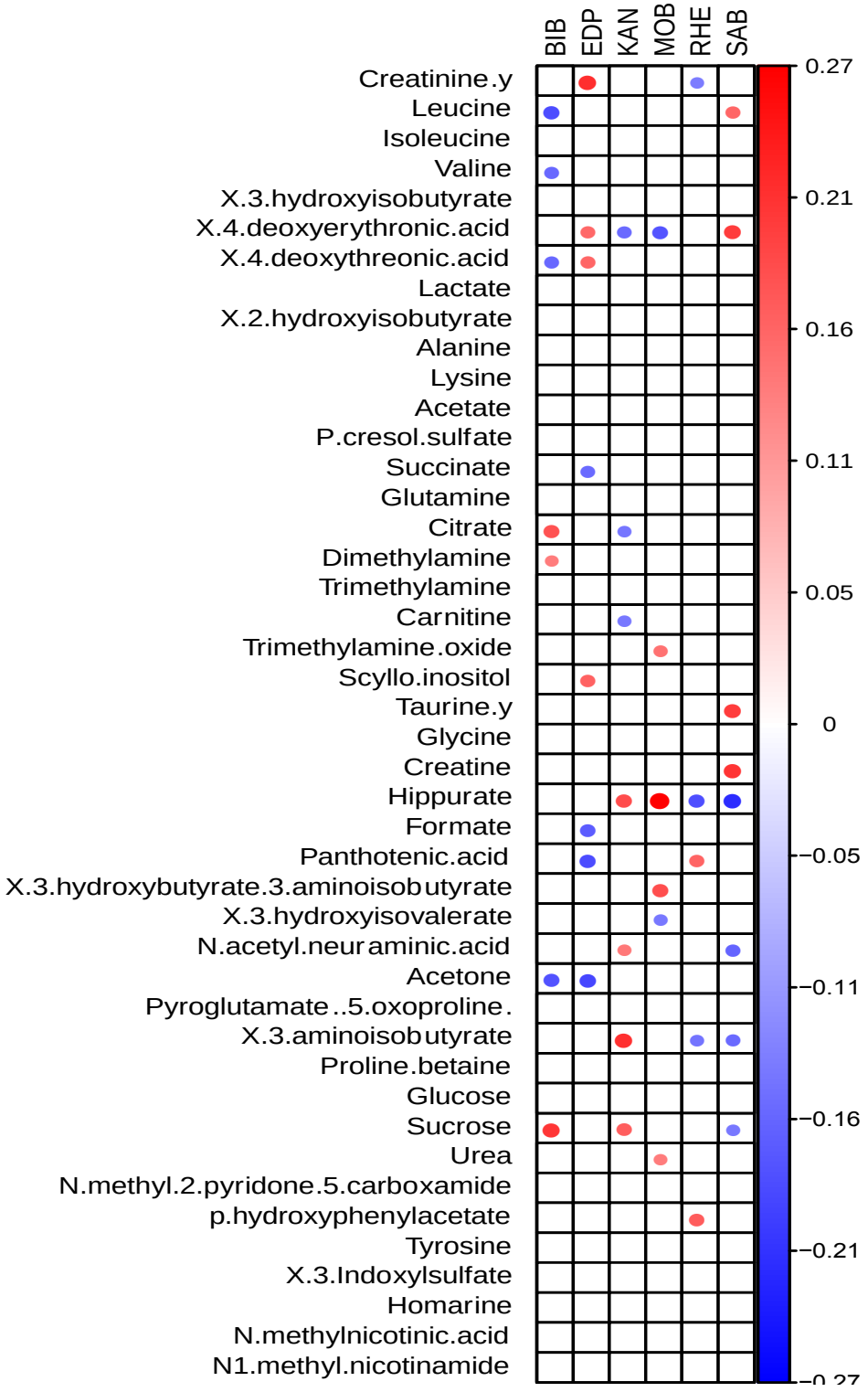

Figure 6. Pairwise spearman correlation coefficients between individual urine metabolites and cohorts

References

Cloarec, O., Dumas, M.E., Craig, A., Barton, R.H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J.C., Holmes, E*., et al.* (2005). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. Analytical Chemistry *77*, 1282-1289.

Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. Analytical Chemistry *78*, 4281-4290.

Tredwell, G.D., Behrends, V., Geier, F.M., Liebeke, M., and Bundy, J.G. (2011). Between-Person Comparison of Metabolite Fitting for NMR-Based Quantitative Metabolomics. Analytical Chemistry *83*, 8683-8687.

Veselkov, K.A., Lindon, J.C., Ebbels, T.M.D., Crockford, D., Volynkin, V.V., Holmes, E., Davies, D.B., and Nicholson, J.K. (2009). Recursive Segment-Wise Peak Alignment of Biological H-1 NMR Spectra for Improved Metabolic Biomarker Recovery. Analytical Chemistry *81*, 56-66.