Supplementary material: Integrating metabolomics profiling measurements across multiple biobanks

Authors: A.D. Dane[(*,1,2)],M.M.W.B. Hendriks[(*,1,2)], T.H. Reijmers[(1,2)], A.C. Harms[(1,2)], J. Troost[(1,2)], R.J. Vreeken[(1,2)], D. I. Boomsma[(3)], C.M van Duijn[(4)], E.P. Slagboom[(5)], T. Hankemeier [(1,2,+)]

[(*)] shared first author

[(+)] corresponding author

[1]Division Analytical Biosciences, Leiden Academic Center for Drug Research, Einsteinweg 55, 2333CC Leiden, The Netherlands
[2]Netherlands Metabolomics Centre, Einsteinweg 55, 2333CC Leiden, The Netherlands
[3]Department of Biological Psychology, VU University Amsterdam, The Netherlands
[4]Department of Epidemiology, Erasmus University Medical School, Rotterdam, The Netherlands
[5]Leids Universitair Medisch Centrum, Albinusdreef 2, 2333 ZA Leiden, The Netherlands

Short description of the cohorts

The Leiden Longevity Study (www.healthy-ageing.nl, www.langleven.net), which includes 3500 persons from families in which at least two long-lived siblings were alive and their parents were of Caucasian descent [Schoenmaker, 2005].

The Netherlands Twin Registry (NTR: www.tweelingenregister.org), which ascertains 2- and 3-generation Dutch families from the entire country based on the presence of twins or higher order multiples in the family and includes nearly 180,000 participants [van Beijsterveldt, 2012, Willemsen, 2012]. The NTR Biobank collected biological samples in >10,000 participants [Willemsen, 2010].

The Rotterdam studies, which consist of a population-based long term follow-up study including 11,800 persons from Rotterdam [Hofman, 2011] and its surrounding area (www.erasmus-epidemiology.nl/rotterdamstudy) and the Genetic Research in Isolated Populations program (GRIP: www.epib.nl/research/geneticepi/research.html#gip) targeting the South Western part of the Province North Brabant.

Table S1 presents a summary of the number of samples measured for each study.

study	batches	unique samples	within study replicates	LLS transfer samples	NTR transfer Samples	acquisition start	acquisition end
LLS	27	2331	448			March 2010	August 2010
NTR	35	2822	536	149*		May 2012	June 2012
ERF	36	2777	405	163*	165	August 2012	November 2012

*= 128 identical LLS samples were used in both NTR and ERF acquisition runs.

Table S1: summary of measurement characteristics

Validation

The quality of the transferred data was assessed on the relative standard deviations (RSDs) obtained from the transfer samples. The transfer RSDs are defined using the response values from the original measurement data after transfer and the response values of the same transfer samples in the reference dataset. The pooled standard deviation of paired replicates can be calculated as [Massart, 1997]:

$$s_p^2 = \frac{\sum_j \frac{d_j^2}{2}}{k}$$

with $d_j = x_{jt} - x_{jr}$ the difference between two measurements of a metabolite in the same (transfer) sample *j* both in the reference study and in the original study, the last measurement data transferred. The number of transfer samples is *k*.

The overall mean of the transfer samples of both the original and reference measurements is given by:

$$\bar{x} = \frac{\sum_j (x_{jt} + x_{jr})}{2k}$$

The RSD of the transfer samples is then calculated as:

$$\mathrm{RSD}_t = {}^{s_p}/_{\bar{x}}$$

Examples of transfer model results

Example results of the transfer models for the LLS and NTR data can be found in Figure S1. The first example (Figure S1a) shows the transfer samples for internal standard corrected ratios of PC(O-36:4) after transfer. The correlation between the samples is high; the resulting transfer RSD is approximately 10%, indicating that the transfer was successful. Figure S1b shows results from a triglyceride, TG(42:0), also with high correlation between IS ratios in the original LLS measurements and the transfer samples in the ERF measurements. The resultant transfer RSD, however, is very high (65%), but is mainly dominated by the outlier at the right bottom of the figure. Lipidomics measurements of triglycerides are known to be complicated; most of the chromatographic peaks are small and badly shaped.

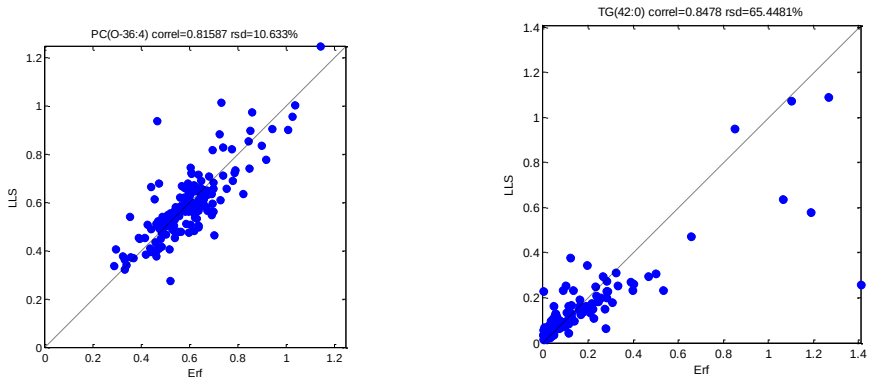


Figure S1 a) PC(O-36:4) IS corrected ratios of LLS transfer samples as measured in the ERF study versus IS corrected values of the same samples in the original LLS measurements. b) TG(42:0) IS corrected ratios of LLS transfer samples as measured in the ERF study versus IS corrected values of the same samples in the original LLS measurements.

Single study workflow

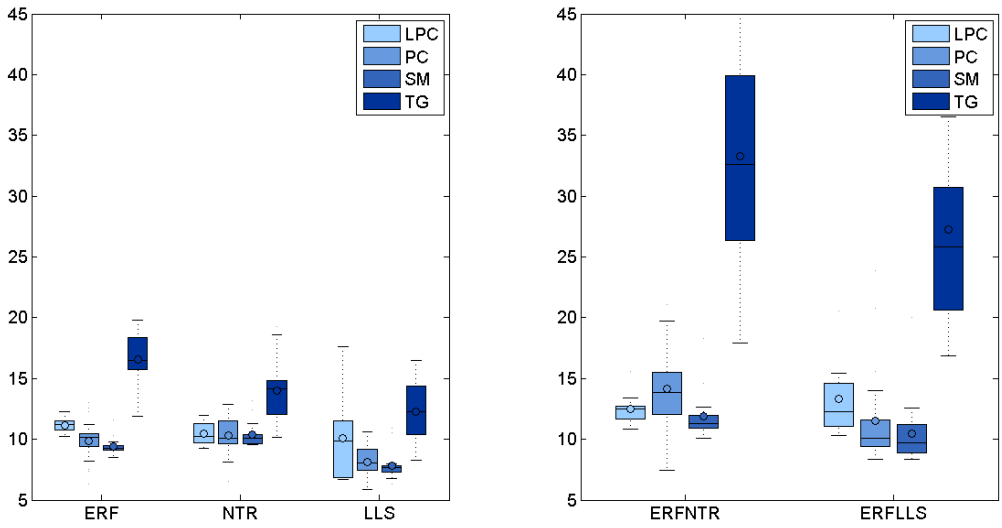
Data are processed in the following manner:

1. Per batch for each metabolite and internal standard the average area in the blanks (those blanks are added to the measurement design) is calculated. For each sample for each metabolite and internal standard this average blank area is subtracted from all peak areas.
2. Each blank corrected area is divided by the blank corrected internal standard peak area.
3. QC correction as described in [van der Kloet 2009] is applied. Briefly: in each batch a smoothed curve is fitted through the QC responses obtained after internal standard correction. Each sample response is then divided by the predicted QC response (predicted at the batch position of the sample) and multiplied by the median QC response in the reference batch.
4. Only metabolites where the RSD in both the QC samples and the replicates is smaller than 20% are reported (this is in line with the quality requirements as described in the standard operating procedures of our metabolomics laboratory).

Reproducibility after transfer

In Figure S2b, the distribution of the transfer RSD_t of all lipids, grouped by lipid class, are summarized. As a comparison, in Figure S2a the RSD_r calculated from replicates in the original measurements are presented. Although the transfer RSD_t , as expected, is always slightly higher than the within study RSD_r (based on within study replicate samples), they are approximately in the same range, for all lipid classes except the TGs. This indicates that with the additional transfer between studies the reproducibility is only slightly lowered. The TGs, however, often show transfer RSDs that are considerably higher than the within study RSDs. The high transfer RSDs are mostly due to one or a few badly performing samples.

Since TG peaks are often very small, these badly performing samples, may actually reflect integration errors.



FigureS2 a) Boxplots showing the distribution of within study RSDs grouped by lipid class. b) Boxplots showing the distribution of transfer RSDs grouped by lipid class. ERFNTR and ERFLLS results are direct results of the transfer models.

Paired principal component analysis of NTR transfer samples

Paired principal component analysis [van Velzen, 2008] can be applied on paired measurements (e.g. two measurements done on the same object/sample). By subtracting the mean of each paired measurement, inter-individual (inter-sample) variation is filtered out and the focus is on differences between the paired measurements. Figure S3 shows the result of paired PCA on the NTR transfer samples. The transfer samples are measured both during the NTR study measurements (red dots) and the ERF study measurements (blue dots). Figure S3a shows the PCA scores before transfer of the NTR study measurements, Figure S3b after the transfer of the NTR study measurements. Clearly it can be seen that before the transfer the samples are located in a different subspace of the two component PCA plane, while after the transfer of the NTR measurements, the clouds of blue and red dots are overlapping.

(note that the blue and red dots are mirrored, this is the effect of the pairwise centering step).

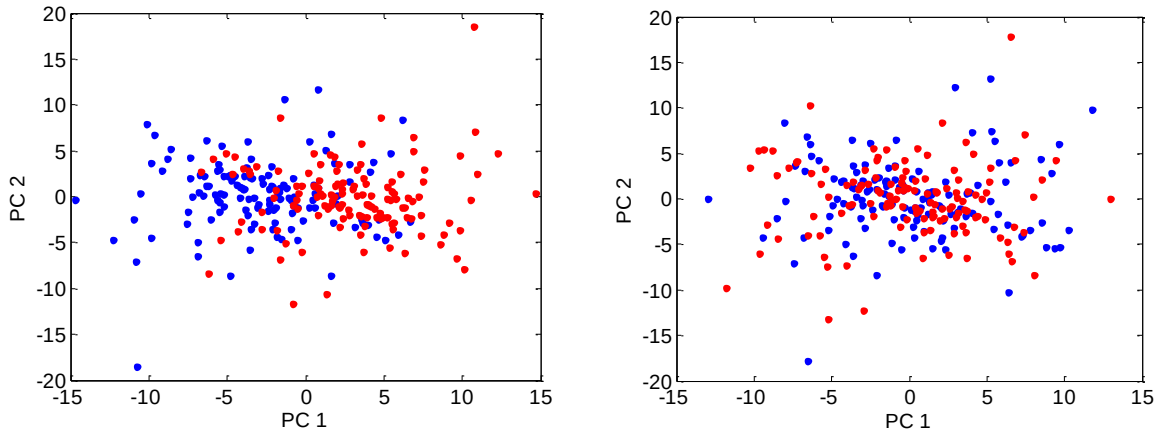


Figure S3 a) scores plot of a paired principal components analysis on two sets of IS ratios of NTR transfer samples. Blue dots are scores of NTR transfer samples as measured during the ERF measurements, red dots are scores of NTR transfer samples as measured during the NTR measurements) b) same as figure a), but now the red dots are scores of the NTR transfer samples measured in NTR and transferred to the ERF domain.

PLS-DA of ERF and transferred NTR data

Differences are found in the spatial distribution of the ERF samples and the NTR samples after transfer. To investigate if the data can reveal a plausible explanation for this observation a PLS-DA [van Velzen, 2008] model was built to investigate which lipids are mostly responsible for the difference between the two studies, ERF and NTR after transfer to the ERF domain. Double cross-validation revealed that a predictive PLS-DA model can easily be obtained with a very low misclassification error (8% misclassified, average number of latent variables used was 3). Figure S4a shows a plot of the selectivity ratio [Rajalahti, 2009] of the lipids in the three component PLS-DA model, Figure S4b a plot of the regression vector of the same model.

Figures S5a and S5b show the same plots, but now represent the results of the comparison of the ERF study samples with the NTR transfer samples as measured during the ERF study measurements. These samples were treated identically and measured at the same time. Remarkably, the model results are very stable and comparable to the results shown in Figures S4a and 4b. This indicates that i) the NTR transfer samples represent the total NTR population well and ii) the transfer model has performed well. The difference found between the NTR and ERF samples therefore has to be the result of either population differences or sample acquisition differences. Suppose that the differences are generated by sampling acquisition then we would expect systematic patterns, e.g whole classes of lipids that change together. This is however not what we observe, hence this explanation is less probable.

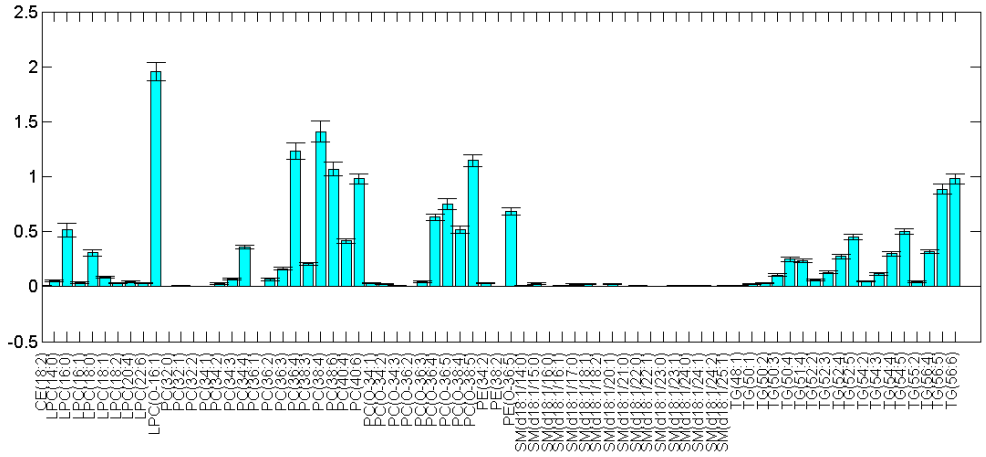


Figure S4a PLS-DA model results of comparison of ERF study samples with transferred NTR study samples. Average selectivity ratios with standard deviations resulting from 10-fold outer cross-validation loop.

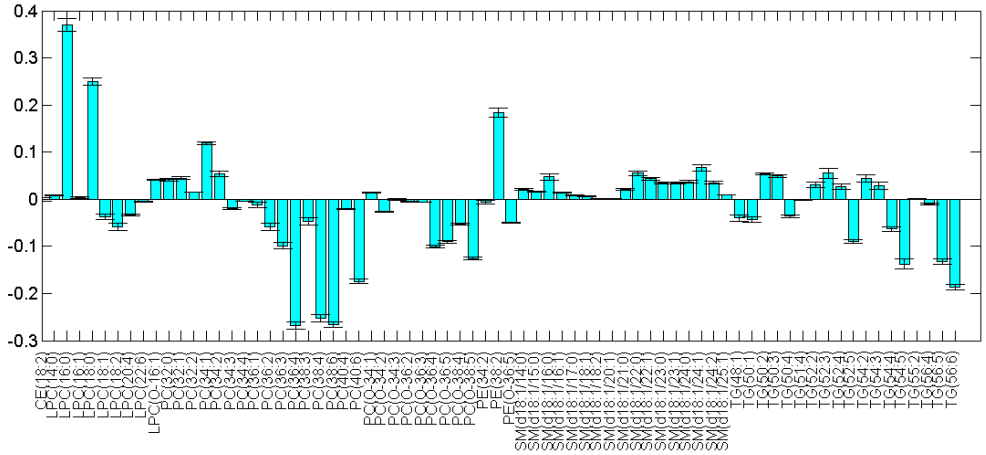


Figure S4b PLS-DA model results of comparison of ERF study samples with transferred NTR study samples. Average regression vectors with standard deviations resulting from 10-fold outer cross-validation loop.

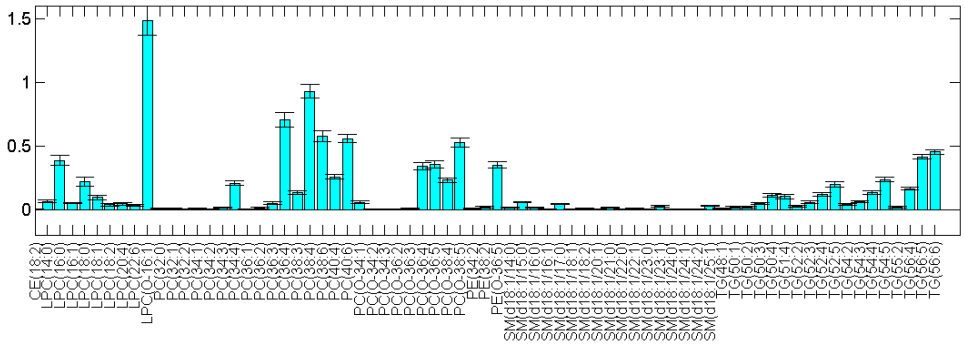


Figure S5a PLS-DA model results of comparison of ERF study samples with NTR transfer samples measured in the ERF study. Average selectivity ratios with standard deviations resulting from 10-fold outer cross-validation loop.

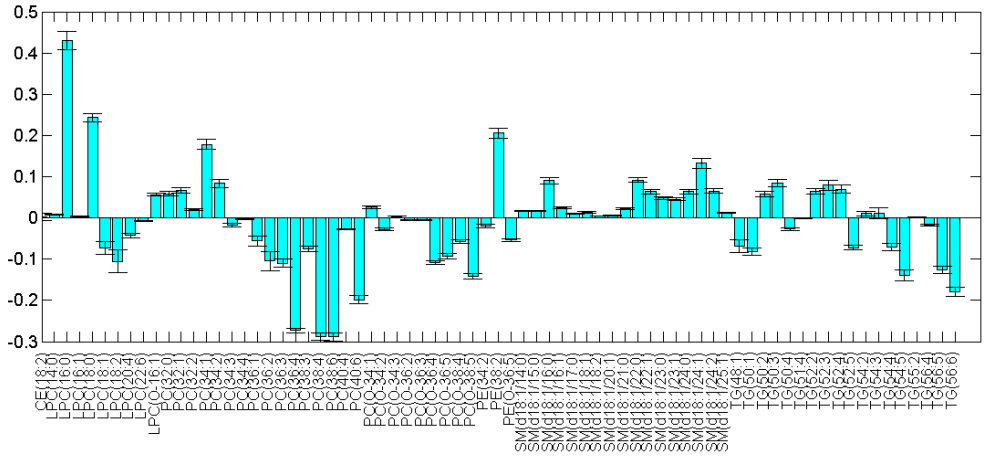


Figure S5b PLS-DA model results of comparison of ERF study samples NTR transfer samples. Average regression vectors with standard deviations resulting from 10-fold outer cross-validation loop.

References

[Schoenmaker, 2005] Schoenmaker M., A.J.M. de Craen , P.H.E.M. de Meijer *et al*, Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study, European Journal of Human Genetics, 14 (2005) 79-84

[van Beijsterveldt, 2012] van Beijsterveldt C.E.M., M. Groen-Blokhuis, J.J. Hottenga *et al*, The Young Netherlands Twin Register (YNTR): Longitudinal Twin and Family Studies in Over 70,000 Children, Twin Research and Human Genetics 1 (2012) 1-16

[Willemsen, 2012] Willemsen G., J.M. Vink, A. Abdellaoui *et al* The Adult Netherlands Twin Register: 25 years of survey and biological data collection, Twin Research and Human Genetics (2012): in press

[Willemsen, 2010] Willemsen G., E.J.C. de Geus, M. Bartels *et al*, The Netherlands Twin Register biobank: a resource for genetic epidemiological studies, Twin Research and Human Genetics, 13 (2010) 231

[Hofman, 2011] Hofman A., C.M. van Duijn O.H. Franco *et al*: The Rotterdam Study: 2012 objectives and design update, European Journal of Epidemiology, 26 (2011) 657-686

[Massart, 1997] Massart, D.L., B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, 1997, Elsevier, Amsterdam

[van der Kloet, 2009] van der Kloet, F.M., I. Bobeldijk, E.R. Verheij, R.H. Jellema, Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping, Journal of Proteome Research, 8 (2009) 5132-5141

[van Velzen, 2008] van Velzen, E.J.J., J.A. Westerhuis, J.P.M. van Duynhoven, F.A. van Dorsten, H.C.J. Hoefsloot, D.M. Jacob, S. Smit, R. Draijer C.I. Kroner and A.K. Smilde, Multilevel Data Analysis of a Crossover Designed Human Nutritional Intervention Study, Journal of Proteome Research, 7 (2008) 4483–4491

[Rajalahti, 2009] Rajalahti, T., R. Arneberg, A.C. Kroksveen, M. Berle, K.-M. Myhr, O.M. Kvalheim, Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles, Analytical Chemistry, 81 (2009) 2581-2590

	LLS				NTR				ERF			
	Mean	*SD*	*min*	*max*	*Mean*	*SD*	*min*	*max*	*Mean*	*SD*	*min*	*max*
Age (years)	59.17	6.85	30.17	80.31	38.60	12.70	18.1	79.6	48.72	14.28	16.65	86.50
Total cholesterol (mmol/L)	5.58	1.18	1.03	10.84	4.96	1.02	2.04	10.88	5.56	1.10	1.80	10.20
HDL cholesterol (mmol/L)	1.43	0.45	0.16	3.31	1.42	0.37	0.46	3.62	1.27	0.37	0.20	3.20
LDL cholesterol (mmol/L)	3.35	0.97	0.47	7.88	2.95	0.93	0.2	7.59	3.72	0.98	1.00	7.40
Triglycerides (mmol/L)	1.82	1.16	0.12	21.16	1.27	0.71	0.1	9.86	1.35	0.79	0.00	7.50
Glucose (mmol/L)	5.92	1.5	2.5	26.3	5.34	0.83	2.7	21.6	4.66	1.14	1.90	15.90
Systolic blood pressure (mmHg)	142.9	20.62	97.75	220.5					140.14	20.31	85.50	239.00
Body mass index (kg/m^2)	25.4	3.57	16.33	46.78	24.30	4.04	14.6	50.7	26.83	4.65	15.54	61.80
	N	*%*			*N*	*%*			*N*	*%*		
Men	999	45.39			956	33.9			1298			
Current smoking					588	20.9			1131	39.30		
Type 2 diabetes	103	4.68			113	4			195	6.80		
lipid medication	167	7.59			90	3.19			366	12.70		

Table S2 Study population characteristics