molecular systems biology

Novel biomarkers for pre-diabetes identified by metabolomics

Rui Wang-Sattler, Zhonghao Yu, Christian Herder, Ana C. Messias, Anna Floegel, Ying He, Katharina Heim, Monica Campillos, Christina Holzapfel, Barbara Thorand, Harald Grallert, Tao Xu, Erik Bader, Cornelia Huth, Kirstin Mittelstrass, Angela Döring, Christa Meisinger, Christian Gieger, Cornelia Prehn, Werner Roemisch-Margl, Maren Carstensen, Lu Xie, Hisami Yamanaka-Okumura, Guihong Xing, Uta Ceglarek, Joachim Thiery, Guido Giani, Heiko Lickert, Xu Lin, Yixue Li, Heiner Boeing, Hans-Georg Joost, Martin Hrabé de Angelis, Wolfgang Rathmann, Karsten Suhre, Holger Prokisch, Annette Peters, Thomas Meitinger, Michael Roden, H.-Erich Wichmann, Tobias Pischon, Jerzy Adamski, Thomas Illig

Corresponding author: Rui Wang-Sattler, Helmholtz Zentrum Muenchen

Review timeline:

Submission date:	18 June 2012
Editorial Decision:	12 July 2012
Revision received:	30 July 2012
Editorial Decision:	07 August 2012
Revision received:	14 August 2012
Accepted:	15 August 2012

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 12 July 2012

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees who agreed to evaluate your manuscript. As you will see from the reports below, the referees find the topic of your study of potential interest. They raise, however, substantial concerns on the work, which should be convincingly addressed in a revision.

The reviewers recognize that the reported findings are of potential interest and the topic of the study interesting. While reviewer #2 is positive, reviewer #1 is more reserved. This reviewer raises serious concerns with the presentation of the data. It is thus essential that the results are described in clearer and much more rigorous way. The recommendation provided by reviewer #1 are important and very clear in this regard.

Please note our policy on *data availability* (http://www.nature.com/msb/authors/index.html#a3.5) and include a section entitled "Data availability" in Materials and Method that provides the links or accession numbers to the metabolomic and functional genomics data reported in this study. If no public resource is appropriate, please include the data in supplementary information in a format that allows others to reproduce and build upon your analysis. As far as possible, the associated clinical and physiological parameters should also be made available, provided it remains compatible with the individual consent agreement used in this study.

As a matter of course, please make sure that you have correctly followed the instructions for authors as given on the submission website.

If you feel you can satisfactorily deal with these points and those listed by the referees, you may wish to submit a revised version of your manuscript. Please attach a covering letter giving details of the way in which you have handled each of the points raised by the referees. A revised manuscript will be once again subject to review and you probably understand that we can give you no guarantee at this stage that the eventual outcome will be favorable.

Yours sincerely,

Editor
Molecular Systems Biology

http://www.nature.com/msb

REFEREE REPORTS

Reviewer #1

The manuscript, "Novel Biomarkers for pre-diabetes identified by metabolomics" details a relatively large scale study of targeted metabolomics in two population based datasets and identifies 3 novel biomarkers that are associated with impaired glucose tolerance and predict future development of impaired glucose tolerance.

Overall, the findings are of potential interest, and the study question is certainly of great interest. The combined use of a targeted metabolomic approach but with a large number of metabolites measures is a strength, and the group is very capable of such laboratory analyses. Importantly, the authors utilize a validation cohort after discovery of the 3 metabolites from the discovery KORA dataset. Also importantly, the authors assess not only independent but incremental capabilities of these biomarkers with relation to clinical risk factors. Finally, the integration of the gene expression data to elucidate potential pathways mediating IGT risk is an excellent addition to the manuscript and helps to take this study beyond just biomarker discovery. I believe that analyses themselves are probably sound, however, for reasons detailed below, it is difficult to completely assess this.

The major weakness of the paper is the fashion in which it is presented. In addition to significant grammatical and some spelling errors throughout the manuscript, the study design is difficult to follow and the results themselves are difficult to follow. The figures are complicated and busy and likewise are difficult to follow. These issues make it hard to decipher the quality and importance of the science.

Some specific questions/issues that arose include:
1) Were any of the individuals taking diabetes medications besides the 2/49 individuals using insulin in KORA?
2) How was "nondiabetic" at baseline defined? Clinically? Or based on the glucose tolerance test?
3) I appreciate that Figure 1 was an attempt to visually represent the complicated study design, however, it did not succeed in that endeavor. Since the KORA S4/F4 samples are analyzed together, perhaps Fig 1a and 1B could be combined. In the results section under "Study participants" (which is very difficult to follow), perhaps first delineating the different phenotypes and then stating the sample sizes would simplify things.
4) For Figure 2, its not clear why the first figure presented (2a) only shows 3 metabolites when this is the first time we are seeing results for the 26 metabolites that were significant in any comparison. Figure 2B should be presented first (and why are only a subset of the 26 metabolites detailed in figure 2b)? Results for each of the 26 metabolites also need to be presented, even if just in the supplement.
5) It is not clear what the results of the random forest analyses were. Did these 3 metabolites emerge from those analyses, or from the stepwise analyses?
6) The definitions for "incident IGT and T2D" need to be included.

7) Presumably for each set of different analyses (i.e. cross sectional vs. incident), all the metabolites were analyzed, but only results for the 3 metabolites are presented.
8) For the incident IGT/T2D analyses, glycine and LPC were combined. It is not clear if this was an a priori defined analysis, and why that was only done for the incident analyses?
9) How were NGT controls defined?
10) The increases in the AUC are not impressive and of doubtful clinical significance. This needs to be noted in the manuscript, and p-values for the comparison of the AUC included.
11) For the five pairwise comparisons, were individuals with known T2D included (it is not clear as the manuscript states the comparisons were with dT2D). If not, why?
12) Did none of the KORA S4 individuals develop T2D or IGT (based on figure 1C)?
13) Could the authors please comment on the potential influence of diet on the 3 key metabolites they have discovered. While all the KORA samples were fasting, the effects of chronic diet presumably may have influenced metabolite levels.
14) While the metabolite-protein network analyses are an excellent addition to the manuscript, its not clear if analyses were done specifically assessing the association between the 3 key metabolites with expression levels of the 4 enzymes?
15) Were the gene expression samples taken at the same time as the metabolomics samples?
16) Why didn't any of the gene expression cohort have already diagnosed T2D?
17) Why did they analyze risk for each SD increase for the metabolites for the incident IGT/T2D analyses but (presumably) presented a per unit increase for the metabolites for the cross sectional analyses?
18) How did the actual levels of the key metabolites compare in KORA vs. ERIC-POTSDAM cohorts?
19) In Table 3, are the OR for a per unit increase in each metabolite? Also, units for the metabolites need to be added.
20) Were the analyses of incident IGT and T2D also additionally adjusted for fasting glucose, insulin and HgA1c like the cross sectional analyses were?

The editorial changes that need to be made include:
1) Attention to spelling and grammatical mistakes and presentation style.
2) Define acronyms the first time they are used (i.e. NGT and i-IFG are presented for the first time on page 5 without any definition)
3) Throughout the manuscript the terms "increased significantly" or "decreased significantly" are used when referring to the comparisons of NGT to IFG to IGT, etc. This makes it seem that there is a change in the actual metabolite levels as an individual transitions from one state to the next. In fact, the proper term for these cross sectional analyses would be "XX levels were higher in individuals with IGT as compared with NGT" (for example).
4) Page 8, 2nd paragraph, last line: I believe this should be referring to Table 4, not Table 3.
5) Page 8, last paragraph, instead of Table S8, I believe it should be Table S6.
6) There needs to be a concluding paragraph to tie the discussion together.
7) The authors should specify that the 2 hr glucose values are after a glucose challenge (currently the methods just state "based on both fasting and 2 -h glucose levels).
8) Define how HOMA-B and HOMA-IR were calculated.

Reviewer #2

The manuscript "Novel biomarkers for pre-diabetes identified by metabolomics" was interesting to read and obviously a large effort in data acquisition and analysis. I have suggested some revisions below as well as some minor corrections and clarifications to the text.

1. Perhaps it was overlooked but could the authors please describe briefly the origin of the KORA cohort as done for the EPIC-Potsdam cohort starting on p. 20.

2. How do the identified metabolites compare to the standards for diagnosis of T2D?

3. On p. 7, the authors mention receiver-operating-characteristic curves and provide a table in the supplementary data. It is suggested that the ROC be provided as well at least to compare the best models to the clinical standards for T2D.

4. In the first paragraph on p. 9, the authors note that they evaluated the predictive value of the five branched-chain and aromatic amino acids reported by Want et al. and found no significant changes. Could the authors speculate as to why it is thought that similar changes were not noted?

5. Only LPC (18:2) is reported to be significant in differentiating the groups. Were other LPC's altered? If not, why is 18:2 more important? Same question for the carnitines; were any others significantly altered?

6. For Figure 2 and Figure S1, how are the adjustments done?

7. For Supplementary Table S5, in the Metabolites combined with T2D risk indicators, why was C2 not included?

8. Table S14 lists the 188 targeted metabolites; what were the criteria used to decide whether a metabolite was used or excluded in the analysis?

1st Revision - authors' response 30 July 2012

Reviewers' comments:

Reviewer #1

The manuscript, "Novel Biomarkers for pre-diabetes identified by metabolomics" details a relatively large scale study of targeted metabolomics in two population based datasets and identifies 3 novel biomarkers that are associated with impaired glucose tolerance and predict future development of impaired glucose tolerance.

Overall, the findings are of potential interest, and the study question is certainly of great interest. The combined use of a targeted metabolomic approach but with a large number of metabolites measures is a strength, and the group is very capable of such laboratory analyses. Importantly, the authors utilize a validation cohort after discovery of the 3 metabolites from the discovery KORA dataset. Also importantly, the authors assess not only independent but incremental capabilities of these biomarkers with relation to clinical risk factors. Finally, the integration of the gene expression data to elucidate potential pathways mediating IGT risk is an excellent addition to the manuscript and helps to take this study beyond just biomarker discovery. I believe that analyses themselves are probably sound, however, for reasons detailed below, it is difficult to completely assess this.

The major weakness of the paper is the fashion in which it is presented. In addition to significant grammatical and some spelling errors throughout the manuscript, the study design is difficult to follow and the results themselves are difficult to follow. The figures are complicated and busy and likewise are difficult to follow. These issues make it hard to decipher the quality and importance of the science.

Reply: We thank this reviewer for her/his overall appreciation of our study and the suggestions she/he has provided. In this revised version, we have removed some results, rewritten parts of the text, and simplified the figures. The current version of our manuscript has been additionally revised by a professional English editor.

Some specific questions/issues that arose include:
1) Were any of the individuals taking diabetes medications besides the 2/49 individuals using insulin in KORA?

Reply: We agree that this is an important aspect. We have now excluded all known type 2 diabetes (T2D) individuals and analyzed only newly-diagnosed T2D patients who do not take any anti-diabetic medication. About 73% of known T2D patients in KORA take medications, in contrast to the newly-diagnosed ones (dT2D), many of whom do not even know about their illness during the course of investigation. Accordingly, we removed the Supplementary Table S4, Figures 2E, 2F and Supplementary Figures S1A, S1C, and we modified the new Supplementary Figure S1D.

2) How was "nondiabetic" at baseline defined? Clinically? Or based on the glucose tolerance test?

Reply: All the classifications of the individuals related to T2D were defined based on fasting and 2-h glucose values according to the WHO diagnostic criteria. We now provide definitions for "nondiabetic", "incident IGT" and "incident T2D". As we explained in the revised manuscript, in page 5, line 12: "*Based on both fasting and 2-h glucose values (i.e. two h post oral 75 g glucose load), individuals were defined according to the WHO diagnostic criteria to have normal glucose tolerance (NGT), isolated IFG (i-IFG), IGT or newly-diagnosed T2D (dT2D) (Meisinger et al, 2010; Rathmann et al, 2009; WHO, 1999) (Supplementary Table S1). The sample sets include 91 newly-diagnosed T2D patients and 1206 individuals with non-T2D, including 866 participants with NGT, 102 with i-IFG and 238 with IGT, in the cross-sectional KORA S4 survey (Figure 1A; study characteristics are shown in Table 1).*" ; and on page 5, line 21: "*Out of these, about 10% developed T2D (i.e. 91 incident T2D) (Figure 1C). From the 641 individuals with NGT at baseline, 18% developed IGT (i.e. 118 incident IGT) seven years later (Figure 1D)*".

3) I appreciate that Figure 1 was an attempt to visually represent the complicated study design, however, it did not succeed in that endeavor. Since the KORA S4/F4 samples are analyzed together, perhaps Fig 1a and 1B could be combined. In the results section under "Study participants" (which is very difficult to follow), perhaps first delineating the different phenotypes and then stating the sample sizes would simplify things.

Reply: To better illustrate our study design, we removed Figure 1B and modified Figure 1C accordingly. The section "Study participants" was rewritten: "*Individuals with known T2D were identified by physician-validated self-reporting (Rathmann et al, 2010) and excluded from our analysis, to avoid potential influence from anti-diabetic medication with non-fasting participants and individuals with missing values (Figure 1A). Based on both fasting and 2-h glucose values (i.e. two h post oral 75 g glucose load), individuals were defined according to the WHO diagnostic criteria to have normal glucose tolerance (NGT), isolated IFG (i-IFG), IGT or newly-diagnosed T2D (dT2D) (Meisinger et al, 2010; Rathmann et al, 2009; WHO, 1999) (Supplementary Table S1). The sample sets include 91 newly-diagnosed T2D patients and 1206 individuals with non-T2D, including 866 participants with NGT, 102 with i-IFG and 238 with IGT, in the cross-sectional KORA S4 survey (Figure 1A; study characteristics are shown in Table 1). Of the 1010 individuals in a fasting state who participated at baseline and follow-up surveys (Figure 1B), 876 of them were non-diabetic at baseline. Out of these, about 10% developed T2D (i.e. 91 incident T2D) (Figure 1C). From the 641 individuals with NGT at baseline, 18% developed IGT (i.e. 118 incident IGT) seven years later (Figure 1D). The study characteristics of the prospective KORA S4 → F4 are shown in Table 2.*"

4) For Figure 2, its not clear why the first figure presented (2a) only shows 3 metabolites when this is the first time we are seeing results for the 26 metabolites that were significant in any comparison. Figure 2B should be presented first (and why are only a subset of the 26 metabolites detailed in figure 2b)? Results for each of the 26 metabolites also need to be presented, even if just in the supplement.

Reply: We agree with the reviewer that the information about metabolites other than the three markers are also important. We removed Figure 2a and added a new Supplementary Table S4 to show the odds ratios (ORs) and *P*-values for the additional 23 metabolites of five pair-wise comparisons with two different models.

5) It is not clear what the results of the random forest analyses were. Did these 3 metabolites emerge from those analyses, or from the stepwise analyses?

Reply: The 3 metabolites were the result from both methods. For the nine significantly different metabolites between IGT and NGT with full adjustment of model 2, random forest was applied first in the selection process, and the stepwise selection was used based on the result from random forest. The intermediate results from random forest were included in this revised manuscript. In page 7, line 21, we added the following sentence: "*Out of the nine metabolites, five molecules (i.e. glycine, LPC*

(18:2), LPC (17:0), LPC (18:1) and C2) were select after random forest, and LPC (17:0) and LPC (18:1) were then removed after the stepwise selection."

6) The definitions for "incident IGT and T2D" need to be included.

Reply: We have now specified the "incident IGT and T2D" in the new Figures 1C and 1D and added the definitions. See reply to point 2) above.

7) Presumably for each set of different analyses (i.e. cross sectional vs. incident), all the metabolites were analyzed, but only results for the 3 metabolites are presented.

Reply: We did cross-sectional analysis for each of the 140 metabolites and identified 3 metabolites. The predictive values of the 3 metabolites were further explored in the prospective analysis as well as in the replication study. In addition to the 3 metabolites presented in Table 3, we added additional 23 metabolites in the Supplementary Table S4, in which the metabolites of the five pair-wise comparisons with two models are presented.

8) For the incident IGT/T2D analyses, glycine and LPC were combined. It is not clear if this was an a priori defined analysis, and why that was only done for the incident analyses?

Reply: We would like to thanks the reviewer for this suggestion. We have now used the combination of all the three identified metabolites (namely, glycine, LPC (18:2) and C2) in the prospective analysis. In our revised version, we state in page 8, line 19: "*Each standard deviation (SD) increment of the combinations of the three metabolites was associated with a 33% decreased risk of future diabetes (OR = 0.39 (0.21-0.71), P = 0.0002).*" and in page 9, line 3: "*When the three metabolites were added to the fully adjusted model 2, the area under the receiver-operating-characteristic curves (AUC) increased 2.6% (P = 0.015) and 1% (P = 0.058) for IGT and T2D, respectively (Supplementary Figure 2, Supplementary Table S7).*" In addition,Table 4, Supplementary Table S7 and Supplementary Figure 2 were changed accordingly.

9) How were NGT controls defined?

Reply: Please see reply to 2) above.

10) The increases in the AUC are not impressive and of doubtful clinical significance. This needs to be noted in the manuscript, and p-values for the comparison of the AUC included.

Reply: We added *P*-values from the likelihood ratio test to indicate the improvement by adding the metabolites to the known T2D risk in the Supplementary Figure S2. They are significant at the 5% level and are predictive with the exception of T2D in model 2. The relatively small increases in the AUC are discussed in page 13, line 8: *"...only 0.4% improvement from the T2D risk indicators as reported in the Framingham Offspring Study) (Wang et al, 2011). ... In contrast, we found that combined glycine, LPC (18:2) and C2 have 2.6% and 1% increment in predicting IGT and T2D in addition to the common risk indicators of T2D. This suggests they are better candidate for early biomarkers, and specifically from NGT to IGT, than the five amino acids.*"

11) For the five pairwise comparisons, were individuals with known T2D included (it is not clear as the manuscript states the comparisons were with dT2D). If not, why?

Reply: To avoid medication effect, we have now excluded known T2D patients. Please see reply to 1) above.

12) Did none of the KORA S4 individuals develop T2D or IGT (based on figure 1C)?

Reply: We modified Figures 1C and 1D to indicate more clearly the incident IGT/T2D individuals. We also added a new description on page 5, line 20: "*…876 of them were non-diabetic at baseline. Out of these, about 10% developed T2D (i.e. 91 incident T2D) (Figure 1C). From the 641 individuals with NGT at baseline, 18% developed IGT (i.e. 118 incident IGT) seven years later (Figure 1D)*".

13) Could the authors please comment on the potential influence of diet on the 3 key metabolites they have discovered. While all the KORA samples were fasting, the effects of chronic diet presumably may have influenced metabolite levels.

Reply: The potential influence of diet on metabolite concentration profiles is a very complex issue. It has been reported that unhealthy eating causes up to one quarter of chronic diseases, including T2D. Our previous study of diet impact on metabolome revealed that the nutrition status with polyunsaturated fatty acids associates with a decrease in saturation of the fatty acid chains of glycero-phosphatidylcholines (Altmaier *et al*, 2011). We added two sentences in the discussion, in page 15, line 23, to address this limitation, as suggested by the reviewer: "*Moreover, the influence from long-term dietary habits should not be ignored, even though we used only serum from fasting individuals (Altmaier et al, 2011; Primrose et al, 2011)*".

14) While the metabolite-protein network analyses are an excellent addition to the manuscript, its not clear if analyses were done specifically assessing the association between the 3 key metabolites with expression levels of the 4 enzymes?

Reply: We started the analysis with the 3 key metabolites and the 46 T2D related genes. We then used public available data bases (HMDB and STRING), as stated in page 21, line 23: "*These enzymes were connected to the 46 T2D-related genes (considered at that point), allowing for one intermediate protein (other proteins) through STRING protein functional interaction and optimized by eliminating edges with a STRING score below 0.7 and undirected paths. The sub-networks were connected by the shortest path from metabolites to T2D-related genes.*"

15) Were the gene expression samples taken at the same time as the metabolomics samples?
16) Why didn't any of the gene expression cohort have already diagnosed T2D?

Reply: The samples for both gene expression and metabolomics analysis were taken at the same time. We rephrased the sentence at page 18 line 11 to: "*Peripheral blood was drawn under fasting conditions from 599 KORA S4 individuals at the same time as the serum samples used for metabolic profiling were prepared.*" We excluded the known T2D patients from the gene expression analysis as we did for the metabolomics analysis. We modified the descriptive sentence in page 18, line 22, to make this clearer: "*The sample sets comprised 383 individuals with NGT, 104 with IGT and 26 with dT2D. The known T2D individuals were removed as had been done for the metabolomics analysis.*

17) Why did they analyze risk for each SD increase for the metabolites for the incident IGT/T2D analyses but (presumably) presented a per unit increase for the metabolites for the cross sectional analyses?

Reply: We added the unit per SD for all ORs in all Tables. We also stated in the method section (page 19, line 13) : "*…all reported OR values correspond to the change per SD of metabolite concentration.*"

18) How did the actual levels of the key metabolites compare in KORA vs. ERIC-POTSDAM cohorts?

Reply: The actual levels of the metabolites have been added to the Supplementary Tables S15 and S12 for the KORA and EPIC-Potsdam cohorts, respectively. They are not exactly the same, as we now discuss on page 10, line 25 "*The absolute levels of these three metabolites were in a similar* *range, with only slight differences that were due probably to the differences of the two cohorts or to potential batch effects of metabolomics measurements (Supplementary Table S12 and S15)".*

19) In Table 3, are the OR for a per unit increase in each metabolite? Also, units for the metabolites need to be added.

Reply: We have now added the units for the metabolites in all Tables; see reply to point 17 above.

20) Were the analyses of incident IGT and T2D also additionally adjusted for fasting glucose, insulin and HgA1c like the cross sectional analyses were?

Reply: The full adjusted model (including HbA_{1c}, fasting glucose and fasting insulin) was also conducted in the revised version of manuscript. We added a Supplementary Table S6 to report the ORs and *P*-values for the 3 metabolites in predicting IGT and T2D. We also added one sentence in page 9, line 1, which stated: "*With the full adjusted model 2, consistent results were obtained for LPC (18:2) but not for glycine (Supplementary Table S6)."*

The editorial changes that need to be made include:
1) Attention to spelling and grammatical mistakes and presentation style.
2) Define acronyms the first time they are used (i.e. NGT and i-IFG are presented for the first time on page 5 without any definition)
3) Throughout the manuscript the terms "increased significantly" or "decreased significantly" are used when referring to the comparisons of NGT to IFG to IGT, etc. This makes it seem that there is a change in the actual metabolite levels as an individual transitions from one state to the next. In fact, the proper term for these cross sectional analyses would be "XX levels were higher in individuals with IGT as compared with NGT" (for example).
4) Page 8, 2nd paragraph, last line: I believe this should be referring to Table 4, not Table 3.
5) Page 8, last paragraph, instead of Table S8, I believe it should be Table S6.
6) There needs to be a concluding paragraph to tie the discussion together.
7) The authors should specify that the 2 hr glucose values are after a glucose challenge (currently the methods just state "based on both fasting and 2 -h glucose levels).
8) Define how HOMA-B and HOMA-IR were calculated.

Reply: We thank this reviewer's carefully reading of our manuscript. We have considered all his/her comments and have revised accordingly. In general, we simplified our manuscript by removing the HOMA-B and HOMA-IR data, and by only showing the fasting insulin values. Finally, we added new conclusion paragraph "*Three novel metabolites, glycine, LPC (18:2) and C2, were identified as pre-diabetes specific markers. Their changes might precede other branched-chain and aromatic amino acids markers in the progression of T2D. Combined levels of glycine, LPC (18:2) and C2 can predict risk not only for IGT but also for T2D. Targeting the pathways that involve these newly-proposed potential biomarkers would help to take preventive steps against T2D at an earlier stage*".

Reviewer #2

The manuscript "Novel biomarkers for pre-diabetes identified by metabolomics" was interesting to read and obviously a large effort in data acquisition and analysis. I have suggested some revisions below as well as some minor corrections and clarifications to the text.

1. Perhaps it was overlooked but could the authors please describe briefly the origin of the KORA cohort as done for the EPIC-Potsdam cohort starting on p. 20.

Reply: We thank the reviewer for this suggestion. We have now provided a description of the origin of the KORA cohort, as stated in page 16 of the "Sample Source and classification" section: "*The Cooperative Health Research in the Region of Augsburg (KORA) surveys are population-based studies conducted in the city of Augsburg and the surrounding towns and villages (Holle et al, 2005; Wichmann et al, 2005). KORA is a research platform in the field of epidemiology, health economics* *and health care research. Four surveys were conducted with 18 079 participants recruited from 1984 to 2001. The survey 4 (S4) consists of 4261 individuals (aged 25-74 years) examined from 1999 to 2001.*"

2. How do the identified metabolites compare to the standards for diagnosis of T2D?

Reply: We compared the correlation between the three metabolites (along with others) and the diagnosis standards (e.g. fasting glucose values and 2-h glucose values). Glycine and LPC (18:2) showed a negative correlation, while C2 showed a positive correlation (Supplementary Table S3). Moreover, we also calculated the improvement of the predictive value when these metabolite markers were combined with common risk factors. In fact, the AUC value for common risk factors is already high. Nevertheless, our metabolite markers still improved this (although with a modest value from 0.818 to 0.828 in our prospective study). Along with the fact that these markers have significantly different concentrations even in the very stringent model 2, we believe that these markers may be affected via different mechanisms or pathways distinct from common risk factors. The new markers can thus help us to better understand the underlying mechanisms of T2D. Moreover, these markers may precede the clinical metrics in the development of T2D and thereby serve as early markers for the disease.

3. On p. 7, the authors mention receiver-operating-characteristic curves and provide a table in the supplementary data. It is suggested that the ROC be provided as well at least to compare the best models to the clinical standards for T2D.

Reply: We provided the ROC curves in a new Supplementary Figure S2 in this revised manuscript. In this figure, we also showed the *P*-values from the likelihood ratio test to indicate the improvement by adding the metabolite markers to the clinical standards (i.e. known T2D risk indicators).

4. In the first paragraph on p. 9, the authors note that they evaluated the predictive value of the five branched-chain and aromatic amino acids reported by Want et al. and found no significant changes. Could the authors speculate as to why it is thought that similar changes were not noted?

Reply: The main difference between our current study and the study published by Wang *et al.* is that they are based on different study designs, which in turn leads to different results. As we discussed in subsection "Different study designs reveal progression of IGT and T2D", on Page 13, line 1, "*Our analysis could replicate four out of the five branched-chain and aromatic amino acids recently reported to be predictors of T2D using nested/selected case-controls samples (Wang et al, 2011).*" and *"…these markers were unable to be extended to the general population (with only 0.4% improvement from the T2D risk indicators as reported in the Framingham Offspring Study) (Wang et al, 2011)* ". The results of the two study are different yet do not contradict each other. Supplementary Figure S1D displayed clearly that these differences in our result might as well reflect the progression of the T2D (i.e., the changes of our metabolite markers occur at an earlier stage, while changes in the branched-chain and aromatic amino acids occur later).

5. Only LPC (18:2) is reported to be significant in differentiating the groups. Were other LPC's altered? If not, why is 18:2 more important? Same question for the carnitines; were any others significantly altered?

Reply: We specified the significantly different names of metabolites in the updated Figures 2A and 2B, and added new Supplementary Tables S4A and 4B, to illustrate all the other 23 metabolites, including other members in the LPC and carnitine families. In our study, we took it further by selecting a subset of metabolites that still contains most information of this group but with only a limited number of metabolites. LPC (18:2) was selected based on random forest and stepwise selection as we stated in page 7, line 21 "*Out of the nine metabolites, five molecules (i.e. glycine, LPC (18:2), LPC (17:0), LPC (18:1) and C2) were select after random forest, and LPC (17:0) and LPC (18:1) were then removed after the stepwise selection. Thus, three molecules were found to* *contain independent information: glycine (adjusted OR = 0.67 (0.54-0.81), P = 8.6 × 10^{-5}), LPC (18:2) (OR = 0.58 (0.46-0.72), P = 2.1 × 10^{-6}) and acetylcarnitine C2 (OR = 1.38 (1.16-1.64), P = 2.4 × 10^{-4}) (Figure 2C)."*

6. For Figure 2 and Figure S1, how are the adjustments done?

Reply: We apologize for not clearly stating this issue. The adjustments were done by taking the residuals of metabolite concentrations from the linear regression model. We rewrote the section "Residuals of metabolite concentrations" on page 20 in the Methods section. It now reads: "*To avoid the influence of other confounding factors when plotting the concentration of metabolites, we used the residuals from a linear regression model. Metabolite concentrations were log-transformed and scaled (mean = 0, SD = 1), and the residuals were then deduced from the linear regression that included the corresponding confounding factors."*

7. For Supplementary Table S5, in the Metabolites combined with T2D risk indicators, why was C2 not included?

Reply: Please see reply to reviewer #1, point 8.

8. Table S14 lists the 188 targeted metabolites; what were the criteria used to decide whether a metabolite was used or excluded in the analysis?

Reply: We apologize if this was not clearly presented. We used two criteria, as described on pages 17 and 18: "*For each kit plate, five references (human plasma pooled material, Seralab) and three zero samples (PBS) were measured in addition to the KORA samples. To ensure data quality, each metabolite had to meet two criteria: (1) the coefficient of variance (CV) for the metabolite in the total 110 reference samples had to be smaller than 25%. In total, seven outliers were removed because their concentrations were larger than the mean plus 5× SD; (2) 50% of all measured sample concentrations for the metabolite should be above the limit of detection (LOD), which is defined as 3× median of the three zero samples. In total, 140 metabolites passed the quality controls (Supplementary Table S15): one hexose (H1), 21 acylcarnitines, 21 amino acids, 8 biogenic amines, 13 sphingomyelins (SMs), 33 diacyl (aa) phosphatidylcholines (PCs), 35 acyl-alkyl (ae) PCs and 8 lysoPCs. Concentrations of all analyzed metabolites are reported in μM. "*

References

Altmaier E, Kastenmuller G, Romisch-Margl W, Thorand B, Weinberger KM, Illig T, Adamski J, Doring A, Suhre K (2011) Questionnaire-based self-reported nutrition habits associate with serum metabolism as revealed by quantitative targeted metabolomics. *European journal of epidemiology* **26:** 145-156

Meisinger C, Strassburger K, Heier M, Thorand B, Baumeister SE, Giani G, Rathmann W (2010) Prevalence of undiagnosed diabetes and impaired glucose regulation in 35-59-year-old individuals in Southern Germany: the KORA F4 Study. *Diabet Med* **27:** 360-362

Rathmann W, Kowall B, Heier M, Herder C, Holle R, Thorand B, Strassburger K, Peters A, Wichmann HE, Giani G, Meisinger C (2010) Prediction models for incident type 2 diabetes mellitusin the older population: KORA S4/F4 cohort study. *Diabet Med* **27:** 1116-1123

Rathmann W, Strassburger K, Heier M, Holle R, Thorand B, Giani G, Meisinger C (2009) Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study. *Diabet Med* **26:** 1212-1219

Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, Lewis GD, Fox CS, Jacques PF, Fernandez C, O'Donnell CJ, Carr SA, Mootha VK, Florez JC, Souza A, Melander O, Clish CB, Gerszten RE (2011) Metabolite profiles and the risk of developing diabetes. *Nat Med* **17:** 448-453

WHO (1999) Definition, diagnosis and classification od diabetes mellitus and its complications. Part 1: Diagnosis and classification od diabetes mellitus. *Report of a WHO consultation*

2nd Editorial Decision 07 August 2012

Thank you again for submitting your revised work to Molecular Systems Biology.

We have now had the chance to consult with our Senior Editors and with Dr. Bernd Pulverer, Head of Scientific Publications at EMBO, about the 'Data availability' statement included in your manuscript and the amended version you sent us on 2.8.2012. Following this consultation and in light of the additional information provided by KORA on 3.8.2012, we would suggest to modify the 'Data availability' statement to specify that the scope of this statement refers to the subset of data that will be published in the present study and to clarify the conditions of re-use of the data included in the present work. Please find enclosed our suggested wording. We would kindly ask you to include the new statement in the Materials & Methods section of the paper.

To ensure reviewers and readers can verify the quality of the data underlying the key findings reported in this study, we would also kindly ask you to include in Supplementary Information the raw data for the novel biomarkers described in the study (glycine, lysophosphatidylcholine (LPC) (18:2) and acetylcarnitine).

*** PLEASE NOTE *** As part of the EMBO Publications transparent editorial process initiative (see our Editorial at http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology will publish online a Review Process File to accompany accepted manuscripts. When preparing your letter of response, please be aware that in the event of acceptance, your cover letter/point-by-point document will be included as part of this File, which will be available to the scientific community. More information about this initiative is available in our Instructions to Authors. If you have any questions about this initiative, please contact the editorial office (msb@embo.org).

Please resubmit your revised manuscript online, with a covering letter listing the amendments made to the manuscipt. Please resubmit the paper **within one month** and ideally as soon as possible. If we do not receive the revised manuscript within this time period, the file might be closed and any subsequent resubmission would be treated as a new manuscript. Please use the Manuscript Number (above) in all correspondence.

Click on the link below to submit your revised paper.

<http://mts-msb.nature.com/cgi-bin/main.plex?el=A5BL1Bsy4C3CII3I7A9GyLj4thuvQrgsdhhoSGgFQZ>

I thank you for your consideration and look forward to receiving your revised manuscript.

Yours sincerely,

Editor
Molecular Systems Biology