Author's response to reviews

Title: 'Health space' visualization and identification of personalized molecular phenotype and treatment responses according to the underlying relevant biological processes

Authors:

Jildau Bouwman (jildau.bouwman@tno.nl)
Jack T.W.E. Vogels (jack.vogels@tno.nl)
Suzan Wopereis (suzan.wopereis@tno.nl)
Carina M. Rubingh (carina.rubingh@tno.nl)
Sabina Bijlsma (sabina.bijlsma@tno.nl)
Ben van Ommen (ben.vanommen@tno.nl)

Version: 2  Date: 11 August 2011

Author's response to reviews: see over
To dr. T. Sands, editor of BMC Medical Genomics  
From: Dr. J Bouwman  
Concerning: revision manuscript MS: 1909425449510263  

Zeist, July 25, 2011

Dear dr. Sands,

Herewith we submit a revised version of our manuscript *Health space visualization and identification of personalized molecular phenotype and treatment responses according to the underlying relevant biological processes.*

Both reviewers had very constructive comments. We have adjusted the article according to their suggestions. We would like to thank them both for their evaluation of our paper, and we thank the editor for the opportunity to send in a revised version.

We sincerely hope that the paper in its revised form will be acceptable for publication in BMC Medical Genomics.

Yours sincerely,

Jildau Bouwman

---

**Detailed response to reviewer’s comments.**  
**Referee 1**

General comment:
Display of multivariate data based on aggregation of variables according to prior knowledge is not as such new. The use of Ingenuity Pathways to do so may be new, but does not represent a revolutionary breakthrough in data visualization. However, it is a valid and useful method. *The comment of the reviewer made it clear that the earlier text is confusing. The parameter selection for the three axes of the health space was based on the parameter selection for the 'overarching processes' as used in Bakker et al., 2011. In our manuscript we tried to summarize their selection method for which amongst others Ingenuity was used. We have adjusted this part of the text in the methods section.*

Minor essential revisions:
1) Improve explanation of the Demonstration dataset in the "Background" section. It is not stated that the study was set up as a crossover trial, which leads to confusion in the results section.  
   *We have added this information to the background section.*

2) Section "Results" - Subtitle "Dietary compounds have a large effect on the response": The claim that "subject 1 [reacts primarily] on the oxidative and inflammatory axis" does not appear supported by the figure. Is this a typo, where a different subject was meant? Furthermore, while for subject 33 the claim clearly mentions a reaction along one axis and the claim for subject 25 explicitly mentions a combined response along two axes, the wording of the claim for subject 1 seems to imply one axis displaying both oxidative and inflammatory pathways (which of course does not exist in the figure). Be careful with plural and singular axis/axes.  
   *Indeed there was a typo; it should be that subject 1 reacts primarily on the oxidative axis*.  
   *Because of the difference in scaling this statement is easier to see in figure 4. We have adjusted this section.*
3) Section "Results" - Subtitle "Constructing the model": The last paragraph states that the vector directions were also visualized and that they were similar across patients. This visualization is missing from the figure. The vectors really must be shown in figure 1. 
We have added arrows to figure 1.

4) Section "Results" - Subtitle "Dietary compounds have a large effect on the response": A claim is made that using a subset / expanded version of the original dataset proves that the method works on different datasets. Are these datasets sufficiently different to justify such a claim?
We did not want to claim that we have tested the method on different datasets, but we demonstrate that different types of data can be used as input for the health space method (e.g. transcriptomics, metabolomics and clinical data: compare figure 2 and 3). We therefore changed datasets into data types.

5) Table 1: The reference to this table does not state that it is in the supplementary data. Furthermore the data shown is a set of three lists, not a table. It would be useful to have further data in a real table. Specifically p-values and parameter class (chemistry, protein or metabolite)
We added this information to the table. The table should be part of the main text.

6) Language: The English is generally good. It is consistently comprehensible and clear, but several systematic mistakes need to be corrected and a general language correction would be helpful. Specifically: prior TO and after; correct usage of built vs. build, than vs. then, extent vs. extend; fragmentary or incomplete sentences.
We have asked a native speaker to correct the English.

Discretionary revisions:
a) The use of the treated group to define the healthy state is unconventional. While the discussion does justify this choice, and there is no problem with it from the point of view of validating the health space concept, it would be helpful to place a short sentence stating that the justification for this choice will follow in the discussion at the point where it is first mentioned.
We have added the comment to the results section 'The separation between 0 and 1 was used to separate relative healthy from more unhealthy subjects, assuming that people become healthier using the anti-inflammatory mix. This separation was further discussed in the discussion section.'

b) Abbreviations: Abbreviations such as PBMC are not defined in the text. For statistical methods such as GSEA this is less of a problem. We have explained all abbreviations at first appearance.

Referee 2
1) Visualization issue. The axes in the visualization (2D or 3D Euclidean space) usually suggest orthogonal relationships in each direction. However, it is clear that the three axes (which represent the biological processes - inflammation, oxidation, and metabolism) are actually inter-correlated. In fact, some of the elements are shared (i.e. based on Table 1, uric acid is shared between Oxidation and Inflammation, while adiponectin is shared between Inflammation and Metabolism). It is not clear whether these processes can be each analyzed independently and then visualized jointly as described.
Indeed we realize that the axes are not orthogonal. The 3D plot is a method to visualize the effect on biological processes. Biologically the processes are interdependent, but to understand how the intervention will improve the individual health status the processes are represented in this way. We have explained the biological reasoning in the discussion and added a statement to the Methods section.

2) Statistical issue. In order to visualize the “Health space”, the authors used several steps to process the data including variable selection, functional enrichment analysis, PLS-DA, data transformation, etc. Many of the steps are not clearly described i) In the Methods section,
please describe how the important variables were selected before performing functional
enrichment analysis. In particular, how the multiple testing issues was addressed

We have added a reference to the original article on the study, the parameter selection in the
article of Bakker et al., was adopted.

ii) If selection of important variables in the above procedure was based on the WHOLE
data, then performing classification/separation using supervised method (i.e. PLS-DA) will
produce over-optimistic (visual or computational) results due to overfitting. The model needs to
be validated by an independent dataset (not used for feature selection).

Unfortunately the dataset is too small to use an independent dataset for the validation of the
model. Therefore, a double cross validation method was used. We have added the reference
to this method.

iii) It is not clear why the data centers of the control group and study group must be set
differently (0 and 1) as the procedure will introduce artificial separation (which may increase
or decrease the actual separation distances among the groups). As a suggestion, authors can
first scale the variance in each axis to 1 (or any other constant) if the variances of each PLS-
DA model are very different. Then plot them in a 3D space. The 3D space can then be
translocated and rotated (similar to manipulating 3D protein structures), so that the data
center of the control group is in the origin, and the data center of the treated group are in the
positive directions. This approach will keep the relative separation of the original data points
in each direction.

We have used the 0 to 1 separation for a biological purpose. We have added the comment to
the results section. The separation between 0 and 1 was used to separate relative healthy
from more unhealthy subjects, assuming that people become healthier using the anti-
inflammatory mix. This separation was further discussed in the discussion section.

The suggestion of the reviewer is not applicable since the scores from the PLSDA models are
not comparable. In other words, a PLS-DA score of three for one model is not equal to a PLS-
DA score of three in another model. To make the scores mutually interpretable, we shifted the
scores from the ‘healthy group’ around zero by subtracting the mean score of the healthy
group. This mean was also subtracted from the scores in the non-healthy group.
Subsequently, all scores were divided by the distance between the group means, resulting in
a mean score for the non-healthy group around zero.

- Minor Essential Revisions

In Table 1, there appear to be duplicate names or redundancies within
the columns:

Adiponectin appears twice in both Inflammation column and Metabolism
column; In the Metabolism column, there are four cholesterol-related terms:
Cholesterol (total), HDL-cholesterol, LDL-cholesterol, Cholesterol.

Indeed some of the measurements are not independent, in the multivariate PLS-DA model
correlations between measurements are taken into account and therefore will have no effect
on the outcome. Doubles within the column derived from different measurement platforms.
Per parameter platform type was added to table 1.