nature portfolio

Peer Review File

A Roadmap to The Molecular Human - Linking Multiomics with Population Traits and Diabetes Subtypes.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The Molecular Human is an interesting project to provide perhaps one of widest arrays of molecular data ever performed in a human cohort. The data provide a unique opportunity to identify how different types of molecules are interrelated, particular in the context of type II diabetes. The authors' adaptation of the mutual best hits (MBH) approach, in combination with Gaussian graphical models, is an intriguing technique that can shed light on how related parts of a biological process transpire across different biofluids. Although perhaps MBH is at times too stringent a filter, it could be a useful method of initially prioritizing some validation targets. Furthermore, the capability of these data to reveal associations among genetic and epigenetics changes along with miRNA transcription was interesting and demonstrates the utility of integrating these types of data. Overall, the paper is well written and clear and should interest a wide audience of researchers. However, I think that the billing of this resource as the "Molecular Human" may be a little oversold, given that the origins of the data in a case control study and the level of missingness for some molecular data.

Minor corrections:

• In the abstract (ll. 50-51) the data source study is simply called "the multiethnic diabetes study," but the actual name should be given.

• There is a sentence fragment on ll. 94-95 that doesn't make sense.

• The abbreviation "20 Mio" on l. 109 is not the standard English abbreviation.

• On ll. 205-207, it is unclear what is meant that "comparing association p-values for QTLs with different omics phenotypes on an identical genetic variant provides an objective measure for comparing readouts from two platforms." Measure of what?

• On ll. 229-230 the statement that MBH "informs on the crosstalk between these matrices," is vague. It is a way of capturing dependencies between matrices, but I'm not sure what is meant by "crosstalk" in this context.

• On l. 295, "To the best of our knowledge, this is the first multiomics TWAS," would have been pretty easy to check on. One example is MOSTWAS: https://doi.org/10.1371/journal.pgen.1009398

• In BOX 1, the text is slightly cut off in the last line.

Comments:

• Related work is really not addressed in the introduction.

• For ll. 123-124 what "technical questions" were answered related to visualization and accessibility of omics data?

• On ll. 220-221, the method for obtaining variance explained is not addressed in the Methods section. Also, the measure does not provide the % variance explained. It is a kind of pseudo-R2. However, pseudo-R2 cannot be compared across datasets, which is what is being done in this case. They are typically used to compare different models on the same dataset, or to provide intuition for the quality of the model fit.

• On ll. 298-299 "gene expressions - lipid/lipoprotein associations were attributed to a group of five gene transcripts". It's unclear how this attribution was made.

• Was the volume of TWAS associations with lipids a result of the underlying cohort?

• On ll. 394-395, the manuscript states "metabolomics and glycomics platforms are characterized by multiple MBH linking common molecular pathways", however, I'm not sure this really conveys how MBH can link different parts of the same molecular pathway assayed in different biofluids using different platforms.

• On ll. 403-404 the manuscript states that "Complex and multifactorial conditions, such as diabetes, cardiovascular and autoimmune diseases require comprehensive characterization for proper diagnostics and treatment." Please give some examples of when this is true.

• On ll. 426-429 the manuscript makes the case that participants were enrolled "continuously on an availability basis…to avoid any possible batch effects between cases and controls." However, this approach does not avoid any possibility of batch effects. Batch effects might be likely, but they could occur due to differential availability.

• On l. 847 the method for calculating the MBH is not clear.

• The fact that study participants were not fasting could bias the results, if the average time between meals differs between those with or without T2D.

• It would be helpful to have a breakdown of what types of omics data are missing (the pattern of missingness) by case control status. This could certainly be in the supplementary materials.

Reviewer #2 (Remarks to the Author):

This is a well-written manuscript from a group with established previous experience in omics. The authors conduct a well-designed attempt to define molecular Interactions linking 6,304

quantitative molecular traits with 1,221,345 genetic variants, methylation at 470,837 DNA CpG sites, and gene expression of 57,000 transcript with disease endpoints. We definitely need more research like this. They combine the results into a network of > 34,000 statistically significant links and use it to construct multi-molecular network of diabetes subtypes, in the Qatari metabolomics study of 391 individuals.

They use mainly three statistical methods, *WAS, MBH and GGMs. By using these three methods and overlapping data, they define a so-called “roadmap” for evaluating multi-omics data. They also define molecular subtypes underlying different diabetic outcomes. Although the sample size is small, very deep phenotyping of the study set allows for new questions, such as consistency across the platforms and omics integration for defining far connections across three tissue types. The results are presented in a website in a user-friendly manner.

Major comments:

While I agree that omics integration is essential and we need instructions and guidelines on how to integrate omics data into meaningful biological pathways, I find it difficult to get the main message of the article. Was it to compare across platform performances, was it to link different molecules from different platforms and tissues, or was it to subtype individuals, or guiding how integration should be done? In its current form, it may also be more suitable for a good database description journal rather than Nature Communications. The website and server will definitely be used widespread by other omics researchers.

Most of the highlight relies on MBH and to my opinion it needs more explanation and justification. There is reference to two clusters of orthologs papers but it was unclear to me how it was implemented for the omics data. Can MBH differentiate between platform similarity versus reactional proximity? Were there any MBHs for differently annotated molecules?. Use of MBH which originally defined from finding ortologue genes and it is used for the first time in the context of human omics, so it should also be explained earlier and more in-depth. Could the authors train an MBH-like algorithm within a platform e.g. by removing the molecular labels in the validation set and test it across the platforms? Or at least show its maximum performance within the same platform, shortly, the MBH application for omics platforms will benefit from benchmarking. If that was already done, the authors could explain it in the manuscript.

With such deeply phenotyped population the reader of this journal might expect novel methodological efforts (and well-explained) on in-silico analyses, such as exploring more formal omics data fusion methods or attempts to address cross platform missingness via multiple imputation and other options of learning, going further than of simple linear models. The data are also suitable for reverse regression analysis i.e multi-phenotype GWAS. The authors have exclusive freedom to address such questions and would be useful for everyone as we are introduced to a new omics platform every other day. For example, instead of using 3 different versions of creatin(for

example), can we use a PC (?) that captures creatin, or select the best one and generate a dataset with unique features only?

Minor comments:

The methods and datasets that were used before can be shorthened and mentined in the supplementary, so that the reader can focus on the novel assets.

Please check for typos, Comincs, or Comics? (Box2)

Reviewer #3 (Remarks to the Author):

The manuscript “The Molecular Human – A Roadmap of Molecular Interactions Linking Multiomics Networks with Disease Endpoints” by Halama et al. uses 391 participants from a multiethnic diabetes study with omic profiling from 18 different platforms to examine the interconnectedness of molecular data using multiple network approaches (partial correlations, GGM, mutual best correlations, etc.). The file “Molecular Human” is a molecular network has ~34,000 significant interconnections. Overall, this study is unique in the use of dense multiomic profiling and the characterization of the inter-relationships between all of these molecular profiles. What is most useful and will serve the overall research community well is the open access website: http://comics.metabolomix.com. This website stores these results and allows outside individuals to explore these interrelationships to further their own understanding, and potentially extended to own research fundings. This is a tremendous asset that to provide. While there are some points that need to be addressed and acknowledged as limitations (sample size, few disease endpoints, etc.). Overall, this study is novel and adds valuably to the scientific community. In addition, this work was performed using a minority population (QMDiab), where the understanding of these molecular relationships is particularly not well-understood.

My specific points follow:

1)These correlation networks shed light on the multimorbidity of omic variants across multiple platforms. This is highlighted well in examples presented (e.g., Figure 4). However, in the discussion, there is little commentary on what might be done with the multiomic networks. It would be useful to elaborate on what understanding/hypotheses may be gleaned from the multiomic correlation networks through an example provided in the results.

2)While there is a description of what omic data types capture specific demographic features the best (e.g., metabolomics and proteomics), can further commentary be provided on the specific strengths of any given omic data type that are observed through these networks? Are there some

omics that seem redundant? Are there important distinctions that are observed between what each omic layer captures? What may be captured by the synthesis of them together? Some of these points might be best summarize by looking at a selection of 2-way comparisons vs. fully multiomics using a subset of the variants. For example, is one omic datatype is missing, does a specific network fall apart because a hub is absent? While these are all general questions, further inquiry to synthesize some of these general points will add substantial impact to this work.

3)While the manuscript does have multiple molecular interactions, the work described does not have multiple disease endpoints. While there are basic characteristics, such as age, sex, BMI, these are not disease endpoints. As such, either "multiple disease endpoints" should be modified to "diabetes" or something referring to population characteristics (age, sex). Otherwise, additional disease endpoints should be added.

4)One concern/quandary with correlation networks is potential residual confounding. Since this is a case-control study where individuals were ascertained for diabetes. The authors did an excellent job describing the diabetes network in the manuscript; however, potential confounding effects of this in the overall correlation networks should be addressed as a limitation.

5)There is little discussion of this as a minority population. Was this adjusted for PCs? Also, is there anything here that may be specific to this minority group? Further comment about this as an advantage (e.g., few minority populations in the literature) and the potential limitations in generalizability should be addressed.

6)While the advantage of this cohort is the tremendous amount of multiomic data and there is no suitable replication cohort with the same omic data, is there any way to replicate some of the strongest observed correlations?

7)A further consideration are the biofluids and omics that are missing. Most notably stool metabolomics and microbiome data. If this is the "Molecular Human" further comment on what is missing should be considered.

8)Minor: The metabolomic acronyms in the tables are almost too short, so it is not easy for the reader to quickly digest what molecular platform is what. Please consider either adding an additional column that lists the specific omic type or adding to the acronym to make it clear (e.g., MetP; MetS, etc.).

The Molecular Human Overture – A Roadmap of Molecular Interactions Linking Multiomics Networks with Population Traits and Diabetes Subtypes

Response to the reviewers' comments

We appreciate the thorough comments and suggestions raised by the reviewers. We thank the reviewers for their time and constructive feedback. Please find below our point-by-point response, including a description of newly added figures and how they reflect on the manuscript. Additionally, we provide a change-tracked version of the revised manuscript where all modifications are highlighted.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The Molecular Human is an interesting project to provide perhaps one of widest arrays of molecular data ever performed in a human cohort. The data provide a unique opportunity to identify how different types of molecules are interrelated, particular in the context of type II diabetes. The authors' adaptation of the mutual best hits (MBH) approach, in combination with Gaussian graphical models, is an intriguing technique that can shed light on how related parts of a biological process transpire across different biofluids. Although perhaps MBH is at times too stringent a filter, it could be a useful method of initially prioritizing some validation targets. Furthermore, the capability of these data to reveal associations among genetic and epigenetics changes along with miRNA transcription was interesting and demonstrates the utility of integrating these types of data. Overall, the paper is well written and clear and should interest a wide audience of researchers. However, I think that the billing of this resource as the "Molecular Human" may be a little oversold, given that the origins of the data in a case control study and the level of missingness for some molecular data.

Response: We thank the reviewers for their time and constructive comments and suggestions and hope to have fully addressed all points raised.

We understand the reviewer's comment feeling that the title "billing of this resource as the "Molecular Human " may be a little oversold, given that the origins of the data in a case control study and the level of missingness for some molecular data". However, please note that our study describes arguably the largest quantification of molecular interactions in the human body, combining 6,304 individual quantitative molecular traits in addition to 1,221,345 directly genotyped genetic variants, methylation at 470,837 DNA CpG sites, and gene expression of 57,000 transcripts at 20M read coverage, captured by 18 different omics platforms. Moreover, all the omics were applied on a relatively large cohort of 391 individuals; the smallest number of individuals per molecular interaction was still 235 subjects for associations between RNA ⇔ Saliva metabolites was monitored. Thus, this is a cohort where multiple phenotypes, including age, sex, BMI, and T2D, could be investigated with a very broad array of omics. As per the reviewer's suggestion, to highlight the uniqueness of our work, we refer to the previous omics study (please see response **R1C1**). Noteworthy, other studies, where extended omics profiling was implemented, are limited in the number of participants, whereas studies, where cohort size exceeds 100 participants, are limited in the number of deployed omics platforms. In the discussion section, we included additional comments on measurements that could be considered in the future to provide an even more complete picture of the molecular human (please see response **R3C7**). Given that this is the first large initiative, setting the stage for other comprehensive multiomics studies, we propose changing the title to: "***The Molecular Human Overture – A Roadmap of Molecular Interactions Linking Multiomics Networks with Population Traits and Diabetes Subtypes***".

COMMENTS

R1C1: Related work is really not addressed in the introduction.

Response R1C1: We agree with the reviewer's comment. Following the reviewer's suggestion, we included an additional paragraph reflecting on previous multiomics works. We think that this additional paragraph is further emphasizes the uniqueness of our study regarding the comprehensiveness of the resources we provide. The following paragraph was added to the introduction:

"Indeed, molecular processes were monitored in human biofluids after integration of multiple different omics including genomics, methylation, transcriptomics, proteomics, and metabolomics [1-10]. For instance, study utilizing broader array of omics, frequently applied to limited number of

subjects (between 1 – 36 individuals), were conducted to uncover dynamic changes in diverse molecular components in response to viral infection[9], spaceflight[7], as well as extensive exercise[8] and could be considered as personalized omics activities. In contrast, bigger cohort study (≥100 individuals) only deployed a limited spectrum of omics measurements [1-3,5]. The molecular processes related to obesity, diabetes, liver function, or cardiovascular disease, were determined in the lifestyle changes study deploying genomics, proteomics and metabolomics[2]. Similarly, disease signatures associating with Schizophrenia or HIV-infection were assessed using proteomics and metabolomics[1] or metabolomics and lipidomics[5], respectively. In the larger cohort (over 1000 subjects) study the limited multiomics array was implemented to reveal metabolite-protein interactions[4] or to define molecular network of Alzheimer disease[6].”

R1C2: For ll. 123-124 what "technical questions" were answered related to visualization and accessibility of omics data?

Response R1C2: We agree with the reviewer that this sentence might be confusing as we visualized the molecular interactions and provided access to the data via the COmics server as well as Excel files (a Cytoscape-based network is also available on GitHub), without testing other options. To clarify this, we amended the text as follows:

“*Here, we combine all data that we ever generated on the QMDiab study with the aim to simultaneously answer technical questions related to omics platform complementarity and data integration. We also asked biological questions related to the interrelationships between these molecular traits and their association with various phenotypes including T2D. Further, we visualized the molecular interactions in the form of interactive network to which we provide interactive and free access.”*

R1C3: On ll. 220-221, the method for obtaining variance explained is not addressed in the Methods section. Also, the measure does not provide the % variance explained. It is a kind of pseudo-R2. However, pseudo-R2 cannot be compared across datasets, which is what is being done in this case. They are typically used to compare different models on the same dataset, or to provide intuition for the quality of the model fit.

Response R1C3: We use the "pseudo-R2" reported by the R package randomForest to estimate how well a given omics phenotype can predict age, sex, BMI, and diabetes state. For continuous variables, this "pseudo-R2" is defined as one minus the mean square error of the regression divided by the variance of the dependent variable. Although this measure is typically used to compare different models on the same dataset or to provide intuition for the quality of the model fit, there is no reason why this measure would not provide an estimate of how well a dependent variable can be modeled using different sets of explanatory variables, as we do here. To make this clearer to the reader, we added the following text to the methods section:

"We use the "pseudo-R2" reported by the R package randomForest as an estimate of how well a given omics phenotype can predict age, sex, BMI and diabetes state. For continuous variables this "pseudo-R2" is defined as one minus the mean square error of the regression divided by the variance of the dependent variable. Note that the "pseudo-R2" is not a strict measure of the explained variance and is used here to provide an intuition for the quality of the model fit that can be obtained using the different omics datasets."

R1C4: On ll. 298-299 "gene expressions - lipid/lipoprotein associations were attributed to a group of five gene transcripts". It's unclear how this attribution was made.

Response: The attribution was made based on the identified associations centered predominantly around five gene transcripts. For the clarification, we rephrased the sentence as follows:

"*Those gene expressions – lipid/lipoprotein associations were found to be grouped predominantly around five gene transcripts including GATA Binding Protein 2 (GATA2), Histidine Decarboxylase (HDC), Fc Epsilon Receptor 1 alpha (FCER1A), Pyruvate Dehydrogenase Lipoamide Kinase 4 (PDK4), and Membrane Spanning 4-Domains A3 (MS4A3), showing association with 590, 516, 36, 18, and 18 lipids/lipoproteins respectively.*"

R1C5: Was the volume of TWAS associations with lipids a result of the underlying cohort?

Response R1C5: The identified TWAS associations could be driven by different factors, including the cohort and the measured molecules. On the other hand, please note that although we used a diabetes case-control study, except for the section where we examine the explained variability,

our investigation is mainly about relationships between the different omics layers. For this purpose, variability in the samples is beneficial as it increases the signal to noise ratio.

We have added the following text to the discussion section:

"*Because our cohort consists of healthy and T2D subjects some of the observed associations could be driven by the molecular alterations which are known features of T2D (e.g. elevated carbohydrates, lipids, and branch chain amino acid levels). For instance, identified TWAS associations, dominated by 5 genes linked to various lipids could reflect on the enrolled participants characteristics, which might be recognized as study limitation. Nevertheless, such an experimental setting enabled us to uncover a range of lipids/lipoproteins with immunostimulatory properties in our lipidomics TWAS, which holds significance for cardiovascular disease.*"

R1C6: On ll. 394-395, the manuscript states "metabolomics and glycomics platforms are characterized by multiple MBH linking common molecular pathways", however, I'm not sure this really conveys how MBH can link different parts of the same molecular pathway assayed in different biofluids using different platforms.

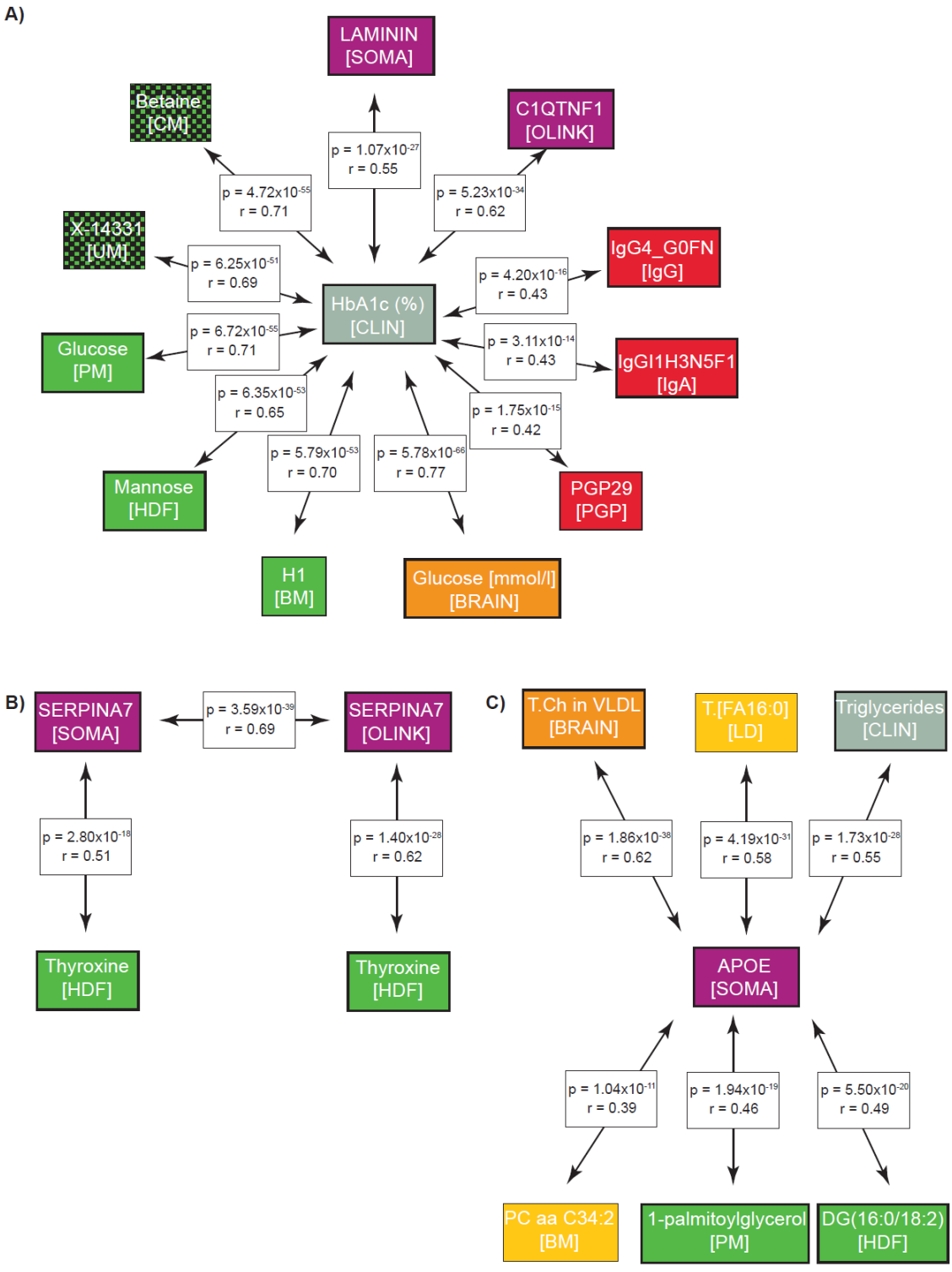
Response R1C6: We agree with the reviewer that the MBH and its reflection on the biological processes were not well outlined in the manuscript, and thus, discussion was limited. Therefore, we included an additional paragraph in the results section highlighting the identified MBH between different omics and platforms and their biological relevance. We further amended the unclear sentence in the discussion section. The additional paragraph of the result section and amendment sentence in the discussion are outlined below:

In Results

"Molecular crosstalk linking omics associations with biological process.

We further investigated the relevance of MBH for capturing biologically relevant process. For example, HbA1C [CLIN], known marker for diagnosis and monitoring of Type 2 Diabetes (T2D) [11,12]*, was showing association with the elevated blood glucose level measured on different platforms as well as other molecules previously described in the context of diabetes including betaine*[13]*, mannose*[14] *and X-14331*[15] ***Supplementary Figure 2 A****. We also found MBH between thyroxine and*

*SERPINA7, a thyroxine-binding globulin, which in the bloodstream carries thyroxine and triiodothyronine into thyroid gland[16]; the MBH was found independently of used technical platform (thyroxine (HDF) ⇔ SERPINA7 (OLINK) (p-value = 1.4x10^{-28}; r = 0.62); thyroxine (HDF) ⇔ SERPINA7 (SOMA) (p-value = 2.8x10^{-18}; r = 0.51)) (**Supplementary Figure 2 B**). The MBHs detected between Apolipoprotein E (APOE), involved in lipid metabolism, and different lipid molecules across various platforms e.g. APOE (SOMA) ⇔ Total cholesterol in VLDL (BRAIN) (p-value = 1.9x10^{-38}; r = 0.65); APOE (SOMA) ⇔ Total [FA16:0] (LD) (p-value = 4.2x10^{-37}; r = 0.62); APOE (SOMA) ⇔ palmitoyl-linoleoyl-glycerol (16:0/18:2) (HDF) (p-value = 5.5x10^{-20}; r = 0.58); APOE (SOMA) ⇔ 1-palmitoylglycerol (1-monopalmitin) (PM) (p-value = 1.9x10^{-19}; r = 0.49); APOE (SOMA) ⇔ PC.aa.C34.2 (BM) (p-value = 1.0x10^{-11}; r = 0.34), further suggest that actual biological processes can be captured by the MBH (**Supplementary Figure 2 C**). The majority of MBH identified between proteomics and glycomics replicated the associations reported by us previously[17]. The MBH's between proteomics and transcriptomics showed frequently the gene transcripts and corresponding proteins SIGLEC14 (RNA) ⇔ SIGLEC14 (SOMA) (p-value = 1.1x10^{-37}; r = 0.60); GNLY (RNA) ⇔ GNLY (SOMA) (p-value = 9.6x10^{-22}; r = 0.49); LILRA5 (RNA) ⇔ LILRA5 (OLINK) (p-value = 2.0x10^{-9}; r = 0.33). This data indicate that associations depicted by the MBH reflect on the actual biological processes. Yet, the MBH could be also used in different capacities. In **Supplementary Information Note 4** we showed that MBH linking lipidomics with metabolomics can provide further insight into the structure of complex lipids."*


A)
LAMININ [SOMA]
Betaine [CM]
C1QTNF1 [OLINK]
p = 1.07x10-27
r = 0.55
p = 4.72x10-55
r = 0.71
p = 5.23x10-34
r = 0.62
X-14331 [UM]
IgG4_G0FN [IgG]
p = 6.25x10-51
r = 0.69
p = 4.20x10-16
r = 0.43
HbA1c (%) [CLIN]
Glucose [PM]
p = 6.72x10-55
r = 0.71
p = 3.11x10-14
r = 0.43
IgGI1H3N5F1 [IgA]
p = 6.35x10-53
r = 0.65
p = 1.75x10-15
r = 0.42
Mannose [HDF]
p = 5.79x10-53
r = 0.70
p = 5.78x10-66
r = 0.77
PGP29 [PGP]
H1 [BM]
Glucose [mmol/l] [BRAIN]
B)
SERPINA7 [SOMA]
p = 3.59x10-39
r = 0.69
SERPINA7 [OLINK]
p = 2.80x10-18
r = 0.51
p = 1.40x10-28
r = 0.62
Thyroxine [HDF]
Thyroxine [HDF]
C)
T.Ch in VLDL [BRAIN]
T.[FA16:0] [LD]
Triglycerides [CLIN]
p = 1.86x10-38
r = 0.62
p = 4.19x10-31
r = 0.58
p = 1.73x10-28
r = 0.55
APOE [SOMA]
p = 1.04x10-11
r = 0.39
p = 1.94x10-19
r = 0.46
p = 5.50x10-20
r = 0.49
PC aa C34:2 [BM]
1-palmitoylglycerol [PM]
DG(16:0/18:2) [HDF]

*"**Supplementary Figure 2. A)** MBH identifies clinically relevant association between HbA1C and molecules involved in pathology of T2D. **B)** MBH reflects on physiological processes such as thyroxine transport and **C)** lipid metabolism. "*

In Discussion

"*Deployment of MBH across omics platforms covering different and overlapping molecular traits, which we investigated here, can indeed be used to define molecular orthology bridging different platforms and omics. While investigating MBH between platforms containing overlapping molecular traits (e.g. SOMA ⇔ OLINK; HDF ⇔ PM; IgG ⇔ IgA) and identifying association between them we showed good platform performance regarding components identification. On the other hand, MBH applied to molecular traits between different omics (e.g. thyroxine (HDF) ⇔ SERPINA7 (OLINK); APOE (SOMA) ⇔ Total cholesterol in VLDL (BRAIN)), reveal biologically relevant molecular interactions. Those examples underscore the value of the measurements and suggests the utility of data integration.*"

R1C7: On ll. 403-404 the manuscript states that "Complex and multifactorial conditions, such as diabetes, cardiovascular and autoimmune diseases require comprehensive characterization for proper diagnostics and treatment." Please give some examples of when this is true.

Response R1C7: As per the reviewer's suggestion, we added examples for clarity reasons. Please see below the newly added text.

"……This is particularly relevant when the disease progression is not well defined, as well as when various comorbidities occur. For instance, treatment of T2D diabetes patients depends on multiple factors including their blood glucose level (tested with HbA1C or 1,5-AG), insulin resistance status (tested with hyperinsulinemia-euglycemic clamp or HOMA-IR), capability to produce insulin (e.g. fasting insulin or C-peptide test) as well as presence of other diseases (e.g. cardiovascular disease, neuropathy, kidney disease, and retinopathy)[18]. The identification of five subgroups among diabetes patients, stratifying individuals with respect to disease progression and diabetic complication risks, adds to the complexity but could further navigate more personalized treatment options[19]. The multiomics analysis offers a powerful framework that

could be utilized to better phenotype patients with complex diseases by defining molecular interactions across omics layers with functional relevance for disease endpoints."

R1C8: On ll. 426-429 the manuscript makes the case that participants were enrolled "continuously on an availability basis…to avoid any possible batch effects between cases and controls." However, this approach does not avoid any possibility of batch effects. Batch effects might be likely, but they could occur due to differential availability.

Response R1C8: The reviewer makes us understand that the sentence was not conveying our thoughts correctly. We amended the sentence as follows:

"….., using identical collection kits and protocols, to avoid batch effects between cases and controls, which could occur during the initial phase of patient enrollment and sample collection."

R1C9: On l. 847 the method for calculating the MBH is not clear.

Response R1C9: Following the reviewer's suggestion, we amended the method section. We also provided an extended description of MBH in the main text, as this concept was not well outlined in the initial submission. Please see the changes below:

In results section

"*In the biological system, homology reflects on molecular, structural, or physiological similarity in different species[20]. Genetic elements inherited in two species by a common ancestor are defined as homologs [21], and are further classified as orthologs if they diverged through speciation or as paralogs if they evolved through duplication [22,23]. The gene orthologs are typically the most similar genes in the respective species in terms of sequence, structure, and function[24]. Among different approaches, deployed for identification of orthology, the most used is bidirectional best hit (BBH), which defines as orthologs all pairs of genes between two species that are reciprocally similar to one another than to any other gene in a sequence similarity search[25,26]. Inspired by this concept, we hypothesized that BBH, which is hypothesis-free approach, could be utilized beyond genomics to identify molecular orthologs between platforms. Here we define the BBH applied for multiomics as MBHs and use this approach to identify ortholog readouts between two platforms.*"

In methods section

"The association between each two platforms was described using mutual best hit (MBH) aiming to identify pairs of features (e.g., genes, proteins) that exhibited a significant correlation with each other. Spearman correlation coefficients between unscaled raw omics data were computed. Next, mutual best hits were identified; only those pairs demonstrating a significant and reciprocal relationships were retained. Reciprocity implies a two-way relationship, where the correlation is bidirectional. Platform-pairwise Bonferroni significance cutoffs ($p < 0.05 / (nPLTA1 * nPLAT2 / 2)$) were used." obtained after cross-platform correlation."

R1C10: The fact that study participants were not fasting could bias the results, if the average time between meals differs between those with or without T2D.

Response R1C10: We understand the reviewer's concern, as the study participants were in a non-defined fasting state. Yet, given the study setting, most participants did not have a major meal for at least 2 h before sampling, and thus, the subjects were not in the postprandial state. Additionally, in our previous study, we showed that increased variability is random and does not tend to bias the associations [15]. Following the Reviewer's suggestion, we further clarified this point.

"Similarly, the fact that study participants were not fasting implies further biological variation in the data, which may strengthen correlation signals that are related to processes confounded by fasting when case-control studies are conducted. Noteworthy, while the average time between meals for individuals with or without T2D was not assessed, and the fasting status of the participants was not defined, our previous study demonstrated that the increased variability is random and does not tend to bias the associations [15]."

R1C11: It would be helpful to have a breakdown of what types of omics data are missing (the pattern of missingness) by case control status. This could certainly be in the supplementary materials.

Response R1C11: Based on the comment, we understand that the reviewer is concerned about the omics data applied to all the case/control samples. We want to clarify that all the samples

from both cases and controls were submitted for all the omics measurements we are reporting in this current manuscript. All the per-platform measurements were conducted simultaneously on both sample types (cases and controls), which were randomized for every measurement to minimize the biases. We commented on this in the materials and methods section to clarify this point.

"*The obtained samples were submitted for deep molecular phenotyping which utilized clinical chemistry parameters along with omics measurements across 18 technically diverse platforms. All the cases and controls were measured simultaneously on each analytical platform to minimize measurements biases*."

MINOR CORRECTIONS:

R1C12: In the abstract (ll. 50-51) the data source study is simply called "the multiethnic diabetes study," but the actual name should be given.

Response: Following the reviewer's suggestion, we added the actual name of the study in the Abstract.

R1C13: There is a sentence fragment on ll. 94-95 that doesn't make sense.

Response: Following the reviewer's suggestion, we amended the sentence as follows:

"*However, with many different technologies and platforms available, questions arise as to the choice of the platforms to use, their complementarity, and in particular, how to integrate these complex data sets once they have been collected.*"

R1C14: The abbreviation "20 Mio" on l. 109 is not the standard English abbreviation.

Response: Following the reviewer's suggestion, we changed Mio to million.

R1C15: On ll. 205-207, it is unclear what is meant that "comparing association p-values for QTLs with different omics phenotypes on an identical genetic variant provides an objective measure for comparing readouts from two platforms." Measure of what?

Response R1C15: As per the reviewer's suggestion, the sentence was amended (please see below).

"Thus, the calculated association p-values for each QTLs with different omics phenotypes, conducted on an identical genetic variant, could serve as an objective measure, enabling the comparison of readouts from two platforms."

R1C16: On ll. 229-230 the statement that MBH "informs on the crosstalk between these matrices," is vague. It is a way of capturing dependencies between matrices, but I'm not sure what is meant by "crosstalk" in this context.

Response R1C16: Following the reviewer's suggestion, we changed the sentence:

"*….. is capturing dependencies between those matrices, and thus inform about the interactions between them*."

R1C17: On l. 295, "To the best of our knowledge, this is the first multiomics TWAS," would have been pretty easy to check on. One example is MOSTWAS: https://doi.org/10.1371/journal.pgen.1009398

Response R1C17: The TWAS, which probed correlations with CpG sites and microRNAs, as defined in the given example, were previously conducted, and we also cite theses in our manuscript. Nevertheless, this is an example of a study claiming to be multiomics, but it focuses only on DNA and RNA aspects, which we do not consider a truly multiomics study. Yet, we understand that our text could be confusing. To clarify, we changed this part of the manuscript as follows:

"To the best of our knowledge, this is the first, conducted to date, multiomics TWAS, that includes proteomics, metabolomics, lipidomics, lipoproteomics and glycomics in addition to methylation and miRNA."

R1C18: In BOX 1, the text is slightly cut off in the last line.

Response R1C18: *Following the reviewer's comment, we enlarged the BOX1 field.*

Reviewer #2 (Remarks to the Author):

This is a well-written manuscript from a group with established previous experience in omics. The authors conduct a well-designed attempt to define molecular Interactions linking 6,304 quantitative molecular traits with 1,221,345 genetic variants, methylation at 470,837 DNA CpG sites, and gene expression of 57,000 transcript with disease endpoints. We definitely need more research like this. They combine the results into a network of > 34,000 statistically significant links and use it to construct multi-molecular network of diabetes subtypes, in the Qatari metabolomics study of 391 individuals.

They use mainly three statistical methods, *WAS, MBH and GGMs. By using these three methods and overlapping data, they define a so-called "roadmap" for evaluating multi-omics data. They also define molecular subtypes underlying different diabetic outcomes. Although the sample size is small, very deep phenotyping of the study set allows for new questions, such as consistency across the platforms and omics integration for defining far connections across three tissue types. The results are presented in a website in a user-friendly manner.

While I agree that omics integration is essential and we need instructions and guidelines on how to integrate omics data into meaningful biological pathways, I find it difficult to get the main message of the article. Was it to compare across platform performances, was it to link different molecules from different platforms and tissues, or was it to subtype individuals, or guiding how integration should be done? In its current form, it may also be more suitable for a good database description journal rather than Nature Communications. The website and server will definitely be used widespread by other omics researchers.

Response: We thank the reviewers for their time and constructive comments and suggestions and hope to have fully addressed all points raised. We are glad that the reviewer recognizes the need to conduct more multiomics research similar to our effort. We also understand that our manuscript might be somewhat overwhelming as we ask multiple questions simultaneously. For instance, we are touching upon technical aspects in a manner enabling future multiomics analysis of, e.g., UK Biobank multiomics data while examining the potential of MBH in such a setting. Using MBH, we further evaluate the platform's complementarity with respect to technical aspects (described in greater detail in the supplementary part of our manuscript) as well as biological aspects, which we included with this revision (please refer to **Response R1C6**). Next, we showed how the multiomics approach evaluated with our methods could provide further insight into complex diseases such as diabetes by improving molecular phenotyping. Finally, we are providing

the scientific community with interactive access to our data via COmics server, which we specifically designed for this purpose. We also provide examples showing how to utilize this data set for hypothesis generation. We constructed this manuscript for the broader community, having in mind especially those who could implement multiomics approaches in their research effort, to show the possibilities of such approaches simultaneously pointing towards the potential discoveries and limitations. Thus, we think that Nature Communication is suitable for our work as we see the Nature Communication readers constitute a broad spectrum consisting of basic and clinical researchers as well as bioinformaticians who will be strongly interested in this manuscript. This view is driven by the feedback we received after oral presentations of this work as we were approached by clinicians, basic and omics scientists, and bioinformaticians who were genuinely interested in our work.

R2C1: Most of the highlight relies on MBH and to my opinion it needs more explanation and justification. There is reference to two clusters of orthologs papers but it was unclear to me how it was implemented for the omics data. Can MBH differentiate between platform similarity versus reactional proximity? Were there any MBHs for differently annotated molecules?. Use of MBH which originally defined from finding ortologue genes and it is used for the first time in the context of human omics, so it should also be explained earlier and more in-depth. Could the authors train an MBH-like algorithm within a platform e.g. by removing the molecular labels in the validation set and test it across the platforms? Or at least show its maximum performance within the same platform, shortly, the MBH application for omics platforms will benefit from benchmarking. If that was already done, the authors could explain it in the manuscript.

Response R2C1: We agree with the reviewer that the concept of MBH was not described extensively. Additionally, we realized that examples highlighting findings from MBH would contribute to a better outline reflecting on the implementation of MBH. We provided a new paragraph in the results section highlighting biologically relevant examples depicted by MBH. We also included an additional paragraph in Supplementary Information showing an example of how MBH was used to reveal complex lipid structures by linking lipidomics with metabolomics. We think that with those examples, we are showing the maximum performance of the MBH approach as well as underscoring the feasibility of MBH in multiomics analysis.

To clarify the reviewer's comment related to the implementation of MBH within the same platform, the MBH is conducted to identify ortholog readouts between two platforms but not within the platform. The GGM was conducted for within-platform analysis.

Please see the new paragraphs which have been included to address reviewer comments:

In results section

To clarify MBH we added:

"In the biological system, homology reflects on molecular, structural, or physiological similarity in different species[20]. Genetic elements inherited in two species by a common ancestor are defined as homologs [21], and are further classified as orthologs if they diverged through speciation or as paralogs if they evolved through duplication [22,23]. The gene orthologs are typically the most similar genes in the respective species in terms of sequence, structure, and function[24]. Among different approaches, deployed for identification of orthology, the most used is bidirectional best hit (BBH), which defines as orthologs all pairs of genes between two species that are reciprocally similar to one another than to any other gene in a sequence similarity search[25,26]. Inspired by this concept, we hypothesized that BBH, which is hypothesis-free approach, could be utilized beyond genomics to identify molecular orthologs between platforms. Here we define the BBH applied for multiomics as MBHs and use this approach to identify ortholog readouts between two platforms."

To highlight MBH as a tool depicting biologically relevant associations:

*"**Molecular crosstalk linking omics associations with biological process.***

*We further investigated the relevance of MBH for capturing biologically relevant process. For example, HbA1C [CLIN], known marker for diagnosis and monitoring of Type 2 Diabetes (T2D) [11,12], was showing association with the elevated blood glucose level measured on different platforms as well as other molecules previously described in the context of diabetes including betaine[13], mannose[14] and X-14331[15] (**Supplementary Figure 2 A**). We also found MBH between thyroxine and SERPINA7, a thyroxine-binding globulin, which in the bloodstream carries thyroxine and triiodothyronine into thyroid gland[16]; the MBH was found independently of used technical*

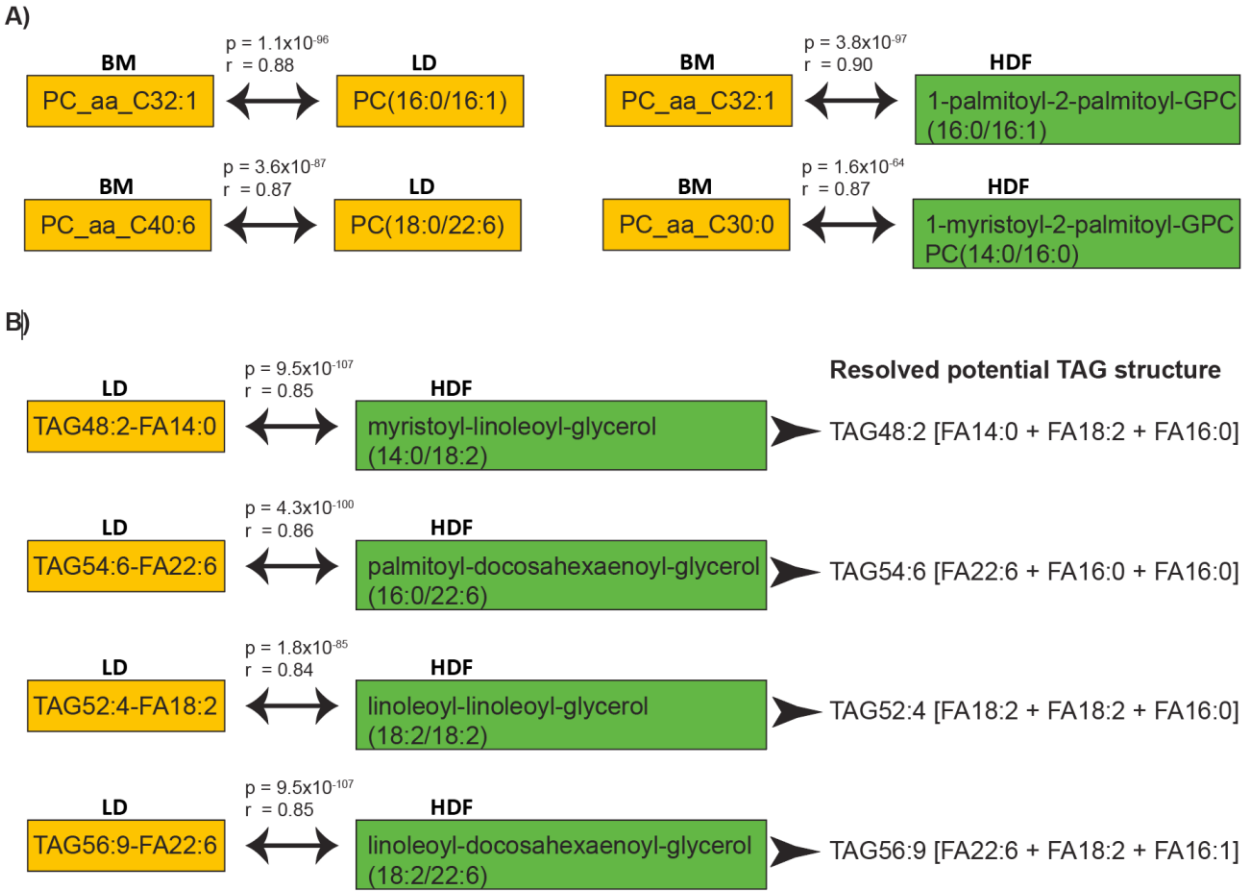
*platform (thyroxine (HDF) ⇔ SERPINA7 (OLINK) (p-value = $1.4x10^{-28}$; r = 0.62); thyroxine (HDF) ⇔ SERPINA7 (SOMA) (p-value = $2.8x10^{-18}$; r = 0.51)) (**Supplementary Figure 2 B**). The MBHs detected between Apolipoprotein E (APOE), involved in lipid metabolism, and different lipid molecules across various platforms e.g. APOE (SOMA) ⇔ Total cholesterol in VLDL (BRAIN) (p-value = $1.9x10^{-38}$; r = 0.65); APOE (SOMA) ⇔ Total [FA16:0] (LD) (p-value = $4.2x10^{-37}$; r = 0.62); APOE (SOMA) ⇔ palmitoyl-linoleoyl-glycerol (16:0/18:2) (HDF) (p-value = $5.5x10^{-20}$; r = 0.58); APOE (SOMA) ⇔ 1-palmitoylglycerol (1-monopalmitin) (PM) (p-value = $1.9x10^{-19}$; r = 0.49); APOE (SOMA) ⇔ PC.aa.C34.2 (BM) (p-value = $1.0x10^{-11}$; r = 0.34), further suggest that actual biological processes can be captured by the MBH (**Supplementary Figure 2 C**). The majority of MBH identified between proteomics and glycomics replicated the associations reported by us previously[17]. The MBH's between proteomics and transcriptomics showed frequently the gene transcripts and corresponding proteins SIGLEC14 (RNA) ⇔ SIGLEC14 (SOMA) (p-value = $1.1x10^{-37}$; r = 0.60); GNLY (RNA) ⇔ GNLY (SOMA) (p-value = $9.6x10^{-22}$; r = 0.49); LILRA5 (RNA) ⇔ LILRA5 (OLINK) (p-value = $2.0x10^{-9}$; r = 0.33). This data indicate that associations depicted by the MBH reflect on the actual biological processes. Yet, the MBH could be also used in different capacities. In **Supplementary Information Note 4** we showed that MBH linking lipidomics with metabolomics can provide further insight into the structure of complex lipids."*

In supplementary material section

"*Deploying platform complementarity to provide further insight into the structure of complex lipids.*

The composition of fatty acid (FA) side chains in complex lipids such as phosphatidylcholine or triacylglycerols play a role in a broad range of biological processes and it is thus critical to determine FA composition in measured lipids. This information is, however, not provided for phosphatidylcholines measured on the BM platform or triacylglycerols measured on the LD platform. We previously showed that the composition of phosphatidylcholines measured on the BM can be resolved by LD platform[27]. Here, we investigated whether MBH may support more

*complete determination of the structural composition of complex lipids. Indeed, after examining MBH between the two lipidomics platforms (LD ⇔ BM) and between the metabolomics and lipidomics platforms (HDF ⇔ BM) and (HDF ⇔ LD), we resolved the FA side chain composition for a number of complex lipids. The side chain composition of phosphatidylcholines measured on the BM platform, for instance, were delineated using MBH (e.g. PC_aa_C32:1 ⇔ PC(16:0/16:1), PC_aa_C40:6 ⇔ PC(18:0/22:6)) (**Supplementary Information Figure 2A**), in line with our previous study (Quell et al., 2019) . The characterization of fatty acid chains in triacylglycerol was similarly refined (TAG48:2-FA14:0 ⇔ myristoyl-linoleoyl-glycerol (14:0/18:2); TAG54:6-FA22:6 ⇔ palmitoyl-docosahexaenoyl-glycerl (16:0/22:6)) (**Supplementary Information Figure 2B**). These examples indicate how combining two technologically similar platforms can add valuable biological information, making them complimentary rather than redundant.*



Supplementary Information Figure 2. Omics platforms overlap and complementarity. A) & B) *The structure of complex lipids was revealed with by complementary platforms connected with MBH."*

Additionally, in the Supplementary information file **Note 1**, a part of the original manuscript, is responding to reviewer comments about differently annotated molecules. We were not including it in the main text, but we are referring to it as follows:

"While analyzing MBH between platforms capturing overlapping set of molecules (PM ⇔ HDF, OLINK ⇔ SOMA, and IgG ⇔ IgA) but utilizing different detection strategies (e.g. GC vs. LC; aptamer vs. antibody) we found that those display between 72% – 93% of concordance in respect

*of detected molecules, which underscores good quality of the selected methods applied in different laboratories (**Supplementary Information Note 1**).”*

“Supplementary Information Note 1

The largest number of MBHs was found between successive generations of the Metabolon platforms (PM ⇔ HDF), which is not surprising, given the technological similarity between them. Out of 369 identified hits, 291 paired identically annotated metabolites, 57 MBHs linked an unknown metabolite measured on the older platform generation to an annotated molecule measured on the more recent platform generation (e.g. X-18601 ⇔ androstenediol (3beta,17beta)-monosulfate), and 21 MBHs linked apparently differing molecules in related pathways (7 molecules; e.g. threonate ⇔ oxalate) or unknowns (14 molecules). As the Metabolon platforms differ with respect to the employed metabolite separation and detection methods (gas chromatography (GC) used for PM platform vs. liquid chromatography deployed for HDF (see methods)), this shows a robust concordance (79%) of platform performance and progressing component identification over time.

Detailed IgG glycosylation was determined by two independent glycomics platforms from two independent labs, referred to as IgG and IgA. Consequently, 29 out of 31 identified MBHs mapped to the same glycan structure, showing excellent agreement of 93% between both platforms.

*Two affinity proteomics platforms were used, one based on the SOMAscan aptamer technology (SOMA, 1129 traits), and the other OLINK technology based on antibody pairs implementing the proximity extension assay (OLINK, 184 traits). Of 72 proteins that overlapped between both platforms, 52 were linked by MBH (72%). We found that overlapping proteins, which were not captured by a MBH, showed low correlation (**Supplementary Information Table 1**), further suggesting that those might be susceptible to different analytical parameters, hence should be validated with alternative technical platforms to ensure the correctness of the measurement.”*

R2C2: With such deeply phenotyped population the reader of this journal might expect novel methodological efforts (and well-explained) on in-silico analyses, such as exploring more formal omics data fusion methods or attempts to address cross platform missingness via multiple imputation and other options of learning, going further than of simple linear models. The data are also suitable for reverse regression analysis i.e multi-phenotype GWAS. The authors have exclusive freedom to address such questions and would be useful for everyone as we are introduced to a new omics platform every other day. For example, instead of using 3 different versions of creatin (for example), can we use a PC (?) that captures creatin, or select the best one and generate a dataset with unique features only?

Response R2C2: We understand the reviewer's view and agree that multiple alternative analyses could be conducted. In fact, we also address computational frameworks in different studies, including recent work [28]. Nevertheless, with this manuscript, we are addressing different needs of the scientific community. We were motivated by the fact that currently, there is an increasing number of multiomics data which, after statistical analysis, requires to be interpreted by biochemists, biologists, and potentially, in the future, by medical doctors. Thus, our focus here was on questions related to molecular interactions, the biological relevance of such interactions, as well as data visualization and the potential for hypothesis generation carried by multiomics data. We were able to address those questions by:

1) Assessment of the MBH as a strategy for identifying molecular interactions across platforms within the same and across omics. We showed that, indeed, this approach could contribute to the identification of biologically relevant associations;

2) Incorporation of GGM's for determination of within-platform molecular interactions;

3) Analysis of multiomics GWAS, EWAS and TWAS;

4) Integration of identified association into auser-friendly server, which the larger scientific community can deploy for hypothesis generation. Additionally, we showed the benefits of multiomics integration on the example of diabetes subtypes.

We consider this work as a fusion resulting from an intensive collaborative effort conducted between bioinformaticians and biochemists dedicated specifically to the scientists interested in omics who are asking questions related to platform complementarity and biological relevance of

the multiomics associations rather than technical data analysis. This manuscript is unique and relevant for the broader readership of Nature Communication journal.

MINOR COMMENTS

R2C3: The methods and datasets that were used before can be shortened and mentioned in the supplementary, so that the reader can focus on the novel assets.

Response R2C3: The methods section was shortened following the reviewer's suggestion.

R2C4: Please check for typos, Comincs, or Comics? (Box2)

Response R2C4: Following the reviewer's comment, the typos were amended.

Reviewer #3 (Remarks to the Author):

The manuscript "The Molecular Human – A Roadmap of Molecular Interactions Linking Multiomics Networks with Disease Endpoints" by Halama et al. uses 391 participants from a multiethnic diabetes study with omic profiling from 18 different platforms to examine the interconnectedness of molecular data using multiple network approaches (partial correlations, GGM, mutual best correlations, etc.). The file "Molecular Human" is a molecular network has ~34,000 significant interconnections. Overall, this study is unique in the use of dense multiomic profiling and the characterization of the inter-relationships between all of these molecular profiles. What is most useful and will serve the overall research community well is the open access website: http://comics.metabolomix.com. This website stores these results and allows outside individuals to explore these interrelationships to further their own understanding, and potentially extended to own research fundings. This is a tremendous asset that to provide. While there are some points that need to be addressed and acknowledged as limitations (sample size, few disease endpoints, etc.). Overall, this study is novel and adds valuably to the scientific community. In addition, this work was performed using a minority population (QMDiab), where the understanding of these molecular relationships is particularly not well-understood.

Response: We thank the reviewers for their time and constructive comments and suggestions and hope to have fully addressed all points raised.

R3C1: These correlation networks shed light on the multimorbidity of omic variants across multiple platforms. This is highlighted well in examples presented (e.g., Figure 4). However, in the discussion, there is little commentary on what might be done with the multiomic networks. It

would be useful to elaborate on what understanding/hypotheses may be gleaned from the multiomic correlation networks through an example provided in the results.

Response R3C1: We agree with the reviewer that the discussion should also focus on the possibility of utilizing a multi-omics network to generate a hypothesis. This will direct the reader's attention to the information included in BOX 1-4, as well as our blog (http://www.metabolomix.com/comics/), where we continue to document new molecular case-studies under the vignette "Comics take on …". In response to the reviewer's comment, we included the following text in the discussion:

"…….*For instance, the multiomics network applied to MOD subgroup resulted in the identification of previously known as well as novel interactions as the one found between leptin and CXCL5, cytokine implicated in the chemotaxis of inflammatory cells (**Supplementary Information Note 10**). Given that CXCL5 was recently implicated in the browning of WAT* [29] *and leptin was determined as a molecule enabling browning of white adipose tissue (WAT)* [30,31]*, our identified association might suggest interplay between CXCL5 and leptin in the processes of WAT remodeling. This further extends our understanding of metabolically healthy obesity, characterized, among others, by high BMI and low insulin resistance*[32]*, which to some extent is characteristic of MOD subgroup. Despite the valuable insights provided by this multiomics integration, it is essential to note that the associations observed are hypothesis-generating in nature. Thus, further study would be required to provide definitive biological conclusions. Nevertheless, the perspective obtained through utilizing multiomics layers in understanding human biology in this study is relevant and can serve as a foundational framework for future multiomics initiatives.*

Additionally, our multiomics network, created based on molecular interactions across 18 platforms, is giving possibilities beyond molecular characterization of diabetes subtypes. It can be utilized in more generalizable approach to better understand molecular milieu (both direct and

*distant) of each measured molecule and consequently, for hypothesis generation as we outlined in **BOX1** – **BOX4** and **Figure 5** under vignette "COmics takes on". For instance, while analyzing the molecular network of 5-methyluridine we pointed out its potential immunosuppressive properties and suggested involvement of methylation and alteration in expression of e.g. IFI44L, EPSTI1, and LY6E genes, relevant in context of autoimmune diseases such as systemic lupus and rheumatoid arthritis [33,34]. We extend the effort of documenting new molecular "case studies" where a multiomics approach provides further insight into given molecule function and their potential involvement in various pathologies through identified omics associations. These case studies are presented in the form of a blog (http://www.metabolomix.com/comics/), depicted as 'Comics take on …'. Finally, we provide the scientific community with access to this multiomics network via the developed webserver COmics (http://comics.metabolomix.com) to facilitate global testing of the interactions of molecules of interest in the context of other omics layers. This can contribute to more rapid hypothesis generation, followed by its testing, and thus progress in the field."*

R3C2: While there is a description of what omic data types capture specific demographic features the best (e.g., metabolomics and proteomics), can further commentary be provided on the specific strengths of any given omic data type that are observed through these networks? Are there some omics that seem redundant? Are there important distinctions that are observed between what each omic layer captures? What may be captured by the synthesis of them together? Some of these points might be best summarize by looking at a selection of 2-way comparisons vs. fully multiomics using a subset of the variants. For example, is one omic datatype is missing, does a specific network fall apart because a hub is absent? While these are all general questions, further inquiry to synthesize some of these general points will add substantial impact to this work.

Response R3C2: The molecular network and spectrum of omics describing the given process/phenotype strictly depend on this process/phenotype, which is a main determinant of the "network hub". For instance, in the diabetes subtypes, for which the molecular network was constructed using multiple platforms (over 11 platforms contributed to the network), proteomics, metabolomics, and lipidomics were mainly assembling the network. Thus, the absence of any of

these three omics (metabolomics, proteomics, lipidomics) would restrain the formation of molecular networks for diabetes subtypes. In contrast, as described for TWAS, a 2-way comparison showing an association between transcriptomics and lipidomics or lipoproteomics was sufficient to reflect on the immunostimulatory process, and other omics were redundant in this context. Hence, the implementation of given omics is context and question-dependent. To make this point clear, we included the following in the discussion:

"*Yet, it is crucial to bear in mind that the implementation of such a broad array of platforms is frequently not feasible and not always necessary. The selection of specific omics and platforms should be driven by the scientific question, as well as the process or phenotype requiring investigation. As demonstrated in this study, phenotypes such as age, sex, or diabetes necessitate omics that closely recapitulate the specific phenotype. For example, metabolomics or proteomics were identified as the main molecular hubs enabling the construction of networks for diabetes subtypes, which was not feasible with transcriptomics or methylation alone, even though they are also components of the network. In contrast, associations revealed by TWAS, reflecting on immunostimulatory processes, were predominantly captured by transcriptomics and lipidomics/lipoproteomics, with other omics not contributing significantly to this discovery. This underscores the importance of 2-way comparisons rather than fully multiomics approaches in capturing certain processes. Now, with access to this data, each investigator has the freedom to monitor the molecular milieu of the molecule or phenotype of interest, allowing them to decide on the most suitable omics/platform approach for their study.*"

R3C3: While the manuscript does have multiple molecular interactions, the work described does not have multiple disease endpoints. While there are basic characteristics, such as age, sex, BMI, these are not disease endpoints. As such, either "multiple disease endpoints" should be modified to "diabetes" or something referring to population characteristics (age, sex). Otherwise, additional disease endpoints should be added.

Response R3C3: Following the reviewer's comment, the title was changed accordingly:

"The Molecular Human Overture – A Roadmap of Molecular Interactions Linking Multiomics Networks with Population Traits and Diabetes Subtypes".

R3C4: One concern/quandary with correlation networks is potential residual confounding. Since this is a case-control study where individuals were ascertained for diabetes. The authors did an excellent job describing the diabetes network in the manuscript; however, potential confounding effects of this in the overall correlation networks should be addressed as a limitation.

Response R3C4: We agree with the reviewer that the nature of our cohort could confound some of the associations. We think that this might be particularly relevant for the identified TWAS associations. In response to the reviewer's comment, we added the following text under study limitations:

"*Because our cohort consists of healthy and T2D subjects some of the observed associations could be driven by the molecular alterations which are known features of T2D (e.g. elevated carbohydrates, lipids, and branch chain amino acid levels). For instance, identified TWAS associations, dominated by 5 genes linked to various lipids could reflect on the enrolled participants characteristics, which might be recognized as study limitation. Nevertheless, such an experimental setting enabled us to uncover a range of lipids/lipoproteins with immunostimulatory properties in our lipidomics TWAS, which holds significance for cardiovascular disease.*"

R3C5: There is little discussion of this as a minority population. Was this adjusted for PCs? Also, is there anything here that may be specific to this minority group? Further comment about this as an advantage (e.g., few minority populations in the literature) and the potential limitations in generalizability should be addressed.

Response R3C5: We agree with the reviewer's comment, and we addressed it as follows:

"*Our study has strengths and weaknesses. The diversity of the QMDiab participants provides access to a wide range of individuals from various ethnicities including Arabs, South Asians, and Filipinos Given that the majority of the study focusses on Caucasian population, multiethnic*

nature of our work especially in multiomics context is truly unique and is adding to the previously conducted omics research on Asian and Middle eastern population[35-37]. Yet, mixed ethnicity in the QMDiab might result in population-specific stratification and thus in inflated p-values. Indeed, our previous study showed that the first three principal components (PC's) of the genotype variants capture self-reported ethnicity [38]. Therefore, we added the first three principal components of the genotyping data (genoPCs) to represent accurately the ethnicity."

R3C6: While the advantage of this cohort is the tremendous amount of multiomic data and there is no suitable replication cohort with the same omic data, is there any way to replicate some of the strongest observed correlations?

Response R3C6: The QMDiab study has served as a replication cohort, enabling a 2-way comparison of outcomes from various GWAS and EWAS studies[38-40], thereby underscoring the relevance of identified associations for traits previously examined. Moreover, several of our identified associations replicate findings from separate omics studies, conducted, however, without the comprehensive multiomics context and molecular network that our approach provides. Therefore, we conducted an extensive literature review and provided references to identified associations, which adds credibility to our study. Nevertheless, we agree with the reviewer that replication, especially in a multiomics context, would be important to conduct but is significantly limited. We hope that future studies, capturing multiple platforms and omics across different populations, will be conducted to enable replication and identification of new/alternative molecular networks that we expect might be discovered.

R3C7: A further consideration are the biofluids and omics that are missing. Most notably stool metabolomics and microbiome data. If this is the "Molecular Human" further comment on what is missing should be considered.

Response R3C7: Following the reviewer's suggestion, we included a short paragraph in the discussion section as outlined below:

"*The Molecular Human could be considered as holistic description of molecular interactions in the human body, which we achieved here by integrating molecules detected across 18 platforms and 8 omics. Although, to date this is the largest effort in terms of the number of measurements conducted in the relatively big human cohort, future attempts extending our understanding of the*

molecular interactome towards process concerning in greater detail secretome by focusing on sweat and tears, exhalome focused on the molecular composition of the breath as well as microbiome aiming to provide comprehensive description of the gut and skin microbiota and their interactions with the host will be critical. Thus, we see our approach as an overture into future large-scale multiomics study for which we are setting a stage."

MINOR

R3C8: The metabolomic acronyms in the tables are almost too short, so it is not easy for the reader to quickly digest what molecular platform is what. Please consider either adding an additional column that lists the specific omic type or adding to the acronym to make it clear (e.g., MetP; MetS, etc.)

Response R3C8: We agree with the reviewer that some acronyms are short. Therefore, we described for each of the acronyms in **Table 1** (please see below) as well as in **Supplementary Table 1**.

Table 1. Overview on applied omics technologies.

Omics	Measurement/Output	Technique/Platform	Matrix	Label
GENOMICS	Genotype	Infinium Human Omni 2.5-8 v1.2 BeadChip kit	DNA extracted from buffy coat fraction from whole blood	DNA
METHYLOMICS	DNA methylation	Illumina Infinium HumanMethylation450 BeadChip kit	DNA, same as for genomics	MET
TRANSCRIPTOMICS	Gene expression	RNA-sequencing based Illumina ~20M reads	RNA extracted from PaxTube	RNA
	microRNA expression	microRNA profiling based multiplex qPCR, Exiqon	RNA extracted from EDTA plasma	miRNA
PROTEOMICS	Relative protein abundance	Slow Off-rate Modified Aptamer (SOMAmer), Somalogic 1,1k	EDTA plasma	SOMA
	Relative protein abundance	Proximity Extension Assay (PEA) based Olink Target 96 Metabolism &	Heparin plasma	OLINK

		Cardiometabolism panels		
GLYCOMICS	Total plasma N-glycosylation	Hydrophilic interaction ultra-performance liquid chromatography (HILIC-UPLC) based Genos pipeline	EDTA plasma	PGP
	IgG glycosylation	Liquid chromatography mass spectrometry (LC-MS) based Genos pipeline	EDTA plasma	IgG
	IgA & IgG glycosylation	LC-MS based [41] pipeline	EDTA plasma	IgA
LIPOPROTEOMICS	Lipoproteins	Proton nuclear magnetic resonance (^{1}H NMR) based Nightingale technology	EDTA plasma	BRAIN
LIPIDOMICS	Absolute lipid concentration	LC-MS based on Lipidyzer technology at Metabolon	EDTA plasma	LD
	Lipids and other metabolite concentration	Flow injection analysis (FIA)- MS based Biocrates technology	EDTA plasma	BM
METABOLOMICS	Metabolite level	(HILIC-MS) & (UPLC-MS) based HD4 Metabolon	EDTA plasma	HDF
	Metabolite level	Gas chromatography (GC)-MS (UPLC-MS) based HD2 Metabolon	EDTA plasma	PM
	Metabolite level	(GC-MS) & (UPLC-MS) based HD2 Metabolon	Saliva	SM
	Metabolite level	GC-MS & UPLC-MS based HD2 Metabolon	Urine	UM
	Metabolite level	^{1}H NMR deploying Chenomx for annotation, based on [42] pipeline	Urine	CM
CLINICAL	Clinical biochemistry and blood counts	Cobas 6000; Roche Diagnostics	Blood/Urine	CLIN

References

1. Campeau, A.*, et al.* Multi-omics of human plasma reveals molecular features of dysregulated inflammation and accelerated aging in schizophrenia. *Mol Psychiatry* **27**, 1217-1225 (2022).
2. Marabita, F.*, et al.* Multiomics and digital monitoring during lifestyle changes reveal independent dimensions of human biology and health. *Cell Syst* **13**, 241-255 e247 (2022).
3. Sailani, M.R.*, et al.* Deep longitudinal multiomics profiling reveals two biological seasonal patterns in California. *Nature Communications* **11**(2020).
4. Benson, M.D.*, et al.* Protein-metabolite association studies identify novel proteomic determinants of metabolite levels in human plasma. *Cell Metab* **35**, 1646-1660 e1643 (2023).
5. Mikaeloff, F.*, et al.* Network-based multi-omics integration reveals metabolic at-risk profile within treated HIV-infection. *Elife* **12**(2023).
6. Shi, L.*, et al.* Multiomics profiling of human plasma and cerebrospinal fluid reveals ATN-derived networks and highlights causal links in Alzheimer's disease. *Alzheimers Dement* **19**, 3350-3364 (2023).
7. Garrett-Bakelman, F.E.*, et al.* The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* **364**(2019).
8. Contrepois, K.*, et al.* Molecular Choreography of Acute Exercise. *Cell* **181**, 1112-1130 e1116 (2020).
9. Chen, R.*, et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293-1307 (2012).
10. Tebani, A.*, et al.* Integration of molecular profiles in a longitudinal wellness profiling cohort. *Nat Commun* **11**, 4487 (2020).
11. Jeffcoate, S.L. Diabetes control and complications: The role of glycated haemoglobin, 25 years on. Vol. 21 657-665 (Diabet Med, 2004).
12. Rahbar, S., Blumenfeld, O. & Ranney, H.M. Studies of an unusual hemoglobin in patients with diabetes mellitus. *Biochemical and Biophysical Research Communications* **36**, 838-843 (1969).
13. Lever, M.*, et al.* Variability of plasma and urine betaine in diabetes mellitus and its relationship to methionine load test responses: An observational study. *Cardiovascular Diabetology* **11**(2012).
14. Mardinoglu, A.*, et al.* Plasma Mannose Levels Are Associated with Incident Type 2 Diabetes and Cardiovascular Disease. Vol. 26 281-283 (Cell Press, 2017).
15. Yousri, N.A.*, et al.* A systems view of type 2 diabetes-associated metabolic perturbations in saliva, blood and urine at different timescales of glycaemic control. *Diabetologia* **58**, 1855-1867 (2015).
16. Contreras, P., Generini, G., Michelsen, H., Pumarino, H. & Campino, C. Hyperprolactinemia and galactorrhea: Spontaneous versus iatrogenic hypothyroidism. *Journal of Clinical Endocrinology and Metabolism* **53**, 1036-1039 (1981).
17. Suhre, K.*, et al.* Fine-Mapping of the Human Blood Plasma N-Glycome onto Its Proteome. *Metabolites* **9**(2019).
18. Richardson, C.R.*, et al.* Management of Type 2 Diabetes Mellitus. (2021).
19. Ahlqvist, E.*, et al.* Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol* **6**, 361-369 (2018).

20. Wagner, G.P. The biological homology concept. *Annual review of ecology and systematics. Vol. 20*, 51-69 (1989).
21. Brown, T.A. Molecular Phylogenetics. (2002).
22. Elemento, O., Gascuel, O. & Lefranc, M.-P. Reconstructing the duplication history of tandemly repeated genes. *Molecular biology and evolution* **19**, 278-288 (2002).
23. Fitch, W.M. Homologya personal view on some of the problems. Vol. 16 227-231 (Trends Genet, 2000).
24. Gabaldón, T. & Koonin, E.V. Functional and evolutionary implications of gene orthology. *Nature reviews. Genetics* **14**, 360-366 (2013).
25. Overbeek, R., Fonstein, M., D'Souza, M., Push, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2896-2901 (1999).
26. Tatusov, R.L.*, et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22-28 (2001).
27. Quell, J.D.*, et al.* Characterization of Bulk Phosphatidylcholine Compositions in Human Plasma Using Side-Chain Resolving Lipidomics. *Metabolites* **9**, 109-109 (2019).
28. AutoFocus: A hierarchical framework to explore multi-omic disease associations spanning multiple scales of biomolecular interaction. *bioRxiv* (2023).
29. Lee, D.*, et al.* CXCL5 secreted from macrophages during cold exposure mediates white adipose tissue browning. *J Lipid Res* **62**, 100117 (2021).
30. Sarmiento, U.*, et al.* Morphologic and molecular changes induced by recombinant human leptin in the white and brown adipose tissues of C57BL/6 mice. *Lab Invest* **77**, 243-256 (1997).
31. Dodd, G.T.*, et al.* Leptin and insulin act on POMC neurons to promote the browning of white fat. *Cell* **160**, 88-104 (2015).
32. Bluher, M. Metabolically Healthy Obesity. *Endocr Rev* **41**(2020).
33. Cooles, F.A.H.*, et al.* Interferon-alpha-mediated therapeutic resistance in early rheumatoid arthritis implicates epigenetic reprogramming. *Ann Rheum Dis* **81**, 1214-1223 (2022).
34. Zhao, M.*, et al.* IFI44L promoter methylation as a blood biomarker for systemic lupus erythematosus. *Ann Rheum Dis* **75**, 1998-2006 (2016).
35. Pan, H.*, et al.* Integrative multi-omics database (iMOMdb) of Asian pregnant women. *Hum Mol Genet* **31**, 3051-3067 (2022).
36. Saw, W.Y.*, et al.* Establishing multiple omics baselines for three Southeast Asian populations in the Singapore Integrative Omics Study. *Nat Commun* **8**, 653 (2017).
37. Yousri, N.A., Albagha, O.M.E. & Hunt, S.C. Integrated epigenome, whole genome sequence and metabolome analyses identify novel multi-omics pathways in type 2 diabetes: a Middle Eastern study. *BMC Med* **21**, 347 (2023).
38. Suhre, K.*, et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications* **8**(2017).
39. Gudmundsdottir, V.*, et al.* Circulating Protein Signatures and Causal Candidates for Type 2 Diabetes. *Diabetes* **69**, 1843-1853 (2020).
40. Sharapov, S.Z.*, et al.* Defining the genetic control of human blood plasma N-glycome using genome-wide association study. *Human molecular genetics* **28**, 2062-2077 (2019).

41. Momcilovic, A.*, et al.* Simultaneous Immunoglobulin A and G Glycopeptide Profiling for High-Throughput Applications. *Anal Chem* **92**, 4518-4526 (2020).
42. Budde, K.*, et al.* Quality assurance in the pre-analytical phase of human urine samples by 1H NMR spectroscopy. *Archives of Biochemistry and Biophysics* **589**, 10-17 (2016).

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have resolved most of the concerns. However, I do have a couple that remain:

In the response for R1C1 the new paragraph addresses the previous concern that related work was not addressed in the manuscript. However, the new paragraph, starting on line 89 of the revised manuscript contains numerous grammatical errors that make understanding the writing difficult. For example, on line 91, "study utilizing broader array of omics, frequently applied to a limited number of subjects," is probably meant to say something like "studies that have utilized a broad array of omics data have often only had a limited number of subjects," or something along those lines. The sentence starting on line 94 has similar issues. On line 96 the incorrect use of an article is confusing, "the lifestyle changes study," leaves one thinking that this study has already been discussed. That issue is also repeated on line 99 and other issues are present as well. This paragraph requires extensive editing.

The response to R1C3 is an improvement. Also, the authors are correct that this particular pseudo-R2 can provide a comparable measure across datasets, because it is based on the MSE and the dependent variable is the same in each case (so the MSEs should be comparable). However, presumably this only applies to the continuous variables. For diabetes state, the question remains as to how variance explained is quantified.

Reviewer #2 (Remarks to the Author):

I thank the authors you for their response. In their response the authors report that “we found that those display between 72% – 93% of concordance in respect of detected molecules” which I think is crucial information and one of the main findings of this paper but it is brushed over. This shows that the data which they pull together in raw form contains apples and oranges, and among every four MBH discovered in this setting, one of them is potentially wrong, depending on the pairs of platforms the analytes are originating from. I suggest they also include all the (dis)concordance results in a table, highlight them and suggest cross platform specific error /confidence rates for the reader and think of removing platforms that yield low concordance.

I am afraid that not emphasizing these facts will definitely mislead the user of the database and the reader of the paper.

Minor comments

Please check for use of abbreviations e.g APOE repeating in the old and new text, and for typos.

Gene names should be italic unless asked differently by the journal.

Reviewer #3 (Remarks to the Author):

The authors have sufficiently addresses my comments and concerns. This is an outstanding manuscript and will serve the scientific community.

The Molecular Human Overture – A Roadmap of Molecular Interactions Linking Multiomics Networks with Population Traits and Diabetes Subtypes

Response to the reviewers' comments

We thank the reviewers for their time and constructive feedback. Please find below our point-by-point response. Additionally, we provide a change-tracked version of the revised manuscript where all modifications are highlighted.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have resolved most of the concerns. However, I do have a couple that remain:

Response: We thank the reviewers for their constructive comments and suggestions and hope to have fully addressed all points raised.

COMMENTS

R1C1: In the response for R1C1 the new paragraph addresses the previous concern that related work was not addressed in the manuscript. However, the new paragraph, starting on line 89 of the revised manuscript contains numerous grammatical errors that make understanding the writing difficult. For example, on line 91, "study utilizing broader array of omics, frequently applied to a limited number of subjects," is probably meant to say something like "studies that have utilized a broad array of omics data have often only had a limited number of subjects," or something along those lines. The sentence starting on line 94 has similar issues. On line 96 the incorrect use of an article is confusing, "the lifestyle changes study," leaves one thinking that this study has already been discussed. That issue is also repeated on line 99 and other issues are present as well. This paragraph requires extensive editing.

Response R1C1: Thank you for the comment the text was extensively revised and was changed as follows:

"Indeed, molecular processes have been monitored in human biofluids through the integration of various omics approaches, including genomics, methylation, transcriptomics, proteomics, and metabolomics [7,13-21]. However, studies that deploy a broader range of omics techniques tend to have a smaller sample size, typically involving 1 to 36 individuals. For instance, these studies investigate dynamic changes in diverse molecular components in response to factors such as viral infection[7], spaceflight[19], as well as extensive exercise[20]. In contrast, larger cohort studies (≥100 individuals) tend to focus on a more limited spectrum of omics measurements [13-15,17]. For example, the impact of lifestyle changes was monitored at the molecular level in processes related to obesity, diabetes, liver function, or cardiovascular disease using genomics, proteomics, and metabolomics[14]. Similarly, proteomics and metabolomics were deployed to determine molecular signatures associated with schizophrenia[13], while metabolomics and lipidomics were used for studying HIV infection[17] . The limited array of omics approaches was also used in a very large population study where the cohort size exceeds 1000 subjects. For instance, genomics, proteomics and metabolomics, were used to monitor metabolite-protein interactions in over 3,600 healthy subjects[16]. Additionally, they were also employed to investigate the molecular network related to Alzheimer's disease based on the molecular alterations measured in over 1200 subjects[18]."

R1C2: The response to R1C3 is an improvement. Also, the authors are correct that this particular pseudo-R2 can provide a comparable measure across datasets, because it is based on the MSE and the dependent variable is the same in each case (so the MSEs should be comparable). However, presumably this only applies to the continuous variables. For diabetes state, the question remains as to how variance explained is quantified.

Response R1C2: The reviewer is correct. We omitted to explain the case of the categorial variables sex and diabetes state. We added the following text to the manuscript and changed the term paragraph header from "Percentage of the variance explained in age, sex, BMI and diabetes state by platform" to "Potential to predict age, sex, BMI and diabetes state by platform" in order to better reflect the specific definition of these measures:

"Potential to predict age, sex, BMI and diabetes state by platform.

For continuous variables we use the "pseudo-R2" reported by the R package randomForest as an estimate of how well a given omics phenotype can predict age, sex, BMI and diabetes state. For continuous variables this "pseudo-R2" is defined as one minus the mean square error of the regression divided by the variance of the dependent variable. Note that the "pseudo-R2" is not a strict measure of the explained variance and is used here to provide an intuition for the quality of the model fit that can be obtained using the different omics datasets.

For categorial variables we use the "Out-Of-Bag Error" (OOBErr) estimate from R, scaled to range between zero and one (1-OOBerr)/(1-OOBerr$_{rnd}$), where OOBErr$_{rnd}$ was estimated by randomizing the sample identifiers."

Reviewer #2 (Remarks to the Author):

R2C1: I thank the authors you for their response. In their response the authors report that "we found that those display between 72% – 93% of concordance in respect of detected molecules" which I think is crucial information and one of the main findings of this paper but it is brushed over. This shows that the data which they pull together in raw form contains apples and oranges, and among every four MBH discovered in this setting, one of them is potentially wrong, depending on the pairs of platforms the analytes are originating from. I suggest they also include all the (dis)concordance results in a table, highlight them and suggest cross platform specific error /confidence rates for the reader and think of removing platforms that yield low concordance.

I am afraid that not emphasizing these facts will definitely mislead the user of the database and the reader of the paper.

Response R2C1: We thank the reviewers for their constructive comments and suggestions and hope to have fully addressed all points raised.

We think that the reviewer might have overseen the description of MBH provided in the **Supplementary Information Note 1**, where indeed we focused on the MBH capturing molecules measured at technologically different platforms within the same omics. We outlined examples of the molecules which were not showing the concordance further focusing on the platform with 72% of MBH concordance namely OLINK and SOMA. We have also listed all the molecules

measured and overlapping between SOMA and OLINK (**Supplementary Information Table 1**) and showed that overlapping molecules which were not captured by MBH had low correlation. To make the **Supplementary Information Table 1** clearer, an additional column informing about MBH status was added. We further suggested to evaluate the aspects related to platform performance by utilization of GWAS as an alternative measure, which was outlined in **Supplementary Information Note 2**.

Additionally in the discussion section we comment on the 28% of the proteins which were not detected by the MBH.

We believe this description sufficiently clarifies any potential confusion for the reader.

Please see below the information captured in Supplementary Information Note 1 and 2 as well as in Discussion.

Supplementary Information Note 1

"*The largest number of MBHs was found between successive generations of the Metabolon platforms (PM ⇔ HDF), which is not surprising, given the technological similarity between them. Out of 369 identified hits, 291 paired identically annotated metabolites, 57 MBHs linked an unknown metabolite measured on the older platform generation to an annotated molecule measured on the more recent platform generation (e.g. X-18601 ⇔ androstenediol (3beta,17beta)-monosulfate), and 21 MBHs linked apparently differing molecules in related pathways (7 molecules; e.g. threonate ⇔ oxalate) or unknowns (14 molecules). As the Metabolon platforms differ with respect to the employed metabolite separation and detection methods (gas chromatography (GC) used for PM platform vs. liquid chromatography deployed for HDF (see methods)), this shows a robust concordance (79%) of platform performance and progressing component identification over time.*

Detailed IgG glycosylation was determined by two independent glycomics platforms from two independent labs, referred to as IgG and IgA. Consequently, 29 out of 31 identified MBHs mapped to the same glycan structure, showing excellent agreement of 93% between both platforms.

Two affinity proteomics platforms were used, one based on the SOMAscan aptamer technology (SOMA, 1129 traits), and the other OLINK technology based on antibody pairs implementing the proximity extension assay (OLINK, 184 traits). Of 72 proteins that overlapped between both platforms, 52 were linked by MBH (72%). We found that overlapping proteins, which were not captured by a MBH, showed low correlation (Supplementary Information Table 1), further suggesting that those might be susceptible to different analytical parameters, hence should be validated with alternative technical platforms to ensure the correctness of the measurement."

Supplementary Information Note 2

"*We have investigated various GWAS associations to evaluate the performance of different platforms. For instance, SNP rs1047891 associated with glycine with a p-value = 7.4 x10^{-18} in 320 samples on targeted lipidomics BM platform, with p-value = 7.4 x10^{-15} in 291 samples on the HDF platform, and with p-value = 7.1 x10^{-14} in 322 samples on the PM platform (**Supplementary Information Figure 1A**). In this example, the targeted assay appears to provide stronger signals, at least compared to the older PM platform, which was measured using an older generation of metabolic profiling described as HD2 by Metabolon. In another example however SNP rs1799958 associate with butyrylcarnitine (C4) with p-value = 7.9 x10^{-9} in 323 samples on targeted lipidomics BM platform, and p-value = 2.3 x10^{-15} in 294 on the HDF platform, and with p-value = 1.6 x10^{-23} in 325 samples on the PM platform, suggesting that platform performance might depend on the molecule properties and how well a given molecule is captured by each platform (**Supplementary Information Figure 1B**). We have observed similar trend across other platforms including measurements of urine metabolome with CM and UM (for SNP rs9922704 with 3-hydroxyisovalerate (**Supplementary Information Figure 1C**)) and with plasma proteome SOMA and OLINK (for rs3896287 with LILRB2 (**Supplementary Information Figure 1D**); for SNP rs8176693 with TIE1 (**Supplementary Figure 2E**); for SNP rs9892586 with CCL14 (**Supplementary Information Figure 1E**)."*

Discussion

"However, approximately 28% of common protein targets were not detected by the MBH for the two affinity-based platforms used for proteomics. This suggests that integrating the measurement approaches can be challenging and may require special attention for the molecules that MBH did not identify. Many of our observations are in line with previous studies assessing proteomics methods in multiple cohorts [93]. They could be linked to differences in their analytical performance for common protein targets."

Supplementary Information Table 1. Correlation levels of molecules measured on both SOMA and OLINK

OLINK	SOMA	N	r	MBH Y/N	p-value
CCL18_P55774 [OLINK CMET]	CCL18 : C-C motif chemokine 18 [SOMA]	322	0.87	Y	$7.9x10^{-100}$
ENPP7_Q6UWV6 [OLINK MET]	ENPP7 : Ectonucleotide pyrophosphatase/phosphodiesterase family member 7 [SOMA]	314	0.85	Y	$3.01x10^{-91}$
MBL2_P11226 [OLINK CMET]	MBL2 : Mannose-binding protein C [SOMA]	322	0.84	Y	$2.7x10^{-88}$
PRSS2_P07478 [OLINK CMET]	PRSS2 : Trypsin-2 [SOMA]	322	0.84	Y	$6.97x10^{-87}$
CNDP1_Q96KN2 [OLINK CMET]	CNDP1 : Beta-Ala-His dipeptidase [SOMA]	322	0.84	Y	$2.26x10^{-86}$
CA3_P07451 [OLINK CMET]	CA3 : Carbonic anhydrase 3 [SOMA]	322	0.8	Y	$1.22x10^{-72}$
IGFBP3_P17936 [OLINK CMET]	IGFBP3 : Insulin-like growth factor-binding protein 3 [SOMA]	322	0.79	Y	$3.98x10^{-70}$
ANG_P03950 [OLINK CMET]	ANG : Angiogenin [SOMA]	322	0.75	Y	$1.29x10^{-60}$
ANGPT2_O15123 [OLINK MET]	ANGPT2 : Angiopoietin-2 [SOMA]	314	0.76	Y	$1.07x10^{-59}$
IGFBP6_P24592 [OLINK CMET]	IGFBP6 : Insulin-like growth factor-binding protein 6 [SOMA]	322	0.74	Y	$1.68x10^{-56}$
CST3_P01034 [OLINK CMET]	CST3 : Cystatin-C [SOMA]	322	0.73	Y	$2.2x10^{-55}$
SELL_P14151 [OLINK CMET]	SELL : L-Selectin [SOMA]	322	0.72	Y	$1.72x10^{-52}$
VCAM1_P19320 [OLINK CMET]	VCAM1 : Vascular cell adhesion protein 1 [SOMA]	322	0.71	Y	$1.98x10^{-51}$
SERPINA7_P05543 [OLINK CMET]	SERPINA7 : Thyroxine-binding globulin [SOMA]	322	0.71	Y	$2.53x10^{-51}$
CCDC80_Q76M96 [OLINK MET]	CCDC80 : Coiled-coil domain-containing protein 80 [SOMA]	314	0.7	Y	$1.64x10^{-48}$
NRP1_O14786 [OLINK CMET]	NRP1 : Neuropilin-1 [SOMA]	322	0.66	Y	$2.23x10^{-42}$

TIE1_P35590 [OLINK CMET]	TIE1 : Tyrosine-protein kinase receptor Tie-1, soluble [SOMA]	322	0.65	Y	$3.09x10^{-40}$
THBS4_P35443 [OLINK CMET]	THBS4 : Thrombospondin-4 [SOMA]	322	0.64	Y	$2.99x10^{-39}$
RTN4R_Q9BZR6 [OLINK MET]	RTN4R : Reticulon-4 receptor [SOMA]	314	0.62	Y	$3.65x10^{-35}$
F11_P03951 [OLINK CMET]	F11 : Coagulation Factor XI [SOMA]	322	0.6	Y	$2.36x10^{-33}$
F7_P08709 [OLINK CMET]	F7 : Coagulation factor VII [SOMA]	322	0.6	Y	$2.49x10^{-33}$
KIT_P10721 [OLINK CMET]	KIT : Mast/stem cell growth factor receptor Kit [SOMA]	322	0.6	Y	$4.14x10^{-33}$
ROR1_Q01973 [OLINK MET]	ROR1 : Tyrosine-protein kinase transmembrane receptor ROR1 [SOMA]	314	0.6	Y	$1.27x10^{-32}$
SIGLEC7_Q9Y286 [OLINK MET]	SIGLEC7 : Sialic acid-binding Ig-like lectin 7 [SOMA]	314	0.59	Y	$2.55x10^{-31}$
C2_P06681 [OLINK CMET]	C2 : Complement C2 [SOMA]	322	0.58	Y	$3.37x10^{-30}$
TIMP1_P01033 [OLINK CMET]	TIMP1 : Metalloproteinase inhibitor 1 [SOMA]	322	0.57	Y	$9.07x10^{-30}$
FAP_Q12884 [OLINK CMET]	FAP : Prolyl endopeptidase FAP [SOMA]	322	0.57	Y	$2.24x10^{-29}$
GP1BA_P07359 [OLINK CMET]	GP1BA : Platelet glycoprotein Ib alpha chain [SOMA]	322	0.57	Y	$9.85x10^{-29}$
CHL1_O00533 [OLINK CMET]	CHL1 : Neural cell adhesion molecule L1-like protein [SOMA]	322	0.55	Y	$1.97x10^{-27}$
NCAM1_P13591 [OLINK CMET]	NCAM1 : Neural cell adhesion molecule 1, 120 kDa isoform [SOMA]	322	0.55	Y	$5.1x10^{-27}$
TNC_P24821 [OLINK CMET]	TNC : Tenascin [SOMA]	320	0.54	Y	$4.23x10^{-26}$
PLA2G7_Q13093 [OLINK CMET]	PLA2G7 : Platelet-activating factor acetylhydrolase [SOMA]	322	0.52	Y	$7.42x10^{-24}$
NID1_P14543 [OLINK CMET]	NID1 : Nidogen-1 [SOMA]	322	0.52	Y	$2.41x10^{-23}$
CA1_P00915 [OLINK CMET]	CA1 : Carbonic anhydrase 1 [SOMA]	322	0.5	Y	$3.49x10^{-22}$
LILRB2_Q8N423 [OLINK CMET]	LILRB2 : Leukocyte immunoglobulin-like receptor subfamily B member 2 [SOMA]	322	0.5	Y	$4.07x10^{-22}$
FCN2_Q15485 [OLINK CMET]	FCN2 : Ficolin-2 [SOMA]	322	0.5	Y	$1x10^{-21}$
LYVE1_Q9Y5Y7 [OLINK CMET]	LYVE1 : Lymphatic vessel endothelial hyaluronic acid receptor 1 [SOMA]	322	0.49	Y	$2.35x10^{-21}$
CFHR5_Q9BXR6 [OLINK CMET]	CFHR5 : Complement factor H-related protein 5 [SOMA]	322	0.49	Y	$1.15x10^{-20}$
ADGRE2_Q9UHX3 [OLINK MET]	ADGRE2 : Adhesion G protein-coupled receptor E2 [SOMA]	314	0.48	Y	$3.58x10^{-19}$
COL18A1_P39060 [OLINK CMET]	COL18A1 : Endostatin [SOMA]	322	0.47	N	$8.85x10^{-19}$

TGFBR3_Q03167 [OLINK CMET]	TGFBR3 : Transforming growth factor beta receptor type 3 [SOMA]	322	0.46	Y	$6.42x10^{-18}$
SPARCL1_Q14515 [OLINK CMET]	SPARCL1 : SPARC-like protein 1 [SOMA]	322	0.44	Y	$8.62x10^{-17}$
DDC_P20711 [OLINK MET]	DDC : Aromatic-L-amino-acid decarboxylase [SOMA]	314	0.43	Y	$7.01X10^{-16}$
MET_P08581 [OLINK CMET]	MET : Hepatocyte growth factor receptor [SOMA]	322	0.4	Y	$1.29X10^{-13}$
CCL14_Q16627 [OLINK CMET]	CCL14 : C-C motif chemokine 14 [SOMA]	322	0.39	N	$4.50X10^{-13}$
CCL5_P13501 [OLINK CMET]	CCL5 : C-C motif chemokine 5 [SOMA]	322	0.39	Y	$5.72X10^{-13}$
CA13_Q8N1Q1 [OLINK MET]	CA13 : Carbonic anhydrase 13 [SOMA]	314	0.38	N	$1.44X10^{-12}$
TGFBI_Q15582 [OLINK CMET]	TGFBI : Transforming growth factor-beta-induced protein ig-h3 [SOMA]	322	0.36	Y	$2.72X10^{-11}$
LCN2_P80188 [OLINK CMET]	LCN2 : Neutrophil gelatinase-associated lipocalin [SOMA]	322	0.36	Y	$3.81X10^{-11}$
FETUB_Q9UGM5 [OLINK CMET]	FETUB : Fetuin-B [SOMA]	322	0.34	Y	$3.90X10^{-10}$
FCGR3B_O75015 [OLINK CMET]	FCGR3B : Low affinity immunoglobulin gamma Fc region receptor III-B [SOMA]	322	0.32	N	$6.62X10^{-9}$
SERPINA5_P05154 [OLINK CMET]	SERPINA5 : Plasma serine protease inhibitor [SOMA]	322	0.3	Y	$5.32X10^{-8}$
ENTPD5_O75356 [OLINK MET]	ENTPD5 : Ectonucleoside triphosphate diphosphohydrolase 5 [SOMA]	314	0.29	Y	$1.01X10^{-7}$
LILRB1_Q8NHL6 [OLINK CMET]	LILRB1 : Leukocyte immunoglobulin-like receptor subfamily B member 1 [SOMA]	322	0.27	N	$1.06X10^{-6}$
SOD1_P00441 [OLINK CMET]	SOD1 : Superoxide dismutase [Cu-Zn] [SOMA]	322	0.24	N	$1.71X10^{-5}$
CDH1_P12830 [OLINK CMET]	CDH1 : Cadherin-1 [SOMA]	322	0.14	N	$9.57X10^{-3}$
GNLY_P22749 [OLINK CMET]	GNLY : Granulysin [SOMA]	322	0.13	Y	$1.61X10^{-2}$
IL7R_P16871 [OLINK CMET]	IL7R : Interleukin-7 receptor subunit alpha [SOMA]	322	0.13	N	$1.64X10^{-2}$
FCGR2A_P12318 [OLINK CMET]	FCGR2A : Low affinity immunoglobulin gamma Fc region receptor II-a [SOMA]	322	0.13	N	$1.77X10^{-2}$
DPP7_Q9UHL4 [OLINK MET]	DPP7 : Dipeptidyl peptidase 2 [SOMA]	314	-0.13	N	$2.61X10^{-2}$
CA4_P22748 [OLINK CMET]	CA4 : Carbonic anhydrase 4 [SOMA]	322	-0.11	N	$5.78X10^{-2}$
GRAP2_O75791 [OLINK MET]	GRAP2 : GRB2-related adapter protein 2 [SOMA]	314	-0.11	N	$5.89X10^{-2}$
CTSH_P09668 [OLINK MET]	CTSH : Cathepsin H [SOMA]	314	0.09	N	$1.23X10^{-1}$
NOTCH1_P46531 [OLINK CMET]	NOTCH1 : Neurogenic locus notch homolog protein 1 [SOMA]	322	0.08	N	$1.77X10^{-1}$

ENG_P17813 [OLINK CMET]	ENG : Endoglin [SOMA]	322	0.07	N	$2.10X10^{-1}$
DIABLO_Q9NR28 [OLINK MET]	DIABLO : Diablo homolog, mitochondrial [SOMA]	314	0.07	N	$2.33X10^{-1}$
ARG1_P05089 [OLINK MET]	ARG1 : Arginase-1 [SOMA]	314	-0.06	N	$2.52X10^{-1}$
ICAM3_P32942 [OLINK CMET]	ICAM3 : Intercellular adhesion molecule 3 [SOMA]	322	-0.05	N	$4.05X10^{-1}$
CDH2_P19022 [OLINK MET]	CDH2 : Cadherin-2 [SOMA]	314	0.04	N	$4.92X10^{-1}$
ANGPTL3_Q9Y5C1 [OLINK CMET]	ANGPTL3 : Angiopoietin-related protein 3 [SOMA]	322	0.03	N	$6.09X10^{-1}$
TYRO3_Q06418 [OLINK MET]	TYRO3 : Tyrosine-protein kinase receptor TYRO3 [SOMA]	314	0.02	N	$6.78X10^{-1}$
ICAM1_P05362 [OLINK CMET]	ICAM1 : Intercellular adhesion molecule 1 [SOMA]	322	-0.01	N	$8.59X10^{-1}$

Minor

R2C2: Please check for use of abbreviations e.g APOE repeating in the old and new text, and for typos.

Gene names should be italic unless asked differently by the journal.

Response: The text was revised, and the gene names were changed to italic.

Reviewer #3 (Remarks to the Author):

The authors have sufficiently addresses my comments and concerns. This is an outstanding manuscript and will serve the scientific community.

Response: We appreciate the reviewers for their time and this positive feedback.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have now satisfactorily answered my previous concerns.

Reviewer #2 (Remarks to the Author):

The authors responded the comments and made the necessary amendments.