Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Ilia Kurochkin

PhD Program: Life Sciences

Title of Thesis: Comparative analysis of human brain based on mass spectrometry data

Supervisor: Prof. Philipp Khaitovich

Chair of PhD defense Jury: Prof. Mikhail Gelfand **Email:** *m.gelfand@skoltech.ru*

Date of Thesis Defense: 26 October 2018

The thesis document includes the following changes in answer to the external review process.

Reviewer: Prof. Christoph W. Turck

Comment 1

Some discussion on the possible applications of the identified lipid and metabolite information would be appropriate. Better patient group stratification tools are in great demand in psychiatry! Specific molecular markers detectable in the periphery would revolutionize current diagnostic procedures that are mainly based on imprecise rating scales.

Answer

Following the Reviewer's suggestion, I have added the possibility of potential diagnostic application of the results into the conclusion section of the thesis.

Reviewer: Dr Amaury Cazenave Gassiot

Comment 1

It is almost a minor point, but the presentation and classification of metabolites and lipids may be a little too general in the introductory paragraphs 2.4.1 and 2.5.1. Specifically, the author rounds up all vitamins as metabolites, although vitamins A, D, E, K are fat soluble and generally classified as lipids. On page 15, it is stated that "lipids are classified into three major classes", this seems like an oversimplification that hardly represents the chemical diversity of lipids. The author recognises already at the end of the same paragraph that TG constitute a fourth class, while the LipidMaps consortium catalogues at least eight classes of lipids.

Answer

Following the Reviewer's suggestion, I have rewritten the chapter using lipid class annotation used in the common resource, LIPID MAPS, and added a description of the missing classes. More general separation of lipids and metabolites in the thesis, for practical reasons, is based on the extraction procedure.

Comment 2

Figure 2.2, a summary of pathway enrichment in ASD in twelve published studies, highlights the limitations of Chapter 2. The authors did a good job at summarising a number of studies but little perspective is given on these results. The differences between the studies are striking, taking the extreme: Ming et al. find all 49 pathways (!) involved in ASD while Cozzolino et al. only find one (and that one is common to only four out of twelve studies). It would have been good to add a final part to the chapter to discuss how the studies overlap (or do not), and what metabolic pathways are most likely to be of practical interest (those found by all studies?).

The same comment is valid for Chapter 3 and the various lipid pathways identified in the studies described in Chapter 4.

Answer

Following the Reviewer's suggestion, I have now added a new paragraph into "Metabolome alterations in autism" section of the thesis discussing the variation in the number of autism-related pathways detected by different studies. Additionally, I added a new paragraph into the conclusion section of the thesis presenting the outcome of my work in the context of other studies.

Comment 3

On page 14, the summary of an article by Graham et al. highlights one of the main caveats of untargeted metabolomics and lipidomics studies: over- or even mis-annotation of molecular species. In that specific example, Graham et al. identified significant changes in the concentration of phosphatidylglycerophosphate (16:1ω7/18:0). However, for that study, the authors seem to have used a high-resolution single stage MS approach. Such an approach yields an accurate mass measurement, which can inform on the sum composition of a lipid (in this case 34:1) but cannot inform in any way about the on the actual fatty composition (16:1/18:0) unless MS/MS is undertaken, and even less yield information on the double bond position (16:1ω7). At best the authors should thus have identified the lipid as phosphatidylglycerophosphate (34:1). If one wants to dig deeper, one may also wonder why a lipid bearing two phosphate group, and therefore likely to yield negative [M-H]- or possibly [M-2H]2- ions was identified is positive ionisation… the same is true for the other lipid identified in the same study: CDP-diacyglycerol (18:1/18:2) that should probably be only reported as 36:3.

Of course, these issues are with the study by Graham *et al.* and not due to a mistake by the author of the present work. However, they highlight a common problem encountered when using single stage MS together with databases searches; an issue that should be mentioned somewhere in the dissertation as is has a direct effect on pathway identification.

Answer

I completely agree with the Reviewer that annotation is a critical problem in untargeted lipidomics and all predicted annotations obtained based on one-dimensional MS experiments need further validation using MS/MS fragmentation. While these types of experiments are outside of the scope of my thesis, I added the discussion of this point in the conclusion section, where I address annotation issues.

Comment 4

Another example of the above issue, directly relevant to this dissertation, is shown in Figure 5.4D. Here the author has chosen to show the concentration of monogalactosyldiacylglycerol (MGDG). A few questions come to mind:

- MGDG is an entire class of lipid, the reference given (LMGL05010014) corresponds to MGDG (18:3/18:4). Can one assume that this is the molecular species that was identified?
- If so, how was the fatty composition elucidated? If only single stage MS was conducted the best annotation for that lipid should be MGDG (36:7). Defined fatty acyl moieties can only be described if MS/MS was done and the respective fragments found.
- Assuming the lipid is indeed MGDG (18:3/18:4), it seems a little odd. MGDG is a plant lipid, I don’t think it has ever been described in mammalian kidney. In addition, 18:3 and 18:4 fatty acids might well be present in kidney but their concentration would be very low (traces level for 18:4).
- To support the identification of this lipid, the author could provide MS info. For example: what ions were detected, was the identification validated in both positive and negative ionisation, is there any MS/MS data?
- In addition, why choose to highlight that specific lipid in the figure while it is not discussed at all in the text?

Answer

As it was mentioned in the previous comment, there are several issues related to the compound annotation step. I have now added discussion of these issues to the thesis. Below please see the answers to a particular question, related to this comment:

- There were multiple molecular species that matched to the above-mentioned lipid peak: MGDG (18:3/18:4); PG(16:0/20:3); PG(14:1/22:2); PG(16:1/20:2) and other PG(36:3), all falling into 10 ppm range. MGDG (18:3/18:4) had the lowest ppm difference, that’s why it was selected. In LIPID MAPS there is only one lipid species with that number of carbon chain (36) and a number of double bonds (7) - MGDG (18:3/18:4). That is the reason we annotated the compound as MGDG (18:3/18:4). Yet, you are completely right to point out that based on our experiment we cannot reliably say which configuration of MGDG (36:7) is correct.
- This particular lipid species was identified only in the negative ionization mode in dataset1 and both positive and negative mode in dataset2. In this

work, there was no any MS/MS data.

- LMGL05010014 was chosen as an example of the lipid species that demonstrates human-specific concentrations differences (decreased in humans compared to the other species) in the kidney in both datasets. We are not discussing this particular molecular species, as we instead concentrated on lipid classes, given the annotation step problem for the data based on one-dimensional MS experiment.

Comment 5

In relation to the above comment, a better description of how lipids were annotated (included maybe the use of internal standards comparison and retention time information) could be added in paragraph 4.1.4. Maybe supplementary data could be made accessible, including the list of identified species (for example what species were used for generating Figure 4.4).

Answer

Following the Reviewer's suggestion, I have now included a more detailed description of the annotation step into the thesis. Specifically, we used a similar annotation approach that is described in the thesis for the metabolite part. At the first step, we used Progenesis QI software to perform peak peaking, next the software automatically detects the different isotope peaks, clustered them together, and reported summed intensity as the monoisotopic mass retention time feature. The resulting features were next searched against LIPID MAPS with a mass tolerance of 10 ppm, allowing [M+H], [M+NH4], [M+Na], [M-H2O+H] modifications as possible adducts in the positive ionization mode, and [M-H] and [M+oAC-H] modifications in the negative ionization mode. The annotated lipid species were not further validated by MS/MS approaches. Furthermore, because we used only one IS in these studies, we haven't used this information for lipid annotation. For the same reason, we haven't used retention time for annotation as well. In our more recent studies, we are using seven different IS, which correspond to different lipid classes, and we use those standards as anchors of the mass and retention time for annotation. For chapter 4, the list of identified lipid species, as well as normalized concentration are available as supplementary information for the original publication: https://data.mendeley.com/datasets/m4dt3z68s5/draft?a=85342504-8750-4703-88bc-83bf0c111935

Comment 6

In the field of lipidomics, normalisation is widely done using class-specific internal standards (IS). In paragraph 4.1.2, the IS 1,2-diheptadecanoyl-sn-glycero-3-phosphocholine is mentioned, but in paragraph 4.1.3, another way of normalisation is mentioned. Was the IS used at all (as it is mentioned in Paragraph 5.1.2)? If so, how? If not, why? Is the IS mentioned in 5.1.2 the same IS?

Answer

In paragraph 4.1.3 we describe normalization procedure used in our studies. We only used one IS in these studies. Consequently, as different lipid classes can behave differently, normalization based on only one IS from a particular lipid class might introduce biases. That’s why we employed a technique to account for mass spectrometry sample ordering effect based on linear regression followed by quantile normalization, without normalization using IS. In our current experiments, we are using seven different IS, representing different lipid classes. Based on these standards, we are conducting standards-based normalization.

In paragraph 5.1.2 the same IS was used, as in paragraph 4.1.2, - 1,2-diheptadecanoyl-sn-glycero-3-phosphocholine. In that study, we tested both normalizations using IS and quantile normalization, with the two procedures yielding similar results.

Comment 7

Randomisation of samples is not mentioned in paragraph 3.1.4, also it is later in paragraph 3.2.1. Add a description of randomisation steps in the method section.

Answer

Following the Reviewer’s suggestion, I have now added the following sentence to paragraph 3.1.3: Sample randomization was performed twice: before the lipid extraction and before the mass spectrometry measurements. Those factors we considered in the process of randomization: diagnosis, age, and species.

Comment 8

The following figures have already been published: 4.1, 4.3, 4.4, 5.2, 5.3, 5.4, 5.5, and 5.6. Should this be acknowledge either in the figures’ legend or in section 1.5 where it could be described which publication corresponds to which chapter.

Answer

Following the Reviewer’s suggestion, I have now added the references to the corresponding publications in figures legends, and connected publications and conference reports with the corresponding chapters.

Comment 9

This comment is somewhat aligned with comment 2 above. In Chapter 3 and in the conclusions, it is stated: “Remarkably, many of these alterations, both at the pathway level and at the level of individual metabolites, coincide with the differences reported in urine and blood of ASD patients.” This is a potentially very interesting point for potential diagnostic applications, but it is only mentioned in passing. What are these specific pathways and individual metabolites? Have they been validated? If not, that would be a perspective for future work.

Answer

Following the Reviewer’s suggestion, I have now added discussion of the potential diagnostic application of our results into the conclusion section of the thesis.

Comment 10

Minor corrections

Answer

I thank the Reviewer for this comment. I have fixed the issues you listed and additionally proofread the text.

Reviewer: Prof. Mikhail Gelfand

Comment 1

What are “enzymes directly linked to metabolites” on p. 31?

Answer

Following the Reviewer’s suggestion, I have now clarified this point. Specifically, enzymes directly linked to metabolites are enzymes, which either process or use as co-factor the following metabolites. Those interactions were directly taken from the KEGG database API.

Comment 2

In the first paragraph of page 32, when defining human-specific metabolite changes, did the authors require the change to be significant? in what sense?

Answer

We defined two sets of human-specific metabolite changes: stringent and relaxed. To identify differences using the stringent cutoff, we performed F-test for each species pair twice, using either species as a reference. If the test was significant for both human/chimpanzee and human/ macaque pairs but not significant for the chimpanzee/macaque pair, then the metabolite was classified as showing the human-specific metabolic difference. The significance was assessed based on BH-corrected p-value. To defined changes using the relaxed cutoff, we calculated distances between species using macaque metabolite concentrations as a baseline. A metabolite was classified as showing the human-specific metabolic difference if its human-macaque distance was larger than chimpanzee-macaque distance, and the direction of changes relative to the macaque coincided in humans and chimpanzees. In this case, we haven’t used any threshold of significance, just the direction of changes.

Comment 3

At the bottom of the same page, was $|\log_2$ fold change$| > 0.2$ the only condition, or was there an additional condition on statistical significance? (At low expression levels one may observe high, but insignificant fold change.) At that, the applied threshold looks very weak: are the biological conclusions robust with regards to the threshold selection?

Answer

The increase of threshold (0.3, 0.4, 0.5) lead to even higher significance of the result, but the power of the test was relatively small (low number of genes linked to significant ASD-related metabolites). It should be mentioned, that changes between ASD patients and healthy humans are relatively small in terms of magnitude levels, which can be seen in Figure 4.2B.

Comment 4

If Fig. 3.3e is based on 500-fold resampling, it might be a good idea to present the ROC curve with error bars or as a distribution.

Answer

Figure 3.2 was added for this purpose, showing the solid line corresponding to average ROC AUC with the area shows the standard deviation.

Comment 5

In the last paragraph of section 3.2, is the observed differences in the fraction of human-specific metabolites in the autism-related modules mirrored by chimpanzee-specific differences, or are some modules enriched in human-specific differences, and other modules, in chimp-specific ones?

Answer

In Figure 3.5b, we plotted the ratio of human-specific and chimpanzee-specific metabolites represented in different categories. That's why some modules demonstrate human-specific behavior (module 2-4), while module 1 is enriched with chimpanzee-specific differences. Also, it should be mentioned that all modules contain metabolites with human-specific and chimpanzee-specific differences, but module 1 contained fewer human-specific metabolic differences compared to the average, while module 4 contained approximately five times more.

Comment 6

In the discussion (section 3.3) it would be instructive to concentrate not on similarities, but on differences between the brain and the blood and urine metabolic changes in ASD, as the latter (unlike the former) may serve as diagnostic markers.

Answer

Following the Reviewer's suggestion, I have now added discussion of the potential diagnostic application of our results into the conclusion section of the thesis.

Comment 7

In section 4.1.5 it is not explained what clustering algorithm has been applied, is it the complete linkage as in chapter 3? Further, why only autism has been analyzed using linear regression?

Answer

Following the Reviewer's suggestion, I have now added the description of the clustering algorithm, which was used: complete linkage. For autism analysis, we used two approaches: Wilcoxon rank sum test and linear regression. It was used only for autism data because the second dataset contained only ASD patients and healthy controls and we wanted to check the robustness of the procedures like it was done in Chapter 3.

Comment 8

Finally, the conclusions chapter would be much more interesting if the author had not just listed the findings of three studies but attempted to integrate them. In particular, precursors of lipids are metabolites: are there any concerted differences in the metabolite and lipid concentrations in various conditions or between species? (Indirectly same links may be observed via transcriptome analysis.) Given that the metabolite study identified some correlation between metabolite differences in autism and between primates, were similar correlations observed for lipids (at that, an integrative analysis of the results of chapters 4 and 5 would be instructive).

Answer

Following the Reviewer's suggestion, I have now made an attempt to combine the results of metabolome and lipidome part. Only a small fraction of metabolites, however, serves as building blocks for lipids. Those building blocks include Acetyl-CoA, which is elongated with Malonyl-CoA. Unfortunately, those metabolites were not detected in our metabolome study. I have now rewritten the conclusion part of the thesis, in which I integrated the lipidome parts and specifically highlighted the finding that glycerophospholipid metabolism was altered in ASD and demonstrated the human-specific behavior. Additionally, I added a paragraph describing a perspective for future work.

Comment 9

"Oligogenic model claims that ASD is caused by a relatively small number of genetic variants, each having a large risk of ASD development. Major gene model claims that ASD development can be caused by genetic variants, each having a large risk" – what is the difference?

Answer

I thank the Reviewer for this comment. Major gene models claim that there is one highly penetrant rare mutation or a limited number of mutations. Sometimes, the last part of this definition (limited number of mutations) is used as a separate model - oligogenic model. To avoid potential misunderstanding, I discarded the oligogenic model from the text.

Comment 10

What dictates the choice of ADNP and ANK2 for specific discussion at the end of section 2.2?

Answer

Those genes were used to demonstrate, that mutation in those genes can cause ASD, but at the same time, they were associated with multiple human disorders.

Comment 11

The selection of discussed metabolites in section 2.4.1 looks completely spurious (glucose and ATP are important, but why specifically these two?).

Answer

I thank the Reviewer for this comment. Glucose and ATP were used as an example of primary metabolites, which are the main energy sources in the cell, but at the same time belonging to two completely different class of molecules: glucose belongs to carbohydrates, which are usually involved in cell energy regulation and ATP belongs to purines, which have a more diverse set of function: building blocks of DNA and RNA, neurotransmitters, extracellular communication.

Comment 12
Metabolome alterations in sections 2.4.2and 2.4.5 would look much better in a tabular form instead of lists in the text or, even better, as Venn diagrams: that would allow a reader to assess the consistency of findings.
Answer
I thank the Reviewer for this comment. Following the suggestion, I have now made new Figure 2.2 and added a new paragraph into “Metabolome alterations in autism” section of the thesis discussing the variation in the number of autism-related pathways detected by different studies.

Comment 13
In 2.5.4 it is not clear whether the observed changes are due to schizophrenia or to the fact that strong drugs have been taken, naturally leading to changes in the lipid content. At that, one would expect a review not merely to repeat the findings but to have a critical component as well.
Answer
I thank the Reviewer for this comment and want to clarify this point. In this work authors compared schizophrenic patients with healthy controls and demonstrated that PE lipid class is altered in disease. Next, they compared drug-treated schizophrenia patients with non-treated patients. Different drugs we used in that experiment, and it was shown that different drugs affect the lipid concentrations in different ways. This type of experiment is particularly interesting, as using this approach an optimal drug combination could potentially be found for a particular patient, i.e. drugs that change only disease altered lipid classes and don’t have off-target effects.

Comment 14
The English style and spelling need to be improved and misprints need to be corrected
Answer
I thank the Reviewer for this comment. I have fixed the issues you listed and additionally proofread the text.

Reviewer: Prof. Georgii Bazykin

Comment 1
The literature review covers much of the relevant literature. It would be interesting to also consider here the genetic determination of the metabolites/lipids level, e.g. in enzymes catalyzing reactions associated with them. Such results are mentioned in the text (e.g. p. 38); is there no prior literature on this?
Answer
I fully agree with the Reviewer that it is interesting to investigate the genetic

determinants of metabolites/lipids levels, but this work lays outside of the scope of this thesis. In the thesis, I briefly mention the following observations relevant to the genetic determinant of metabolites/lipids levels:

1) There are several genetic alterations in ASD, which are found in glutathione metabolism. The SNPs are found in enzymes gamma-glutamyltransferase (*GGT1),* glutathione synthetase (*GSS),* glutathione peroxidase (*GPX1),* glutathione S-transferase mu (*GSTM1),* glutathione S-transferase alpha *(GSTA2),* which either produce or catabolize glutathione, the metabolite that is decreased in ASD.
2) Also I described that the alteration of pyrimidine pathway genes, such as uridine monophosphate synthase (*UMPS*), dihydropyrimidine dehydrogenase (*DPYD*), Dihydropyrimidinase (*DPYS*), beta-ureidopropionase 1 (*UPB1*) leads to neulogical aberration and based on our metabolomics data pyrimidine pathway is altered in autism as well. That's why there is a possible link between alterations in genes and metabolite concentration.
3) Additionally, I described a few studies, in which authors identified loci affecting serum levels of TC, LDL-C, HDL-C, and TG.

Currently there are only a few studies, in which up to thousand people were genotyped and concentrations of metabolites or lipids were measured in the same individuals, in order to identify connection between genetic variants and metabolites or lipids in different populations (10.1038/s41467-017-01972-9; 10.1038/s41467-017-00413-x). Currently, those types of studies are limited to population studies and only small group of metabolites or lipids is investigated. To the best of my knowledge there is no large-scale analysis of metabolic QTL in cognitive diseases.

Comment 2

The finding that only a tiny fraction of lipids evolves in a clock-wise fashion is very interesting; however, I do not see the ground to reject immediately the "neutral" explanation for this evolution as done by the author. Perhaps the rest of the lipidome is functional, and much of its divergence is also functionally relevant, and therefore strong. Even for genomic data, there are cases out there when just a very small fraction of the genome evolves neutrally, with the rest being functional (e.g., in Drosophila, just short sequences at the middles of introns are considered functional).

Answer

I thank the Reviewer for this comment. I completely agree with the point that the majority of the lipidome might be functional. This idea is mentioned in the thesis in Figure 5.6 B. In addition, 95% of lipids show significant age-dependent concentration differences (10.1038/s41380-018-0200-8). This observation further supports the functionality of the large proportion of the lipidome.

Comment 3

At the bottom of p. 44, I am confused. Are the overrepresented changes in module

4 in the same direction in ASD as in human as a species? This looks interesting, but unclear.

Answer

I performed the analysis of ASD direction changes and analyzed it in the context of human comparison to non-human primates. Specifically, in the case of module 4, the concentration differences observed in ASD coincides with the direction of the difference between control humans comparing to non-human primates, but the difference magnitude is lower.

Reviewer: Prof. Andrey Mironov

Comment 1

Here, however, it should be noted that in the analysis of lipidome mass spectrometry data is used, which are qualitative rather than quantitative.

Answer

MS data is indeed qualitative if we speak about absolute concentration values. In our work, however, we always analyze relative lipid levels across samples or groups: such as affected samples vs. controls, etc. In such analyses, MS data differences are quantitative and could be compared across studies or sample group comparisons within one study.

Comment 2

Moreover, the work used data from different sources. This means that artifacts associated with the characteristics of laboratories are possible. It would be useful to compare data from different laboratories obtained for the same types of biomaterials.

Answer

All studies described in the thesis represent an analysis of the data obtained in the same laboratory: all data was produced by the laboratory of Patrick Giavalisco in Max Planck Institute for Molecular Plant Physiology using the same experimental protocol. Similar data produced by other laboratories for brain samples is not publicly available and possibly does not exist. Yet, we made comparisons between datasets generated independently, such as two control-ASD datasets, yielding a comparable set of pathways affected by ASD in both datasets.