Skolkovo Institute of Science and Technology

as a manuscript

Anna Tkachev

ANALYTICAL PIPELINE FOR LC-MS-BASED GLOBAL LIPIDOMICS DATA ANALYSIS AND ITS APPLICATION TO THE INVESTIGATION OF METABOLIC SIGNATURES OF PSYCHIATRIC DISORDERS

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow - 2022

The PhD dissertation was prepared at
Skolkovo Institute of Science and Technology.

Academic Supervisor:
Philipp Khaitovich, PhD,
Skolkovo Institute of Science and Technology,

Introduction

Significance of the Work

Technical advances have been indispensable to progress in biological sciences. Prior to the developments in mass spectrometry methods, detailed molecular compositions could not be quantified. In particular, these developments have pushed lipids – essential building blocks of all living cells and key players in energy metabolism – out of the obscurity they have been placed in previously. When genetics have not provided a comprehensive understanding of particular diseases, the study of these complex traits from a novel perspective, such as the lipid phenotype, has the potential of valuable insights.

Mental illnesses, debilitating conditions that affect a substantial proportion of the population, are examples of such traits that are known to have a strong genetic component, for which, however, no satisfactory genetic model has been discovered. Indeed, while it is well accepted that mental disorders are inseparably tied to their physical manifestations in the brain and body of the suffering individual, many questions remain. We do not know what causes mental illness, or the specific mechanisms by which drugs can help improve symptoms. Mental illnesses are separated into distinct disorders, such as schizophrenia, depression, and bipolar disorder, but both the clear-cut separation between the disorders and the homogeneity within the disorders are questionable. Since mental illnesses are currently characterized on the basis of particular behavioral symptoms, more objective physical metrics of disease are required to advance mental health research.

Lipid profiles in both the brain and blood are relevant in the context of mental disorders. There is a growing awareness of the important role lipids play in the brain, a particularly lipid-rich organ. At the same time, mental illnesses are not just disorders of the brain – they are intrinsically connected to physical health, exemplified by the high prevalence of cardiometabolic comorbidities among psychiatric patients. The blood plasma contains important indicators of metabolic health, of which blood lipids seem to be essential players. Due to the relative novelty of the field and associated technological complexity, studies concerned with lipids in the brain and blood of psychiatric patients remain scarce. In this thesis, I present results on lipid profiles in the blood of psychiatric patients in relation to their disease phenotype and medication response, as well as medication effect on the lipid composition of the brain of non-depressed macaques, modeling medication therapeutic and side effects in humans.

While rapid progress is being made in the field of lipid research, its growth has only recently begun. With developments of experimental methods, the wealth and complexity of the information output increases. Within different analytical approaches, global lipidomics analysis is associated with higher data complexity,

since it is aimed at quantifying the lipid compounds in biological systems in as much of a broad and unbiased manner as practicably achievable. The untargeted nature of the analysis is simultaneously its strength and the source of potential drawbacks. While it has the potential of being extremely informative, computational methods and data processing standards are lagging behind the technological advancements. Because corresponding data analysis methods are not yet well-established, they present an additional challenge to the lipidomics analysis. Accordingly, a considerable part of this thesis is concerned with data processing specific to global lipidomics. Developments of new lipidomics methods, including data processing tools, will facilitate lipidomics analysis and expand its application outside of its niche field, resulting in a better understanding of the role lipids play not only in mental illness, but in health and disease, in general.

Project Objectives

In this thesis, I have aimed to develop a pipeline for global lipidomics data analysis, as well as apply global lipidomics analysis methods to several specific studies of lipid profiles in psychiatric disorders. The particular questions addressed in this work were:

- Methodological aspects of a global lipidomics data analysis pipeline development and optimization.
- The investigation of lipid profiles in the blood plasma of psychiatric patients diagnosed with schizophrenia, depression, and bipolar disorder.
- The analysis of association between individual changes in blood lipid profiles and symptom severity in individuals with schizophrenia before and after hospital treatment.
- Long-term fluoxetine treatment-induced effects on the molecular composition of the brain of juvenile macaques, including alterations in lipid abundances, in the context of therapeutic and side effects of antidepressants reported in children and young adults.

Main Results Summary

As a result of the studies included in this thesis, the following propositions are put forward:

- The main processing steps of a global lipidomics data processing pipeline were described, followed by the discussion of gaps existing in the current analysis protocols. The computational approaches developed for this thesis were essential for the results of the lipidomics studies included in this thesis.
- Signal duplication in untargeted lipidomics experiments remains one of the most overlooked steps in existing untargeted lipidomics data analysis

pipelines. The data-driven approach proposed in this thesis for the reduction of signal duplication removed more than twice as many redundant features compared to the standard adduct filtration approach.
- Patients with schizophrenia exhibit robust alterations in levels of individual lipid compounds compared to healthy controls.
- These changes are common for other psychiatric disorders, such as bipolar and depressive disorders.
- Patients with schizophrenia exhibit a highly specific lipid profile distinguishing them from healthy controls, as demonstrated by the high predictive power of the proposed classification model.
- Poor treatment response in patients with schizophrenia is associated with changes in particular lipid compound levels in the blood of these individuals.
- Lipids are altered in the brain of juvenile macaques as a result of fluoxetine treatment. The lipid profile is related to PUFA reduction.

Scientific Novelty

- Evidence-based methods for accounting for signal duplication and unexpected adduct formation in global lipidomics, such as described in this thesis, have not been proposed before.
- The investigation of lipid profiles in blood plasma of psychiatric patients was the most comprehensive, to date, in terms of cohort and sample sizes, lipidomics coverage, and transdiagnostic comparison.
- Reported markers of treatment response in schizophrenia are scarce, and the triglyceride lipid profile described in this thesis was not reported in relation to treatment response in schizophrenia.
- The investigation of effects of fluoxetine on lipid profiles in juvenile macaque brains was the first study of the effect of long-term fluoxetine treatment on the molecular composition of primate brains.

Practical Implications

The practical implications of the obtained results is related to the need of more objective physical metrics of disease for the advancement of mental health research.

The reported lipid-based diagnostic model for schizophrenia has been validated in several independent test datasets, demonstrating its potential applicability in practice. While such a diagnostic model for the diagnosis of schizophrenia has limited practical use as-is, it might possibly be expanded to more nuanced conditions, such as less severe mental illnesses or the prediction of future disease development.

The investigation of responses to psychopharmacology is a highly relevant issue due to currently unpredictable treatment outcomes. By investigating associations between changes in individual lipid profiles and symptom improvement after

treatment, a particular lipid profile associated with poor medication response has been described. Further study of treatment response markers can potentially improve not only the understanding of medication effects, but enable more personalized therapy selection in the future.

The analysis of the effect of the antidepressant fluoxetine on juvenile macaque brains was related to a lipid signature of reduced polyunsaturated fatty acids (PUFAs). These lipid components are presumed to be important both for mental health and proper brain development in children. While the results based on model organisms cannot be directly generalized to humans, they call attention to the importance of lipids and PUFAs in depression research.

Finally, the global lipidomics data analysis is a challenge in itself due to limited literature resources. This can be explained by the novelty of the field, and its interdisciplinary nature at the intersection of analytical chemistry and data analysis. The chapter on data processing for global lipidomics presented in this thesis can be a useful practical guide, as well as a pointer for further investigations and developments.

Personal Contributions

All of the statistical analysis based on lipid data was performed by the author. All of the lipidomics data processing, excluding some particular software steps, were performed by the author.

Publications Related to this Thesis

1. Tkachev A., Stekolshchikova E., Anikanov N., et al. Shorter Chain Triglycerides Are Negatively Associated with Symptom Improvement in Schizophrenia. Biomolecules. 2021;11(5):720. Published 2021 May 11. doi:10.3390/biom11050720

2. Tkachev A., Stekolshchikova E., Bobrovskiy D.M., et al. Long-Term Fluoxetine Administration Causes Substantial Lipidome Alteration of the Juvenile Macaque Brain. Int J Mol Sci. 2021;22(15):8089. Published 2021 Jul 28. doi:10.3390/ijms22158089

Other Publications by the Author

Publications by the author that are not included in this thesis:

3. Tkachev A, Stepanova V, Zhang L, et al. Differences in lipidome and metabolome organization of prefrontal cortex among human populations. Sci Rep. 2019;9(1):18348. Published 2019 Dec 4. doi:10.1038/s41598-019-53762-6

4. Kurochkin I, Khrameeva E, Tkachev A, et al. Metabolome signature of autism in the human prefrontal cortex. Commun Biol. 2019;2:234. Published 2019 Jun 21. doi:10.1038/s42003-019-0485-4

5. Khrameeva E, Kurochkin I, Han D, et al. Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. Genome Res. 2020;30(5):776-789. doi:10.1101/gr.256958.119

6. Stepanova V, Moczulska KE, Vacano GN, et al. Reduced purine biosynthesis in humans after their divergence from Neandertals. Elife. 2021;10:e58741. Published 2021 May 4. doi:10.7554/eLife.58741

Thesis Structure

The **first chapter** of the thesis introduces the context of the work, and postulates the project objectives.

Chapter 2 contains background information and is divided into three main sections. Section 2.1 provides general information concerning psychiatric disorders. It provides an overview of three psychiatric disorders, schizophrenia, bipolar disorders, and major depressive disorders. It discusses current approaches to diagnosis, available biomarkers, similarities and differences between disorders, as well as somatic comorbidities. A brief overview of psychopharmacology is also presented. Section 2.2 introduces the topic of lipids and lipidomics. A short but comprehensive introduction to lipid nomenclature is provided. Evidence from lipidomics studies in different mammalian tissues is discussed and a general overview of lipid metabolism is provided, with the goal of facilitating the understanding of lipid research implications for readers not familiar with the topic. An overview of lipidomics studies in the context of psychiatric disorders is provided, as well, for blood plasma and brain, in particular. Section 2.3 provides general information on experimental techniques for lipidomics, in particular, liquid chromatography mass spectrometry (LC-MS). It includes a critical review of the available literature concerning data processing of LC-MS-based lipidomics.

Chapters 3, 4 and 5 contain the main results for the studies of lipid profiles related to psychiatric disorders. Each chapter contains a discussion of the presented results, as well as a methods section describing the experimental details and statistical analysis.

Chapter 6 is concerned with data processing for untargeted LC-MS-based lipidomics. In section 6.1, I describe the main processing steps in global lipidomics data processing pipeline, including gaps existing in the current analysis protocols. In particular, I discuss removing features with poor biological signal, high technical variability, correction for the influence of experimental confounding factors, and alignment of different datasets for the comparison of results from different experiment. In section 6.2, I propose a data-driven approach for dealing with a particular issue that remains poorly resolved in typical untargeted lipidomics data

analysis workflow: the removal of extensive signal duplications in untargeted LC-MS lipidomics.

Chapter 7 comprises thesis conclusions.

Main results

Cross-cohort lipidomics analyses in individuals with severe mental disorders

Summary

There is fragmentary evidence that common psychiatric disorders, schizophrenia (SCZ), bipolar disorder (BPD), and major depressive disorder (MDD), substantially alter brain lipidome composition, as well as the lipid composition of blood plasma. Here, an extensive analysis of blood plasma lipid alterations in these three disorders is presented, involving 1,361 lipid features measured in 1,354 blood plasma samples collected from unique individuals as separate cohorts in China, Western Europe (Germany and Austria), and Russia. A signature of 77 lipid intensity alterations shared by all three cohorts was identified (SCZ-associated lipids). This signature was also present in BPD and MDD patients, aligning with the reported symptomatic and genetic overlap among the three disorders. Moreover, a predictive model based on lipid intensities was proposed, separating SCZ patients from controls (CTL) with high diagnostic ability (Area under the Receiver Operating Characteristic Curve, ROC AUC=0.86-0.95), and further validated on two different independent datasets.

Study setup

To assess whether blood lipidome composition is altered in common psychiatric disorders, schizophrenia (SCZ), major depressive disorder (MDD), and bipolar disorder (BPD), in a robust and reproducible manner, we conducted a study involving 876 patients and 478 control (CTL) individuals from three cohorts sampled at different geographic locations: Chongqing, China (CN), several locations in Germany and Austria (DE-AT), and Moscow, Russia (RU) (CN: n = 170, 222, 153 for SCZ, MDD, and CTL, respectively; DE-AT: n = 184, 148, 187 for SCZ, BPD, and CTL, respectively; RU: n = 82, 36, 34, 138 for SCZ, BPD, MDD, and CTL, respectively). Additionally, we collected a sample group consisting of 104 first psychotic episode patients (FEP) at a single location (fepRU dataset). In total, we assessed lipid abundances in 1,552 plasma samples.

To obtain systematic coverage of the blood plasma lipidome, we measured all samples using liquid chromatography coupled with untargeted mass spectrometry (LC-MS) in both negative and positive ionization modes. This approach yielded 1,361 reproducibly detected lipid features. Of them, we annotated 395 lipid compounds based on their retention time and mass-to-charge properties, including, for a subset

of the compounds, an examination of LC-MS^2 fragmentation patterns. Annotated lipids covered 16 lipid classes, aligning well with previously reported blood plasma lipidome components (Figure 1).

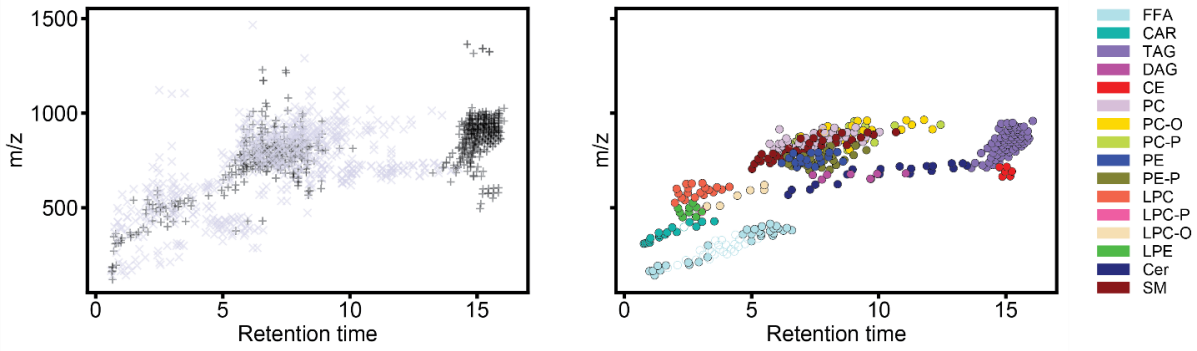

Figure 1: Detected features and annotated compounds. Left: The retention time and mass-to-charge values for all reproducibly quantified 1,361 lipid features in positive (black, "plus" markers) and negative (grey, "x" markers) ionization modes. Right: same for computationally annotated lipids. Empty circles indicate lipids discarded from downstream analysis as food-intake related. Colors correspond to lipid classes, denoted on the right.

Blood lipidome alterations associated with schizophrenia

Comparison of lipid intensities between a total of 436 SCZ and 478 CTL samples, after exclusion of effects potentially attributable to confounding variables such as age, sex, fasting period prior to blood donation, body mass index (BMI), co-recorded with sample information, revealed significant differences for 38-61% of detected features for each of the three cohorts (permutation $p < 0.001$, Wilcoxon rank-sum test Benjamini-Hochberg corrected FDR = 10%). Among them, 213 features showed significant intensity differences independently in all three sample cohorts – CN, DE-AT, and RU (SCZ-associated features; permutation test, $p < 0.00001$; Figure 2A). Further, the direction of the intensity changes in SCZ correlated positively and significantly between all three cohort pairs (Pearson correlation, $r \geq 0.75$, $p \leq 0.00001$; Figure 2B).

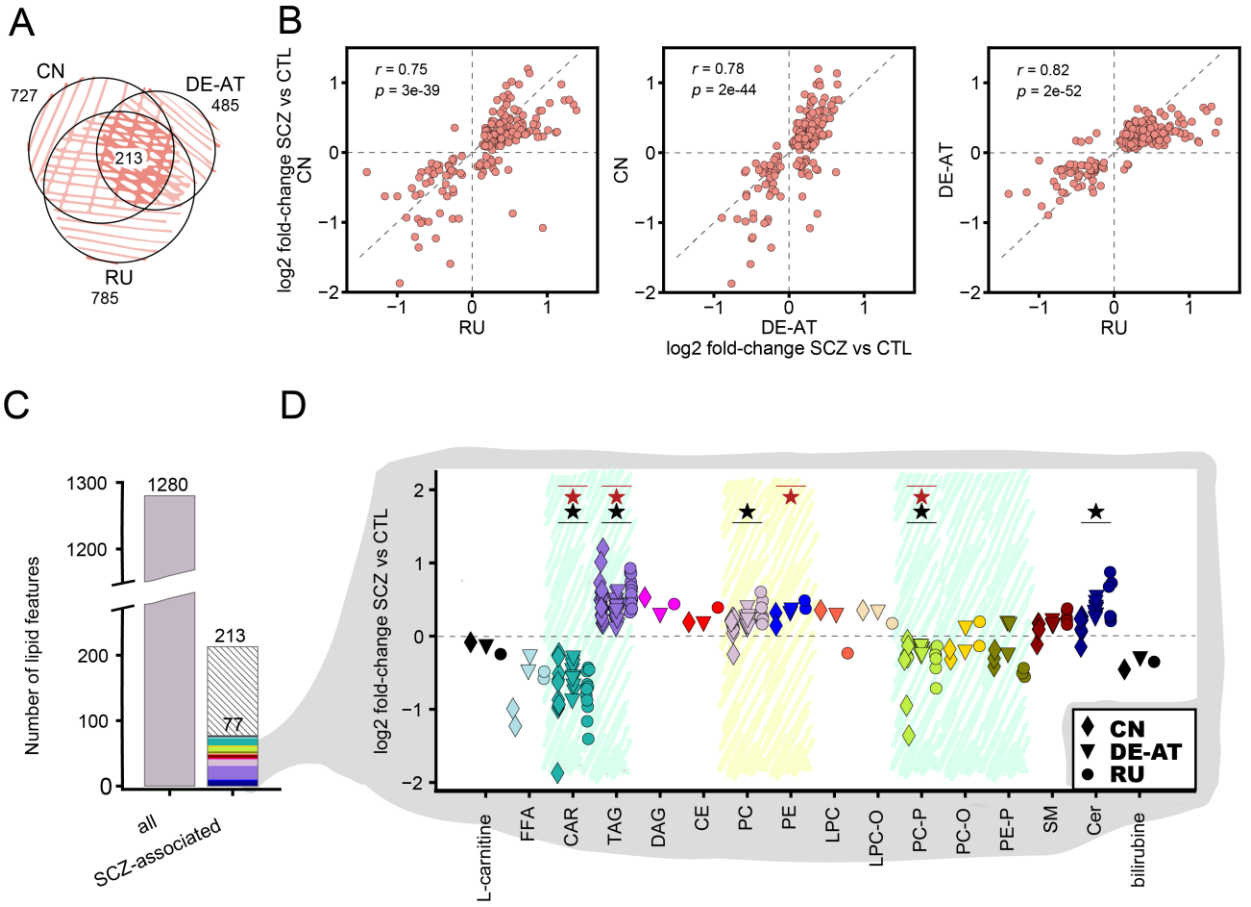

Figure 2: SCZ-associated lipidome alterations. (A) Number of lipid features showing statistical differences in each of the sample cohorts – CN, DE-AT, and RU – and their intersection (SCZ-associated lipid features). (B) Pairwise comparisons of the lipid intensity differences between schizophrenia (SCZ) and control (CTL) samples between the cohorts. The intensity differences are displayed as the base two log-transformed fold change (log2 fold-change). Circles represent the 213 SCZ-associated lipid features. Pearson correlation coefficients and corresponding p-values are marked in the top left corner. The horizontal and vertical dashed lines indicate log2 fold-change = 0, the diagonal dashed line indicates the y-axis = x-axis line. (C) The numbers of all analyzed confounder free lipid features (left), SCZ-associated lipid features (right, hashed), and annotated SCZ-associated compounds (right, colored). The colors correspond to the compound classes displayed in panel D. (D) Intensity differences between SCZ and CTL individuals of the 77 SCZ-associated compounds sorted according to their lipid classes (marked by color and lipid class labels at the bottom) and shown separately for each cohort (marked by symbol shapes). Stars on the top denote lipid class level significance of the difference, described in subsection 3.1.2 of the thesis.

The 213 SCZ-associated lipid features included 77 unique lipids validated by their LC-MS^2 fragmentation patterns covering 14 large lipid classes (SCZ-associated lipids; Figure 2C; Table 1). Analysis of these 77 lipids revealed a systemic effect of SCZ, leading to a coordinated shift of the lipids within a class towards either lower or greater intensity in all three cohorts (Figure 2C).

Lipid class	Lipid species	Molecular subspecies, when available
TAG	TAG(47:1) TAG(47:2) TAG(48:1) TAG(48:3) TAG(49:0) TAG(49:1) TAG(49:2) TAG(49:3) TAG(50:1) TAG(50:2) TAG(50:3) TAG(51:1) TAG(51:2) TAG(51:3) TAG(52:1) TAG(52:2) TAG(54:2) TAG(54:3) TAG(55:6) TAG(56:5) TAG(56:6)	
CAR	CAR(10:0) CAR(10:1) CAR(11:1) CAR(12:0) CAR(12:2) CAR(13:0) CAR(13:1) CAR(14:0) CAR(14:2) CAR(16:2) CAR(18:2)	
FFA	FFA(12:2) FFA(13:1)	
CE	CE(22:5)	
DAG	DAG(36:2)	DAG(18:1_18:1) DAG(18:0_18:2)
PC	PC(33:1) PC(34:1) PC(36:1) PC(37:3) PC(38:3) PC(38:4) PC(38:5) PC(40:4) PC(40:5) PC(42:5)	PC(18:1_15:0) PC(16:1_17:0) PC(16:0_17:1) PC(16:0_18:1) PC(18:0_18:1) PC(17:0_20:3), PC(19:1_18:2) PC(18:0_20:3) PC(16:0_22:4) PC(18:0_20:4), PC(18:1_20:3) PC(16:0_22:5) PC(18:0_22:4) PC(20:0_20:4) PC(18:0_22:5) PC(18:0_24:5)
PE	PE(36:4) PE(40:5)	PE(16:0_20:4) PE(18:0_22:5)
PC-O, PC-P	PC(O-34:0) PC(O-38:6) PC(P-34:2) PC(P-34:3) PC(P-35:3) PC(P-38:4) PC(P-38:5) PC(P-38:6) PC(P-40:7)	PC(O-18:0/16:0) PC(O-16:0/22:6) PC(P-16:0/18:2) PC(P-16:1/18:2) PC(P-35:3)[unresolved fatty acid composition] PC(P-18:0/20:4) PC(P-16:0/22:5) PC(P-16:0/22:6) PC(P-18:1/22:6)
PE-P	PE(P-36:4) PE(P-38:5) PE(P-38:6)	PE(P-16:0/20:4) PE(P-18:1/20:4) PE(P-16:0/22:6)
LPC	LPC(22:4)	
LPC-O	LPC(O-26:1)	
Cer	Cer(d34:1) Cer(d34:2) Cer(d36:1) Cer(d36:2) Cer(d40:2) Cer(d40:3) Cer(d42:2) Cer(d42:3)	Cer(d18:1/16:0) Cer(d18:2/16:0) Cer(d18:1/18:0) Cer(d18:2/18:0) Cer(d18:2/22:0) Cer(d40:3)[unresolved fatty acid composition] Cer(d18:1/24:1) Cer(d18:2/24:1)
SM	SM(d35:2) SM(d36:1) SM(d36:2) SM(d37:2) SM(d38:2)	
other	bilirubine, L-carnitine	

Table 1: The 77 SCZ-associated lipids, including fatty acid composition, when available.

Predictive classification and inter-cohort reproducibility assessment

While most of the lipid intensity differences between SCZ and CTL had moderate amplitudes, they were sufficient to separate SCZ individuals from controls with reasonable accuracy using a predictive model. Specifically, a logistic regression model trained on the three merged cohorts and 395 putatively annotated lipids separated SCZ from CTL with good diagnostic ability (Area under the Receiver Operating

Characteristic Curve (ROC AUC) = 0.95 ± 0.02 for RU test cohort, 0.88 ± 0.03 for DE-AT test cohort, and 0.86 ± 0.03 for CN test cohort). Further, the model performed just as well in classifying FEP and CTL samples from the fepRU dataset (ROC AUC = 0.89 ± 0.025). Subsection 3.1.3 of the thesis further includes analysis of cross-cohort reproducibility of lipid alterations in SCZ at the individual lipid level, as well as the level of predictive modeling.

Lipid alterations in SCZ correlate with other psychiatric disorders

We next examined lipid intensity alterations associated with MDD and BPD, each represented by two independent cohorts (CN: n = 222 for MDD; DE-AT: n = 148 for BPD; RU: n = 36, 34 for BPD, MDD, respectively). We identified 97 lipid features showing significant intensity differences in both MDD cohorts (MDD-associated lipids; Wilcoxon test, BH-corrected FDR = 10%), and 47 lipids altered in both BPD cohorts (BPD-associated lipids; Wilcoxon test, BH-corrected FDR = 10%).

Comparison of disease-associated lipids in section 3.1.4 indicated that while lipids identified as significantly associated with particular psychiatric disorders were different, the shifts in lipid abundances between disease and control were highly correlated for the three disorders (Pearson correlation, r = 0.94 and r = 0.88, p < 0.00001 for MDD-SCZ comparison in CN and RU cohorts, respectively; Pearson correlation, r = 0.66 and r = 0.93, p < 0.00001 for BPD-SCZ comparison in DE-AT and RU cohorts, respectively), indicating a shared lipid profile. Additional analysis of reproducibility of SCZ effect between cohorts (described in subsection 3.1.3 of the thesis) suggested that we could not confidently report lipid abundance differences specific to just one of the disorders. Comparison to the available literature further supported common alterations for the three disorders. Subsection 3.1.5 of the thesis included an analysis of the relationship between disease-associated genes and genes related to lipid metabolism, based on published data. The results indicated that genes associated with the blood plasma lipidome variation overlapped strongly with the genetic markers shared by the three disorders, suggesting a link between the well-described shared genetic component of the psychiatric disorders and shared lipidome alterations reported here.

Relationship between SCZ-associated lipidome alterations and medication.

Most of the SCZ patients assessed in our study received long-term antipsychotic medication treatments, shown to have effects on blood plasma lipids. To assess the extent of medication effect on our results, additional analysis was included in section 3.1.6 focusing on first psychotic episode patients and SCZ that have not received antipsychotic medication during at least six months prior to blood sampling.

Model validation follow-up study

In the results above, a lipid-based diagnostic model was defined that could distinguish SCZ from CTL with an AUC of up to 0.95. While the performance of the model was

robust across the different datasets, the applicability of a lipid-based biomarker panel in practice remains to be studied further. As the first step of addressing this question, we performed a follow-up study. A total of 119 blood plasma samples from psychiatric patients and healthy controls were collected and analyzed using the same lipidomics methods as in the original study. The sample labels, however, were concealed during the entirety of the statistical analysis to exclude model hyperparameter optimization and emulate real-world generalization capabilities of the model.

The performance of the model in the original study could be translated to an expected 0.77 sensitivity and 0.86 specificity. Accordingly, using the model trained on the lipid abundance of the 914 SCZ and CTL samples from the original studies, we predicted the disease status of the validation samples. A 0.5 threshold for the model probability was used to define SCZ and CTL prediction labels (Figure 3A).

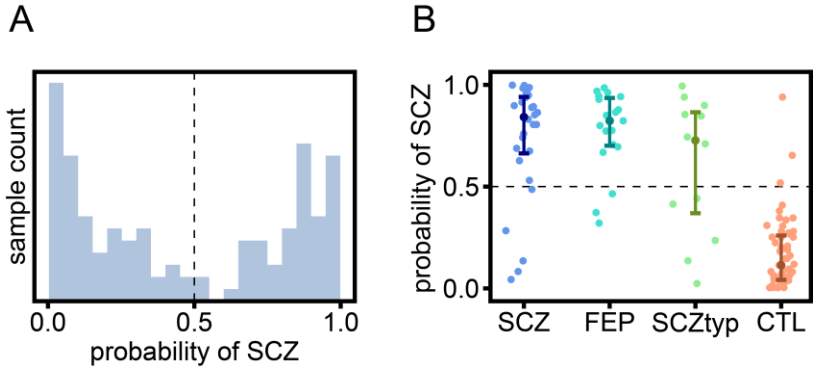

Figure 3: Prediction probabilities of SCZ for validation samples. (A) Distribution of prediction probabilities of SCZ for the 119 samples. Dashed black line corresponds to the defined threshold for label predictions. (B) Distribution of prediction probabilities of SCZ separately for SCZ, FEP, SCZtyp, and CTL samples. The four darker points correspond to median values, and whiskers extend to the inter-quartile ranges.

Of the 119 validation samples, 28 corresponded to SCZ individuals, 19 to FEP individuals, 55 to healthy CTLs, and 12 to schizotypical (SCZtyp) individuals – a personality disorder with certain similarities to SCZ, but without pronounced psychiatric symptoms. For five samples, diseases labels could not be recovered. The proportion of correctly identified SCZ and FEP patients (sensitivity) was 0.82 and 0.84, respectively, while the specificity was found to be 0.95 – higher than the performance expected from the original study (Figure 3B). Interestingly, SCZtyp individuals showed a high variability in prediction labels, as the distribution of predicted probabilities for these samples was quite different than the distributions for both SCZ/FEP and CTL (proportion of SCZtyp individuals predicted as SCZ: 0.58; Figure 3B). These results are in line with the classification of SCZtyp disorder as an intermediate schizophrenia-spectrum phenotype.

Shorter chain triglycerides are negatively associated with symptom improvement in schizophrenia.

This chapter follows the results published in [1]. Some passages have been quoted verbatim from [1], and figures were reproduced or modified with permission from [1].

Summary

Schizophrenia is a serious mental disorder requiring lifelong treatment. While medications are available that are effective in treating some patients, individual treatment response can vary, with some patients exhibiting resistance to one or multiple drugs. Currently, little is known about the causes for the difference in treatment response observed among individuals with schizophrenia, and satisfactory markers of poor response are not available for clinical practice. Here, I studied the changes in the levels of 322 blood plasma lipids between two time points assessed in 92 individuals diagnosed with schizophrenia during their inpatient treatment and its association with the extent of symptom improvement. Twenty (20) triglyceride species were found to be increased in individuals with least improvement in Positive and Negative Syndrome Scale (PANSS) scores, but not in those with largest reduction in PANSS scores. These triglyceride species were distinct from the rest of triglyceride species present in blood plasma. They contained a relatively low number of carbons in their fatty acid residues and were relatively low in abundance compared to the principal triglyceride species of blood plasma.

Study setup

We assessed the abundance of 322 lipid species in the blood plasma of 92 individuals diagnosed with schizophrenia collected at two time points: at the beginning and end of their hospitalization at a psychiatric clinic (average ± standard normal deviation: 37 ± 19 days). Samples were represented by female and male individuals (58 % female) of age ranges 17 – 43 years, and information on medication regimen was collected. Symptom severity was assessed by Positive and Negative Syndrome Scale (PANSS) score at the two time points. Lipidomics measurements were produced using mass spectrometry coupled with liquid chromatography in negative and positive ionization modes. From the reproducibly quantified lipid features, a set of 322 unique lipid compounds was annotated based on their mass-to-charge, retention time values, as well as fragmentation patterns. Assessed lipid species covered 14 lipid classes and aligned well with expected blood plasma lipidome composition.

Association between changes in lipid abundances and symptom improvement

All but one patient displayed symptom improvement from the first to the second time point, demonstrated by the reduction in PANSS scores (Figure 4A). However, the improvement in PANSS varied considerably, with top 25% best responders displaying PANSS scores improvement from -102 to -39 point differences (n = 23) and

bottom 25% worst responders displaying PANSS score improvement -14 to 0 point differences(n = 24) (Figure 4A). The extent of changes in lipid abundances differed depending on PANSS score improvement. While for worst responders, 22 lipids showed significant changes from first to the second time point (worst-response-associated lipids; Wilcoxon signed-rank test, Benjamini-Hochberg correction FDR 5%; Figure 4B), the effect in best responders was lower (Wilcoxon signed-rank test, no significant lipids at Benjamini-Hochberg FDR 5% threshold; Figure 4B). Accordingly, although the levels of worst-response-associated lipids at baseline were similar for best and worst responders, best responders did not display a statistically significant increase from first to the second point (respectively: Mann–Whitney U test, 1 of 22 $p < 0.05$; Wilcoxon signed-rank test, all 22 $p > 0.05$; Figure 4C-D). Among the 22 worst-response-associated lipids, 20 were triglycerides, 30% of the total number of triglycerides (Figure 4E). Triglycerides with lower carbon number (40-48 carbons in fatty acid residues) were most affected (Figure 4F).

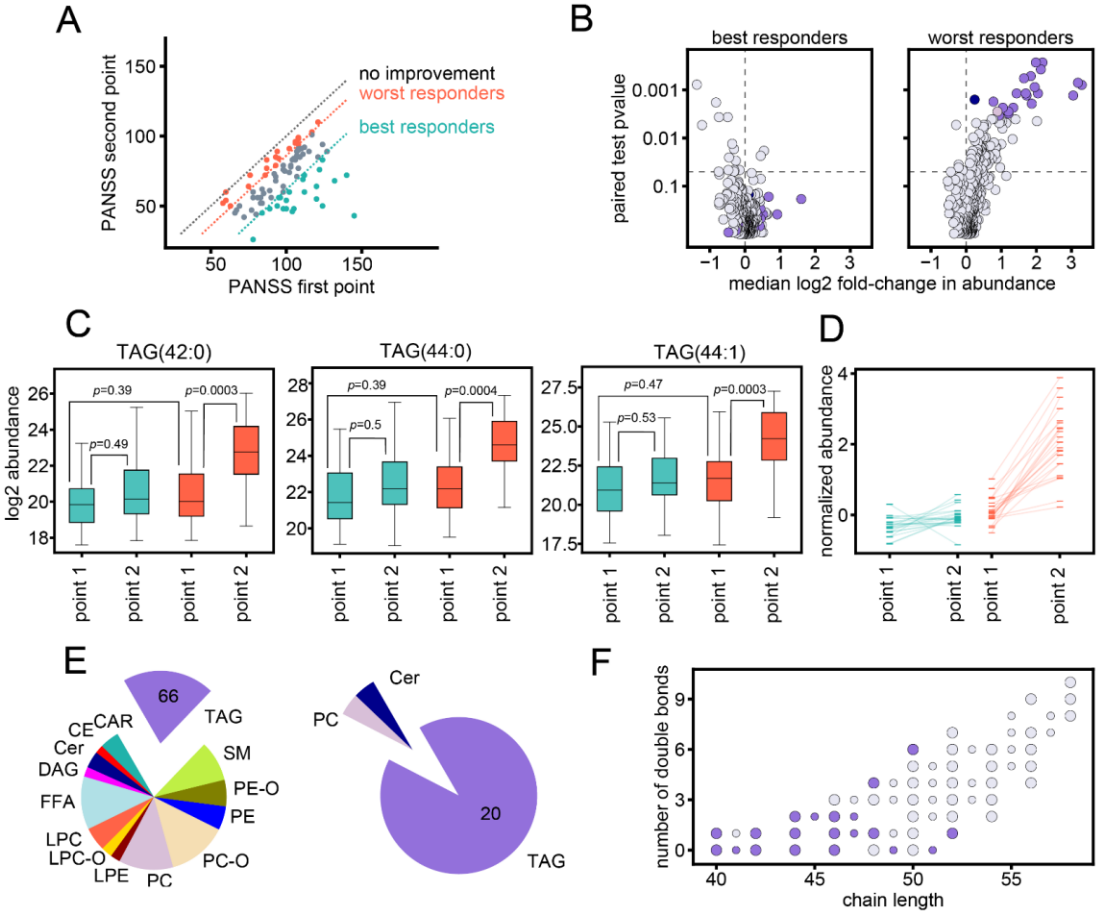

Figure 4: Significant changes in worst responders. (A) PANSS scores at first and second time points for the 92 individuals. Individuals with least improvement in PANSS score (worst responders, n = 24) and most improvement in PANSS score (best responders, n = 23) are marked in orange and green, respectively. Orange and green dashed lines demarcate the upper and lower quartile of PANSS differences, dashed grey line corresponds to identical PANSS values at the two points. (B) P-values of the Wilcoxon signed-rank test plotted against the median base two log-transformed fold change (log2 fold-change) between the two

time points for individuals with most improvement (left) and least improvement (right). The 22 worst-response-associated lipids are marked in color according to lipid class, TAG (purple), PC (dusty pink), Cer (dark blue). Dashed lines demarcate log2FC = 0 and nominal $p = 0.05$. (C) Log2 abundances for best and worst responders at first and second time point. Three worst-response associated lipids with strongest statistical effect ($p < 0.0005$ for Wilcoxon signed-tank test for changes in worst responders) are plotted: TAG 42:0, TAG 44:0, TAG 44:1. Noted p-values correspond to Wilcoxon signed-rank and Mann-Whitney U test p-values for comparisons between groups. Boxplot whiskers and fliers correspond to standard boxplot definition. (D) The median values of the normalized log2 abundances for best and worst responders at first and second time points, for all 22 worst-response associated lipids. For each lipid, the log2 abundances were normalized by the lipid mean value for all patients at the first time point. (E) Left: the 322 annotated lipids, grouped by lipid class. Right: the 22 worst-response-associated lipids, grouped by lipid class. The number of respective triglyceride species are indicated on the plot. (F) The number of carbons in the fatty acid residues (chain length) and number of double bonds for the annotated triglycerides. Worst-response-associated lipids are colored in purple. Larger and smaller circle sizes correspond to even and odd chain triglycerides, respectively.

Influence of medication and sex

Because the association between changes in lipid abundance levels and symptom severity could be confounded by medication regimens, subsection 4.1.3 contains analysis aimed to exclude the influence of medication regimen variation on reported results. Similarly, subsection 4.1.4 contains analysis aimed to exclude the effect of possible sex imbalance on reported results.

The subsection 4.1.5 of the thesis further includes the analysis of lipid profiles at first time point as predictive markers of symptom improvement after treatment, with no reported statistically significant results.

Long-Term Fluoxetine Administration Causes Substantial Lipidome Alteration of the Juvenile Macaque Brain

This chapter follows the results published in [2]. Some passages have been quoted verbatim from [2], and figures were reproduced or modified with permission from [2].

Summary

Fluoxetine is an antidepressant commonly prescribed not only to adults, but also to children for the treatment of depression, obsessive-compulsive disorder, and neurodevelopmental disorders. The adverse effects of the long-term treatment reported in some patients, especially in younger individuals, call for a detailed investigation of molecular alterations induced by fluoxetine treatment. In the presented study, we assessed residual effects of a two-year fluoxetine administration on the expression of genes, abundances of lipids and polar metabolites in the prelimbic cortex of 10 treated and 11 control macaques. Analysis of 8871 mRNA transcripts, 3608 lipids, and 1829 polar metabolites revealed substantial alterations of the brain lipid content, including significant abundance changes of 106 lipid features, accompanied by subtle changes in gene expression. Lipid alterations in the

drug-treated animals were most evident for polyunsaturated fatty acids (PUFAs). A decrease in PUFAs levels was observed in all quantified lipid classes excluding sphingolipids, which do not usually contain PUFAs, suggesting systemic changes in fatty acid metabolism.

Study setup

In this study, we assessed alterations of gene expression, polar metabolite, and lipid abundance in the prelimbic cortex (PLC; a part of the medial prefrontal cortex) of macaques treated with fluoxetine using RNA-sequencing (RNA-seq), Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR-MS), and high precision mass spectrometry coupled with liquid chromatography (LC-MS), respectively (Figure 5A). The fluoxetine and vehicle administration began at one year of age, which is equivalent to 4–6 years of age in humans, and continued uninterrupted for two years, followed by a one year post-dosing period that ended at four years of age, just before puberty. Our analysis included 10 monkeys treated with fluoxetine and 11 control monkeys administered with a vehicle.

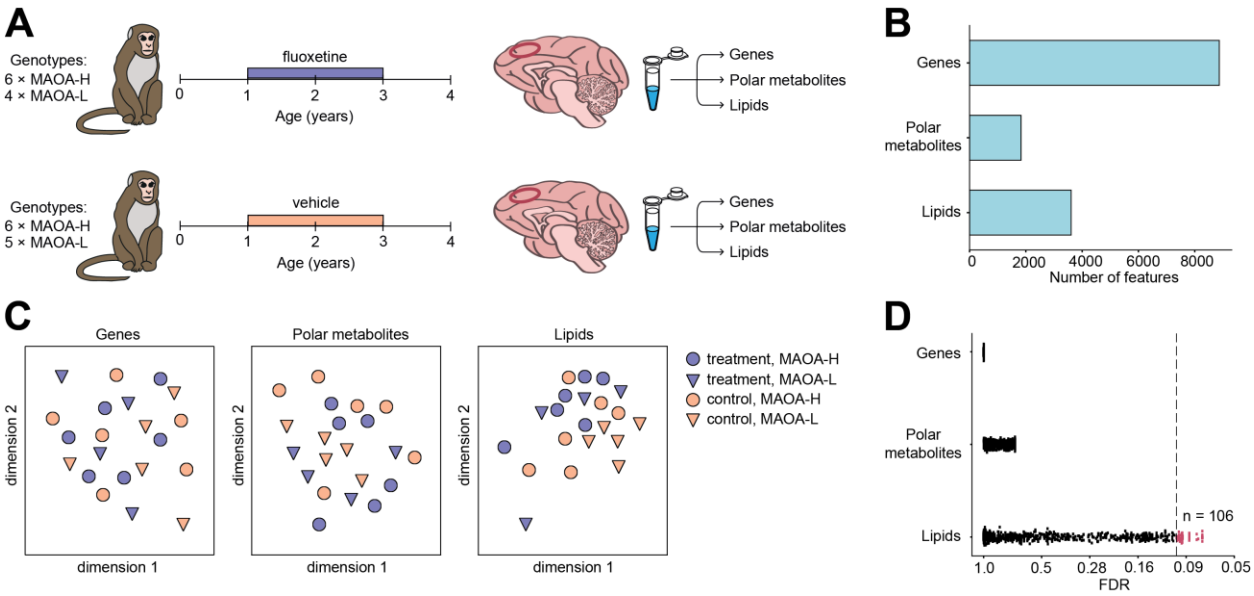

Figure 5: Experimental design and quantified molecular phenotypes. (A) Schema of experimental design. (B) Number of quantified genes, polar metabolites, and lipids. (C) Multi-dimensional scaling plots visualizing variation among samples calculated based on abundance levels of 8871 genes, 1829 polar metabolites, and 3608 lipids. Colors correspond to treatment status, shapes to the *MAOA* genotype. (D) Distribution of t-test FDR-adjusted p-values calculated for 8871 genes, 1829 polar metabolites, and 3608 lipids in the comparison between treated and control animals. Dashed line corresponds to FDR cutoff of 10%. The number of lipids passing this FDR cutoff (red) is marked on the right.

RNA-seq yielded quantitative gene expression measurements for 8871 protein-coding genes annotated in the macaque genome. Mass spectrometry analyses generated abundances for 1829 polar metabolite and 3608 lipid features, with 514 polar metabolite features and 373 lipid features putatively annotated (Figure 5B). Visualization of variation among samples based on these three data indicated the

separation of treated and control macaques at the lipid abundance level, but not at the gene and polar metabolite levels (Figure 5C).

Effect of fluoxetine on lipids

Statistical analysis revealed 106 lipid features (treatment-associated lipids) showing significant abundance differences between fluoxetine-treated and control monkeys (permutations, $p = 0.008$; Benjamini-Hochberg adjusted FDR = 10%). By contrast, the treatment effect was substantially weaker at the gene expression and polar metabolite abundance levels, and statistical effects were considered too low to define any reasonable false-discovery rate (FDR) threshold for treatment-associated genes or metabolites (Figure 7D; minimal observed q-value = 0.9986 and 0.68740 for gene expression and polar metabolite levels, respectively). We further conducted group-based analysis assessing the significance of the treatment effect at the level of lipid classes. This approach revealed significant abundance differences for five lipid classes: free fatty acids (FFA), phosphatidylethanolamines (PE), lysophosphatidylethanolamines (LPE), and hexosylceramides (HexCer;O2 and HexCer;O3) (Figures 6A; Gene Set Enrichment Analysis, adjusted $p < 0.05$). Accordingly, lipids within these classes showed coordinated treatment response, exhibited by the abundance of HexCer;O2 and HexCer;O3 lipids increasing in the PLC of the treated monkeys, and FFA, LPE, and PE lipids decreasing as a result of the treatment.

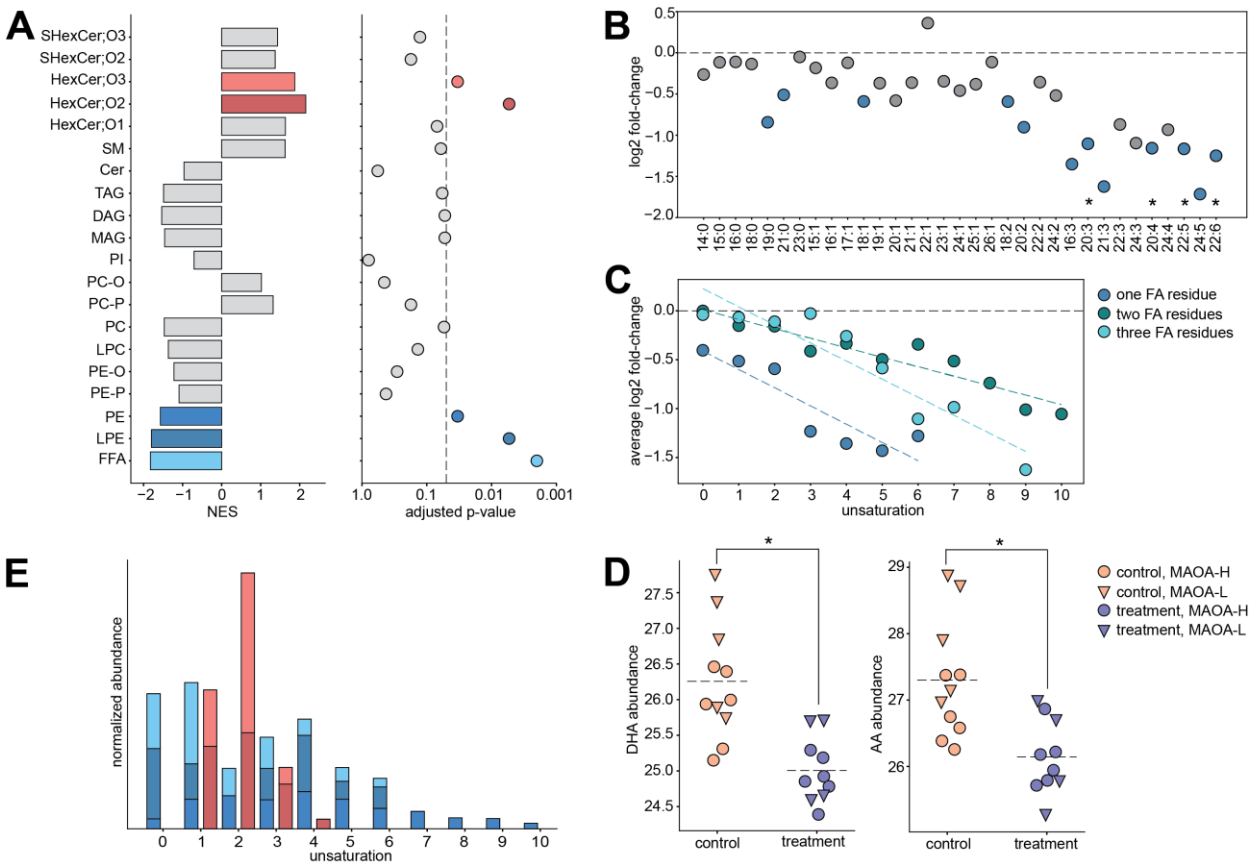

Figure 6: Effect of fluoxetine on brain lipids. (A) Lipid class enrichment analysis results based on the comparison between treated and control animals. Left: normalized enrichment scores (NES). Positive NES corresponds to an increase in the lipid abundances in treated animals, and negative NES corresponds to decrease. Colors mark lipid classes demonstrating significant enrichment in the treatment-control differences. Right: adjusted enrichment *p*-values. The dashed line corresponds to the adjusted *p*-value cutoff of 0.05. (B) Base two log-transformed fold change (log2 FC) values calculated between treated and control animals for lipids in free fatty acid (FFA) class. *X*-axis labels indicate fatty acid chain length and number of double bonds. FFAs are ordered by increasing unsaturation. Blue symbols mark compounds with nominal *t*-test $p < 0.05$. Asterisks mark statistically significant compounds (*t*-test; FDR = 10%). The dashed line indicates log2 FC = 0. (C) Log2 FC values calculated between treated and control animals, averaged for all the lipids with the same level of unsaturation (number of double bonds). Lipids were first separated into groups based on the number of fatty acid residues, one (dark blue), two (green), or three (light blue), and sphingolipids were excluded. The black dashed line indicates log2 FC = 0. Colored dashed lines indicate linear regression lines fitted to each group ($p = 0.0054$, 0.54×10^{-5} and 0.00022 for lipid classes containing one, two and three fatty acid residues, respectively). (D) Abundance levels of docosahexaenoic acid (DHA) and arachidonic acid (AA) in treated and control animals. Symbols represent individual samples. Colors correspond to treatment status, shapes to the *MAOA* genotype. Asterisks mark statistically significant differences between treatment and control (*t*-test; FDR = 10%). (E) The cumulative abundances of lipids contained in FFA, LPE, PE, HexCer;O2, and HexCer;O3 lipid classes (colors as in panel A) for each level of unsaturation (number of double bonds per compound). Cumulative abundances were normalized between the lipid classes.

The effect of fluoxetine was not uniform within a lipid class, but depended strongly on its fatty acid residue composition. Specifically, lipids containing fatty acids with multiple double bonds, or polyunsaturated fatty acids (PUFAs), were affected by the

treatment to a greater extent. This effect was most obvious for free fatty acids (Figure 6B), but also evident for the other detected glycerophospho- and glycero- lipid classes (F-test for linear regression, p = 0.0054, 0.00004, and 0.00022 for lipid classes containing one, two, and three fatty acid residues, respectively; Figure 6C). Notably, the two PUFAs constituting up to 25% of all fatty acid content in the brain, docosahexaenoic acid (DHA, FFA 22:6) and arachidonic acid (AA, FFA 20:4) differed significantly between treated and control animals at the individual compound level (Benjamini-Hochberg adjusted FDR = 10%; Figure 8D). In contrast to glycerophospholipids and glycerolipids, which contain substantial amounts of PUFA residues, lipid classes depleted in PUFAs, HexCer;O2 and HexCer;O3, increased significantly as a result of the treatment (Figure 8A,E; Gene Set Enrichment Analysis, adjusted $p < 0.05$). Subsection 5.1.3 of the thesis further includes an analysis of interaction effects between MAOA genotype and fluoxetine treatment on the levels of lipid abundances in the brains of the study animals.

Group based analysis for metabolite and gene expression data

While we did not detect significant gene expression and polar metabolite abundance differences in the PLC of fluoxetine-treated macaques, below-the-threshold effects might still be informative. Using group-based analysis designed to reveal sub-threshold effects, we identified significant treatment effects at the gene expression level, but none for polar metabolites. Specifically, 87 gene groups defined using Gene Ontology (GO) biological process terms differed in fluoxetine-treated macaques (Gene Set Enrichment Analysis, adjusted $p < 0.05$), discussed in subsection of the thesis 5.1.4.

Lipidomics Data Analysis

Section 6.1 of the thesis describes the proposed workflow for global lipidomics data analysis.

A shortened version of section 6.2 of the thesis, “Extensive Signal Duplications in LC-MS-based Lipidomics”, is presented below.

Summary

In untargeted LC-MS-based lipidomics, the same lipid compound can be represented multiple times in the data. The source of this signal duplication are the different adducts and other types of compound modifications that can be formed during electrospray ionization (ESI). Common adducts are expected, such as [M + H]+, [M + NH4]+, [M + Na]+, [M + K]+ in positive ionization mode or [M-H]-, [M-H+HCOO]-, [M-H+CH3COOH]-, [M-Cl]- in negative ionization mode, but these adducts account for only a small fraction of the observed signal duplications. The difficulty in defining possible adducts and compound modifications lies in their dependence on experimental conditions. Here, I present a data-driven approach for their estimation.

Specifically, a set of 29 chemically diverse internal standards was used to determine signals in the data associated with their addition to the sample. A filtration procedure was proposed that resulted in a substantial reduction in the number of features in negative and positive ionization modes, while retaining most of the putatively annotated features.

Results

A set of diverse internal standards (n = 29; detailed list can be found in subsection 6.2.1 of the thesis) was used to assess the possible adduct formation (or other signal duplications) in the untargeted LC-MS method. Five replicate plasma samples were spiked with these standards and compared to five replicates without the additions.

First, Student's t-test was used to compare the lipid abundances in the samples with added standards (spiked samples) and those without additions (non-spiked samples). Features that had lower average abundances in spiked samples compared to non-spiked samples and p-values passing the 1% false-discover-rate (FDR) were defined as spike-derived features (n = 4790 and 1561 for positive and negative modes, respectively; Students' t-test, Benjamini-Hochberg correction).

Next, the spiked-derived features were used to define the m/z differences for subsequent adduct filtration. For each standard, the m/z differences between its main adduct (as described in subsection 6.2.1 of the thesis) and co-eluting spiked-derived features (in a 0.02 minute interval) were calculated. The union of all the m/z differences for all the spiked standards was used to define common differences using a sliding-window histogram along the m/z values and peak detection function, discussed in detail in subsection 6.2.2 of the thesis. The m/z differences present for at least 3 different standards were retained for subsequent filtration (n = 69 and 23 for positive and negative modes, respectively; Table 2). While some masses were expected, others were data-derived and could not be characterized from the literature.

NEGATIVE ionization mode		POSITIVE ionization mode	
m/z difference	number of standards	m/z difference	number of standards
82.00199997	13	21.98199998	19
164.006	11	58.05199997	14
67.98699997	10	83.95199997	11
246.009	9	106.959	9
149.99	8	157.957	8
-24.04400002	8	111.949	8

-14.01500002	8	89.96899997	7
157.989	8	219.926	7
84.97799997	8	79.93999997	7
-74.03700002	7	179.936	7
231.994	7	43.96299998	7
239.993	7	233.943	7
		225.944	7
		128.941	7

Table 2: The top m/z differences defined using spiked-derived features.

Removal of unwanted adducts from the whole set of features detected in plasma samples was performed by searching for co-eluting compounds (in a 0.02 minute interval) with these defined m/z differences. Because the m/z differences were calculated based on one main adduct and one redundant feature, the choice of which feature should be retained, and which should be discarded, was straightforward. After performing standard data filtration procedures, including blank filtration, QC filtration, and ^{13}C isotope removal, 3672 and 2039 features remained in positive and negative modes, respectively (Figure 7). The described signal duplication siltration discarded an additional 1414 and 1060 features in positive and negative modes, respectively (Figure 7). Importantly, the standard adduct removal, when only well-described common adducts were considered, performed poorly (Figure 7). Only 35 and 45 % of features removed by the above-described method would have been found by common adduct filtration in the two modes, respectively.

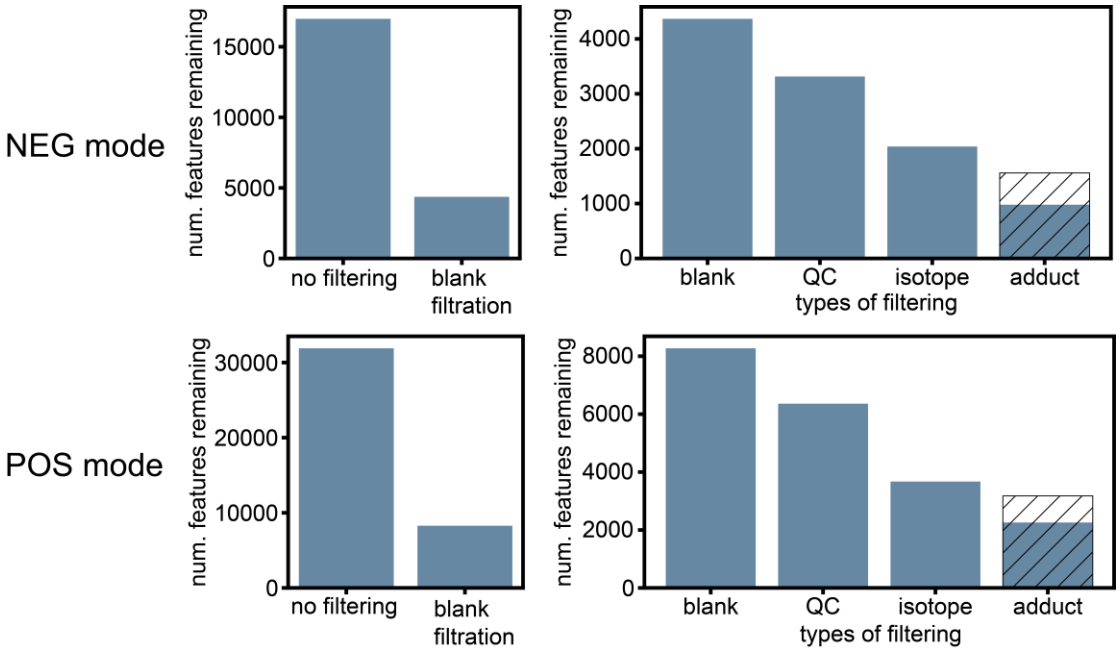

Figure 7: The number of features remaining in non-spiked plasma samples after different filtration procedures. Left: no filtering and filtering by blank samples (removing of contaminants/false-detects). Right: Filtration by blanks, technical replicates (QCs, removing features with technical variability), filtration of isotopes, and filtration of adducts/signal duplications. The dashed bar indicates the number of features remaining after filtration by common adducts alone, described in detail in subsection 6.2.1 of the thesis.

To assess the performance of the adduct filtration procedure, a set of putatively annotated compounds was considered. Of the 258 and 223 annotated lipids in positive and negative mode, 15 and 3 were removed as redundant adducts (Figure 8).

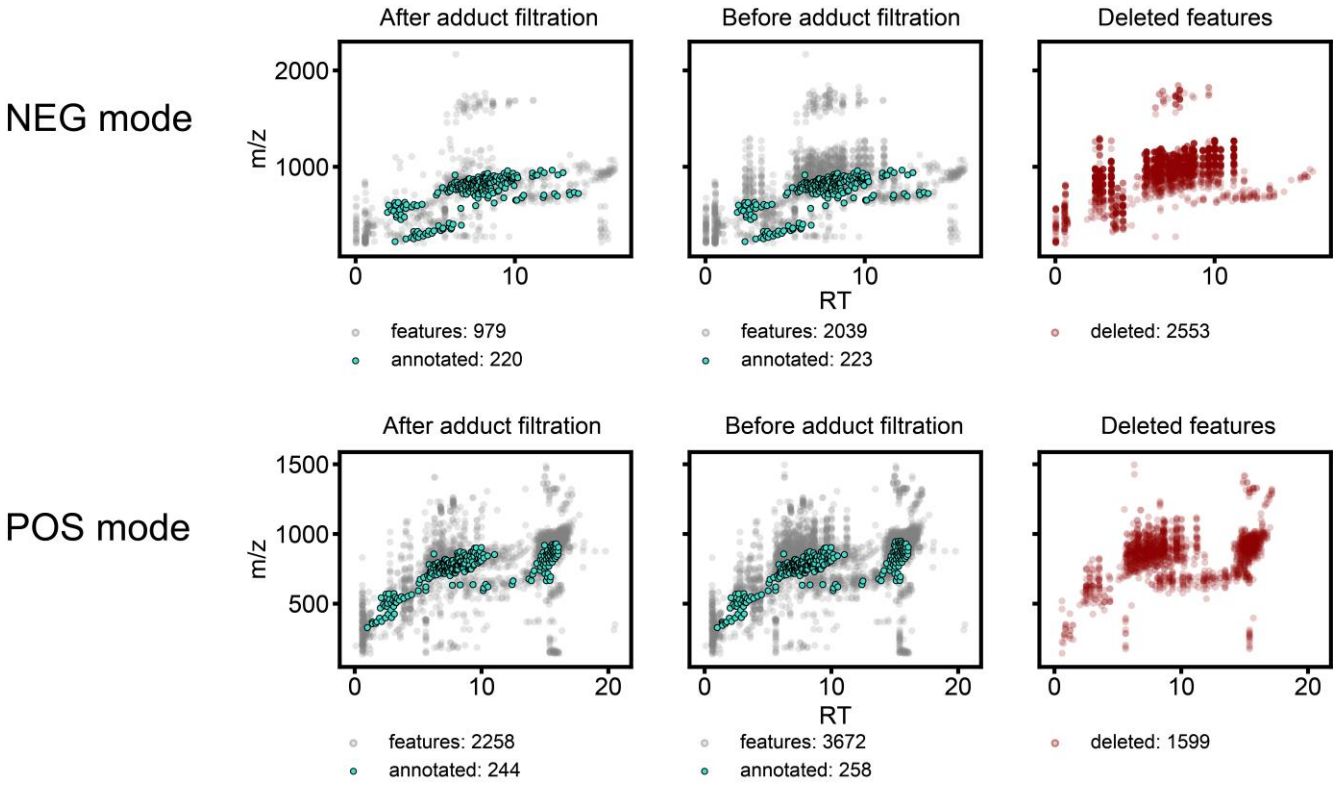

Figure 8: The m/z and RT values of the quantified features in positive and negative ionization modes. From left to right: 1) the features remaining after standard data cleaning and signal duplication filtration, non-annotated (grey points) and the annotated (blue points) features; 2) the features remaining after standard data cleaning before signal duplication filtration, non-annotated (grey points), and annotated (blue points) features; 3) the features deleted by signal duplication filtration, red points. The number of features are indicated below the plots.

As an alternative to the signal duplication filtration proposed here, common m/z differences can be defined from the whole set of quantified features [3]. Subsection 6.2.2 further contains discussion on the advantages of the method proposed here compared to the existing alternatives.

Conclusions

The main results of the thesis are summarized below:

- A workflow for global lipidomics data analysis was proposed, with key processing steps outlined and critically assessed, providing a practical guide for global lipidomics data analysis.
- A data-driven approach was proposed for the reduction of signal duplications in untargeted lipidomics experiments, which removed more than twice as many redundant features than the standard approach of adduct filtration, while retaining most of the annotated features.
- A profile of lipid alterations in the blood plasma of individuals with schizophrenia was discovered, robustly reproduced across several independent sample cohorts
- The described lipid profile was shown to be consistent for other psychiatric disorders: major depressive and bipolar disorders.
- A lipid-based predictive model was proposed, separating individuals with schizophrenia from controls with high diagnostic ability (Area under the Receiver Operating Characteristic Curve, ROC AUC=0.86-0.95) and validated on two separate test datasets.
- A particular lipid profile associated with poor medication response in schizophrenia was described, namely, an increase in shorter-chained triglyceride lipid species was shown to be associated with poor symptom improvement after treatment.
- Lipids were shown to be altered as a results of fluoxetine administration in juvenile macaque brains, while other data modalities, gene expression and polar metabolite, did not show significant alterations. Among lipids, polyunsaturated fatty acids (PUFAs) were shown to be decreased in the brain of macaques that have undergone fluoxetine treatment, with free fatty acids showing strongest effects, but other PUFA-containing lipids exhibiting decreased levels, as well.

References

1. Tkachev A, Stekolshchikova E, Anikanov N, Zozulya S, Barkhatova A, Klyushnik T, et al. Shorter chain triglycerides are negatively associated with symptom improvement in schizophrenia. Biomolecules. 2021. 2021. https://doi.org/10.3390/biom11050720.
2. Tkachev A, Stekolshchikova E, Bobrovskiy DM, Anikanov N, Ogurtsova P, Park DI, et al. Long-term fluoxetine administration causes substantial lipidome alteration of the juvenile macaque brain. Int J Mol Sci. 2021. 2021. https://doi.org/10.3390/ijms22158089.
3. Mahieu NG, Patti GJ. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. Anal Chem. 2017. 2017. https://doi.org/10.1021/acs.analchem.7b02380.