Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Alina Chernova

PhD Program: Life Sciences

Title of Thesis: Integrating high-throughput genotyping and lipidomic profiling for discovery of genetic determinants of cultivated sunflower seed oil content
Supervisor: Prof. Philipp Khaitovich

Chair of PhD defense Jury: Prof. Georgii Bazykin ***Email*** *g.bazykin@skoltech.ru*

Date of Thesis Defense: 29 January 2021

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

I would like to thank all jury members for the careful review of my thesis and valuable comments. I have addressed all comments and questions and have made several changes reported in detail below.

Professor Loren H. Rieseberg

1) **Correcting for multiple tests is tricky with genomic data because there are many markers, which are not independent. Your approach of using LD blocks seems reasonable to me, but it was not clear to me how exactly you estimated them. That is, what thresholds did you use to call an LD block? Also, did you consider using the sunflower trait ontology tool to harmonize your phenotypic data (p. 39) with that generated by other groups (https://www.cropontology.org/terms/CO_359:ROOT/Sunflower%20traits)?**

Thank you for this comment. This info is missing in the text. Now I have added the information about how LD blocks were estimated into the methodology section. I have used the default parameters of Haplowiew software except for the window size, which was increased from 500 Kb to 1500 kb. To estimate the LD, the R2 parameter was used. Regions in strong LD were observed by heat-maps constructed in Haplowiew.
But if to talk about correcting for multiple tests, we just use the average length of LD block (among all chromosomes) determined from the LD decay plot and then the estimation of the total LD block amount based on sunflower genome size.
Unfortunately, I have not used the tool you suggested, but I will consider it for future studies.

2) **Unfortunately, the XRQ reference genome (which was employed for SNP calling) has a large number of mis-assemblies, stemming from the use of a faulty physical map during the assembly process. This can lead to false positives, especially if there are broad and strong GWA peaks. An example is the GWA peak for branching on chr. 10. (p. 54). The branching locus is known to be near the bottom of chr. 10, which is clearly seen in the GWA analysis. However, the additional hits scattered across the rest of the chromosomes are likely false positives. Todesco et al. (2020; Nature 584, 602-607) corrected for this by transferring SNPs to a new reference genome that is assembled correctly, so this is something you could consider as well (in the future!).**

Thank you for this comment. Unfortunately, at the time when most of the data analysis was performed, the reference HanXRQr1.0 was the best available option. But for the future, we definitely consider the SNP transferring to the better references.

3) **With respect to the restorer locus, you should mention that the locus is an introgression from H. petiolaris, which was characterized by Baute et al. (New Phytologist 2015; 206:830-838). Also, Owens et al. (Evolutionary Applications 2019; 12:54–65) reported on copy number variation in cultivated sunflower and found that one of your candidate genes (aldehyde dehydrogenase) shows a 10-fold increase in copy number in restorer relative to maintainer lines. Thus, it is our favored candidate for the restorer of fertility gene. The same paper reports on a PPR gene on chr. 8 that shows copy number differentiation between maintainer and restorer lines. Thus, it represents a candidate gene for the restorer locus on chr. 8.**

Thank you for mentioning this. I have added suggested references to chapter 4.4. This chapter is based on the paper we published in Jan 2019, the same month when Owens et al. published their study. This is the reason why this reference was missed. According to a candidate gene for the rf gene in chr. 8 I can hypothesize that it is dependent on the genetic background, and in some lines, we can find rf in chr. 13 and in other lines in chr. 8 and it can be explained by sunflower whole-genome duplication.

.

Professor Elena Potokina

1) **In chapter 4.2.2. there is an overview of quantitative agronomically important traits variation among genotyped lines from VNIIMK e.g. plant height, head diameter, length of the planting date to flowering, DTF, days; planting-physiological maturity period, 100-seed weight etc. The raw data are also listed in Annex for chapter 4. The question is, if both genotyping and comprehensive phenotyping data were accumulated for the same set of lines, why GWAS results were described only for branching trait? Has the variation of other quantitative morphological and phenological traits also been mapped using GWAS but revealed no significant associations?**

Thank you for this question. That's true that many different phenotypes were collected, and we tried to run GWAS for them as well. But it didn't reveal significant associations. We see the main reasons for this: there were not enough plants and variation among them to get something meaningful for such complex quantitative traits. We tried to fill the gap with plants from other collections, but phenotyping was performed using slightly different protocols, making it challenging to combine the results.

2) **It is understood that for cross-pollinated species such as sunflower, GWAS analysis ideally requires the collection of both genotyping and phenotyping data for the same plant. The point is that in agrobiological practice, one-year phenotyping data for quantitative traits (e.g. plant height) is considered insufficient to assess the variability of a trait for the particular accession (line). What approach would you suggest to solve the problem at least for sunflower?**

I agree this is a complicated problem, and one-year data is considered not sufficient. But there are two ways to collect the data for at least three years or plant the material in 3 different locations and take an average value for the phenotype. But it is not always possible for each GWAS study, and it works just with pure lines. Another way is to go from the trait and understand the GxE interactions for each trait on one cohort of samples and then extrapolate it for other samples. if the trait was shown to be not affected by the environment significantly, it is possible to assess it during one year. However, we understand that this approach is suboptimal but could be used for experiments carried out for sufficiently large cohorts where the high number of traits are estimated. This approach was used in my study, where I work with FA content. To estimate GxE is very important in both scenarios since working with the trait highly affected by the environment can result in false-positive associations.

3) **There are several sections in Dissertation where "LD block" or "LD block length" were mentioned. In Material and Methods chapter it was just mentioned that "LD block analysis was performed using Haploview software". I am wondering what criteria have been set for**

determining where an LD block starts and ends? What threshold (r2?) was required for an LD decay to assume that SNPs still belong to the same LD block?

Thank for mentioning this. I missed this information in the Material and Methods chapter, and I have added it. I have used the default parameters of Haplowiew software except for the window size, which was increased from 500 Kb to 1500 kb. To estimate the LD, the R2 parameter was used. Regions in strong LD were observed by heat-maps constructed in Haplowiew.

Just minor technical comments:
In the "GBS library preparation…" section (page 37) it is written: "To perform the second restriction digestion with NlaIII, Master Mix including 0.7 µl of NlaIII (NEB, USA) with a working concentration of 20U / lL, 2 µL of CutSmart buffer (NEB, USA), and 16.1 µl of mQ + DNA mix was prepared. 2 µl of another adapter (5mM) called "common" 30 was added into each well to be ligated to the overhanging ends generated by NlaIII afterward ".

I was confusing by "adapter called" common "30". Another thing is that one step before in this protocol it was mentioned that "Then, 10 µl was taken from each sample and pooled in one Eppendorf tube", so there are no more wells after this step to which 2µl of another adapter could be added ... Should a common adapter be added to the pooled DNA sample instead?

Thank you for mentioning this. First point you mention is a typo. Adapter was called "common". Now it is corrected.
Yes, you are absolutely right. Second restriction was performed in one tube with pooled DNA after first restriction. And common adapter was added to this tube. I have corrected the text.

Page 35. The sunflower was sowed following the preceding crop, fall wheat… - winter wheat was mentioned?

Yes, winter wheat was considered here. Now it is corrected.

<u>Professor Laurent Gentzbittel</u>

1) **Chapter 2 presented a good review of the literature on the subject area. The reviewer would have benefit from some synthetic and critical review of the literature to provide additional insights, the current text being too descriptive to some extent.**

Thank you for mentioning this. I decided to modify a final part of chapter 3 and add some finalizing notes, which help introduce my study's scope.

2) **Bioinformatics or statistical methods would have benefit from more details and developments to reinforce confidence in the edited results.**

I have extended on LD analysis and primary data processing.

3) **Chapter 5 provided a comprehensive and technical analysis of lipidome profiles in sunflower and some comparisons with rapeseed. Notwithstanding its clear technological**

interest, the research questions in this chapter are not very apparent for a scientist who is not a biochemists or a specialist of lipids. In particular, the value added of these methods in breeding programs (costs, practical aspects, throughput, repeatability, relationships with breeding targets or market demands) would have been interesting to develop.

I agree with this comment. Perhaps it was worth paying more attention to this point in the literature review.

4) **Chapter 6 provided a detailed analysis of a large collection of sunflower lines, with the aim to provide understanding of the genetic bases of lipidomics patterns. A recent method of UPLC-MS was used to carry out the detailed analysis of the profiles. A large number of different lipids and fatty acids were identified. The genetic control of the quantity of some lipids or fatty acids was described. It would have been interesting to try to combine the information using some multivariate methods for example. The provided results may be very important but need additional studies and replications to be validated. The experimental design did not allow evaluating environmental variation with enough details. The fact that open-pollination of plants was allowed (chapter 3, section 3.4.1 page 40) – and thus the genotype of the seed being possibly different from that of the mother plant - raises some interesting questions about the mapping of the phenotype of the seed fatty acid profile to the genotype of the mother plant if cross-pollination occurred during the field trials.**

I fully agree with the reviewer with his first comment. Unfortunately, it was impossible to plant all lines in different locations or one location for at least three years. And we did the small experiment for environmental variation evaluation, but I agree that it cannot fully substitute the data for all lines environmental variation.
According to the second comment about open-pollination: yes, that's true that open pollination was allowed during the phenotyping experiment since it was easier for our collaborators who performed the fieldwork. But it should not affect the fatty acid profile since it is determined by the mother plant (endosperm), pollen with different genotype will not make any difference. Actually, for mass spectrometry measurements, which are central in my work, I used seeds that were obtained in conditions where open-pollination was not allowed (pure line seeds).

<u>Professor Lee Hickey</u>

I want to thank the reviewer for the separate file, including the thesis text with some changes and corrections. I found it very useful and applied the changes in the final text.
Below I address the comments from the review.

1) **Firstly, the paragraph structuring could be revised and improved. This is an issue throughout the thesis and makes it difficult for the reader to follow, particularly when many very short paragraphs occur (i.e. paragraphs containing 1 or 2 sentences only). Remember, each paragraph should be a single idea. Each paragraph should be between 4-8 sentences long, start with an introductory sentence and end with a concluding sentence.**

Thank you for this comment. Paragraphs structure was revised along the manuscript.

2) **The literature review could benefit with a final section that highlights the research gaps or priorities for research. This would help guide and introduce the reader to the topics/research questions to be studied in this thesis.**

I am grateful for this idea. I have added the paragraph at the end of the literature review which helps to introduce the aims of the study

3) **Chapter 3 reports the materials and methods used for the research conducted as part of this thesis. I must say that I am not familiar with this type of structure for a PhD thesis (normally, in Australia the research chapters each comprises an Introduction, Methods, Results and Discussion section), but I think this is just a style thing and the arrangement works for this thesis, so this is fine with me! I think the "samples" section is quite confusing because there is a mix of a description of sunflower accessions, a description of how the seeds were produced for analyzes and also some mention of rapeseed accession… please consider how this section could be presented more clearly and perhaps divided into 2 or 3 smaller sections to help guide the reader?**

Thank you for this comment. I did some rearrangements in "Samples" section and made changes you suggested inside the text.

4) **Chapter 4 reports new insight of the phenotypic and genetic diversity of diverse sunflower accessions, in particular the Russian accessions that were genotyped for the first time. It is critical for researchers and breeders to understand the trait diversity if it is to be harnessed for crop improvement. Highlights from this research include GWAS that detect marker-trait associations for linoleic acid content on chromosomes 8, 9 and 17, and narrowing down a 7.7 Mb region on chromosome 13 associated with fertility restoration (specifically Rf1). A number of candidate genes were identified for the chromosome 13 region associated with Rf1. Based on the haplotype of this segment, and the passport information for the accessions you have genotyped, could you determine the likely origin of the Rf1 gene? Further, it would be very interesting to report whether any Rf1 haplotypes are associated with agronomic traits… Such as association could be possible via pleiotropic effects or through LD as a result of long term**

Sunflower has more than 70 sources of CMS, but in commercial hybrids breeding, the PET1 CMS was obtained by P. Leclercq from the interspecific hybrid H. petiolaris Nutt. × H. annuus is used predominantly. Seed production of sunflower hybrids is based on the extensive use of the Rf1 and Rf2 genes, which interact with each other, giving the effect of restoring pollen fertility. It is believed that the primary gene is Rf1, which is responsible for the fertility restoration and is present in the vast majority of CMS PET1 fertility restoring lines.
Due to CMS's extended use in sunflower breeding, there is a possibility to find Rf1 haplotypes associated with agronomic traits, and we confirmed it for branching. Restorer lines are usually branched and sterility maintainer not. This is because to obtain F1 hybrids, non-branched lines with a single large apical head are often used as female parents, and lines with a recessive type of branching, with multiple small heads located on the lateral branches, are used as male parents. This approach allows an increase in the length of the flowering period of male parents due to the difference in the flowering times of the heads on the plant and getting F1 plants with a single large head. It is known from the literature that the branching locus is localized on

chromosome 10, and we see the associations with Rf1 in this chromosome, which are probably linked to branching.
Also, restorer lines are usually resistant to specific plant diseases, and sterility lines are susceptible (mildew is an example). This is also a consequence of artificial selection.

5) **Should the "Conclusions and future perspectives" section at the end of the thesis be a numbered chapter, like the Introduction (chapter 1)?**

I really do not know why, but in the template which is recommended to follow for the thesis "Conclusions" section is not numbered, so I decided to stick to this structure.

Professor Eric Bishop von Wettberg

1) **The first overarching theme that Chernova briefly addresses in her introduction (page 12) is that with population growth and climate change, immense increases in food production are required. This is a common theme in the contemporary crop genetics literature. It serves a motivation and rationale for characterizing germplasm collections to find adaptive alleles. As common as this trope has become in the literature, I believe it requires some careful examination. Although I think it is still a good motivation, we should not lose sight of the larger dynamics of the food system. Outside of Russia, incredible amounts of food are wasted, particularly in the West. Considerable strides could be made in improving food access and security were 25-40% of food not wasted in the US, Western Europe, and countries with similar food distribution systems due to spoilage. In many market economies, food production is also highly skewed towards production of animal products, which can require 10 calories of grain to provide a single calorie of meat. Furthermore, widespread income inequality means that a few have disproportionate access to food. If our aim is to reduce hunger, efforts on any of these three factors are likely to have a far large and much more immediate impact than raising productivity of any crop. The scope of expected hunger with climate change may still exceed these steps, so raising yields is still imperative. But even with raising crop yields, genetic improvements will likely only account for a quarter of potential for improved yields. Improved agronomic management may do far more to bridge the gap between yields obtained on research stations and those achieved on actual farmers. All of this is not a reason to not also invest in crop genetics, but it is important context. We should keep in mind the relative contribution of different actions, and appreciate that genetics is just one tool. Likely improvements in crop genetics are more important for adaptation to climate change and its associated impacts (new diseases and pests, shifting patterns of abiotic stress). There is some broader historical context for this in a recent book called "The Prophet and the Wizard" (Charles Mann, 2018) that looks at the worldviews that go into how we conceptualize scientific responses to hunger and resource scarity.**

Thank you for your comment. I can only agree with it. Of course, I am aware of the problem of wasted food and understand that the problem is much more complicated than what I explained in my introduction. Maybe it was a mistake to make it so biased, but the idea was to highlight the importance of breeding and genetic studies integrated into breeding programs. Because in Russia, compared to Europe, the US, and Australia, this understanding is very poor. And plant

genetics studies are very unpopular. But yes, the approaches we use should be complex, and we should see the full story. Thank you for this book recommendation. It will be interesting to read.

2) **A second "big-picture" question is about the biological constraints on oil seed production. Why is it that ~60% oil content is the limit of seed oil content? What factors impact natural variation in seed oil content? How do these patterns help guide our efforts to find natural variants in seed oil content that may be agronomically useful? There is a rich ecological/evolutionary literature to be accessed on these topics. Some of this literature is on "economic spectra" in plant traits. For species like sunflowers, where wild relatives in Helianthus and many other taxa in the Compositae family have wind-dispersed seeds, there is a trade-off between dispersal capacity of seeds by wind and the likelihood of germination. Larger seeds are more dispersive, but may individually have a lower germinate rate. In cultivated sunflowers this trait has been lost in domestication, but still has an evolutionary legacy. Oil composition may be shaped for selection on dispersal. A second important aspect of selection on seed fatty acid content is likely temperature. There are clear latitudinal gradients in seed fatty acid content that track background temperature. An older review from Randy Linder (Linder, C.R., 2000. Adaptive evolution of seed oils in plants: accounting for the biogeographic distribution of saturated and unsaturated fatty acids in seed oils. The American Naturalist, 156(4), pp.442-458.) is a good starting point for this literature, as a meeting point between agricultural and natural systems.**
   Understanding natural selection on fatty acid content and ratios can help design approaches to find naturally occurring variants with particular compositions in germplasm accessions. Some frameworks already exist for such searches, such as FIGS (Focused Identification of Germplasm Strategy) and Gap Analyses already exist, or genotype-environment association tests. But these methods are relatively crude, leaving open the possibility of deploying more precise tools to associate climatic conditions in particular environments with genetic variants leading to unique fatty acid profiles. I would be intrigued to see the data collected by Chernova analyzed by some of these tools, such as BayEnv, Bedassle, or Gradient Forest analysis. I suspect these general tools could be adapted to be more precise for looking for oil seed content by adding in formation on environment matching of oil seed content to different ecologies and thermal regimes. Although human impacts on distribution of landraces and cultivars may make associations of environment and genotype less precise, they may still provide insight when paired with predictions about oil seed content.

Thank you for the literature you share and suggested methods. This aspect stays a little bit apart from my study's scope since I was more focused on the methodological part of fatty acid data collection. But I will consider it and use it in my future work. I am sure it can provide exciting results.

3) **A third theme, of narrower scope, is that of minor fatty acid components. I think the discussion of this topic, which is important for the production of specialty oils, deserves more context. What biological role might these FA constituents have? What processes would lead to their low abundance in seeds or other tissues? What are the likely limits on selection in these fatty acids? Would a seed be viable with arachidic or behenic acids as a major constituent? How would its properties change?**

Thank you for this suggestion. I have expanded the discussion on this topic in chapter 6 VLCFA are very interesting minor oil components. They are essential for plant development and act as signal molecules. Considerable progress in the understanding of VLCFAs has been made thanks to the isolation of the elongase mutants. Mutants lack of enzymes essential for VLCFA production showed that they involved in very diverse lipid pools and take part in cellular functions, cell proliferation, differentiation, or death. It is known that their amount can vary between the species. For example, Moringa *olifera* seeds are rich in docosanoic acid and are the primary source of it for the cosmetic industry.

There were experiments with the dramatic increase of VLCFAs in Arabidopsis (More than 30% from total FA fraction) and this cause dramatic alterations in plant morphology. One recent study suggests that overexpression of 2 proteins AKR2A and KCS1, participating in VLCFAs synthesis, increases these FAs in Arabidopsis and leads to cold tolerance. It requires further research, but there is a possibility of producing sunflower breeds with elevated levels of minor fatty acids in the future.

4) **Cytoplasmic Male Sterility has proven a valuable and durable breeding tool in sunflowers. Is there any evidence of CMS breaking down in sunflower? In maize, another crop where CMS was once widely deployed, a disease susceptibility locus (to southern blight) was closely linked to the CMS locus. Is there anything similar in sunflowers? If there is CMS breakdown, or deleterious alleles linked, then finding new CMS genes may be particularly useful. In other crops (pigeonpeas, Cajanus cajan, for example), many new CMS loci have come from wild relatives? Would an examination of more divergent Helianthus species from the secondary genepool be helpful in this regard?**

Thank you for this interesting question. CMS is a hot topic for the sunflower since it predetermined sunflower development as a hybrid crop. I have not seen any records about CMS breaking down in sunflower. But there is a high risk, and it is a common problem for all hybrid crops since sterility loci are usually liked with susceptibilities for different diseases. Sunflower CMS PET1 system, widely used, also came from the wild relative (interspecific hybrid H. *petiolaris* Nutt. × H. annuus). Sunflower has more than 70 sources of CMS already discovered, and research is continuing (they all came from introgressions). But in commercial hybrids predominantly, one system is used because there are no stable restorer genes available for other sources. So, this is extremely important to search for new Rf genes.

Prof. Georgii Bazykin

1) **Overall, the work has few mistakes and typos (and those I have spotted in the first review round have been fixed).**

Thank you for mention this. I performed the proof reading again. Typos and mistakes have been corrected.