

## Jury Member Report – Doctor of Philosophy thesis.


**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:** Professor Philipp Khaitovich

**Name of the Reviewer:** Professor Elena Potokina

<p>I confirm the absence of any conflict of interest</p> <p>(Alternatively, Reviewer can formulate a possible conflict)</p>	<p><b>Signature:</b></p>  <p><b>Date: 28-12-2020</b></p>
---	---

*The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.*

*If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.*

### Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

### **Brief evaluation of the thesis quality and overall structure of the dissertation**

The Dissertation is presented in a good scientific style, combining impressive experimental results and their interpretation. The structure of the thesis is classical for dissertation manuscripts. Chapters 4-6 describing the original results of the study are well illustrated by figures and tables. Each Results' chapter ends with a "Discussion" section that summarizes the main findings.

Chapter 1 provides a short overview of how genomics can help humanity to overcome global challenges such as food security risks. It explains why sunflower took a special place in the world of cultivated plants becoming a model crop to trace introgressions and genome reorganizations occurring on the way for domestication. The aim of the study is defined in the chapter.

Chapter 2 presents Literature Review providing information about sunflower cultivation history in Russia and worldwide, genetic resources, modern trends and peculiarity of sunflower breeding. The special attention was paid to the importance of sunflower oil, its chemistry, biosynthesis and composition. The newest information about progress in understanding of sunflower genome and studies applying NGS for *Helianthus* species are perfectly reviewed. The special section is focused on approaches to understand the genetic basis of quantitative traits in sunflower such as resistance to diseases, abiotic stress tolerance and yield-related traits. Advances in plant phenotyping technologies, including molecular phenotyping (e.g. metabolomics) are discussed. It was suggested that in the post-genome era genomic selection may determine the future of sunflower breeding. Overall, the chapter provides a consistent overview of all the relevant information to understand and evaluate the author's original research.

Chapter 3 describes Materials and Methods. The plant material used for the study looks impressively extensive. 638 samples including 292 accessions from the VIR world germplasm collection, 199 inbred lines from the leading Russian sunflower research Institute VNIIMK and 147 oil-producing sunflower lines from private breeding company Agroplasma Seed provides a very good solid base for the study. Procedure of genotyping performed by NGS sequencing was described in detail, including DNA extraction, GBS library preparation and sequencing and computational analysis of sequencing data. Agrobiological traits scored in the field were listed. The procedure of lipid extraction and UPLC-MS profiling was described. The special attention was paid to the lipidomic primary data analysis and annotation, Figure 1 helps to understand the procedure. Computational approaches for population structure estimation, LD analysis, and association search are described referring all the relevant software employed. The Materials and Methods chapter is well structured and it separately describes the approaches that were used in each of the three subsequent Results chapters.

Chapter 4 presents the results of genetic and phenotypic diversity of the 186 sunflower lines from the VNIIMK collection and 134 lines from Agroplasma Seed Company.

65,553 SNVs revealed for the VNIIMK collection allowed to estimate LD block length distribution. Remarkably, mean of the LD block length (110.517 Kb) differs greatly from median (0.053 Kb) suggesting significant bias of recombination frequency that occurs in the chromosomes of sunflower. Combining genotypes and phenotypes of inbred lines from the VNIIMK collection, GWAS analysis revealed significant association for branching trait and linolenic acid (18:3) content in the seeds.

28,153 SNP discovered in 134 sunflower accessions from Agroplasma Seed Company split the accessions in two clusters corresponding to sterility maintainers and restorer lines. The main loci associated with fertility restoration, according to performed GWAS, were located on LG 10 and 13. The QTL on LG 10 was affected by the population structure and coincided perfectly with the locus associated with branching trait. However, in chromosome 13 a 7.72 Mb long section was defined where eight highly significant SNPs are located. Remarkably, the identified 7.72 Mb region was located within segment of chromosome 13 flanked by SSR markers previously described as co-segregating with Rf1 locus. Based on this finding the Section 4.3.3 describes the procedure of Rf1 candidate genes Identification using tools of ExPASy SIB Bioinformatics Resource Portal. As the result, the list of 21

identified candidate genes is shown in Table 2 and their arrangement within the 7.72 Mb region is shown by Figure 11.

Chapter 5 is rather a methodological chapter that presents results of the implementation of Ultra-performance liquid chromatography-mass spectrometry (UPLC-MS) technology for fatty acids (FAs) profiling in sunflower and rapeseed seeds. Using as a model “a small cohort of samples” (50 sunflower and 50 rapeseed accessions), the author described the optimized method of lipid extraction and MS profiling protocols intending to use the obtained results further in FAs profiling for GWAS. Results obtained for the same set of plants with commonly used GC-FID (gas chromatography-flame ionization detection) were compared with UPLC-MS data. Comparison of different techniques for quantitative assessment of FAs in sunflower and rapeseeds is well illustrated by Figures 16-17.

In the chapter the special section is devoted to the UPLC-MS assessment of triacylglycerides (TAGs) in 50 sunflowers and 48 rapeseed lines. TAGs, which were common for the two crops as well as TAGs exhibited significant differences between sunflower and rapeseed were identified. Further section describes results of TAG comparison between winter and spring rapeseed.

Chapter 6 described discoveries of the novel genetic determinants of oil fatty acid content in sunflower based on genotyping and lipid profiling of 601 cultivated sunflower lines. Two to three plants per line were used for GBS, thus totally 1490 genotypes were assessed. SNP calling was preformed using HanXRQr1.0 reference genome; 2,360,111 SNPs spanning all 17 chromosomes were identified. The population structure analysis revealed a distinct group of genotypes derived from the Agroplasma collection. The revealed genetic diversity of Russian sunflower germplasm was compared with those collected worldwide using public available data based on 2345 SNPs shared between the datasets.

The story then moved on to preparing molecular phenotyping data for GWAS purposes. Since the ecological and biological reproducibility of FA and TAG data estimation is a key issue for association analysis, the model setup experiment was conducted. 6 sunflower inbred lines, each in 5 biological replicates per each of 3 years, yielding a total of 89 accessions were genotyped (GBS) and phenotyped (UPLC-MS estimation of FAs and TAGs) in the same way to test the effects of the genotype-environment interaction using ANOVA. Next, computational annotation of intact lipidome of the oil samples extracted from 601 sunflower lines was performed yielding 687 lipids. Finally, GWAS was performed for 543 accessions for which both genotype and lipid intensity data were available. Only computationally annotated 27 fatty acids (FAs) were subjected to the association study, 23 of them satisfied the criteria for GWAS. Significant associations for eleven FAs were identified, for which SNP annotation and candidate gene identification were performed using the boundaries of the corresponding LD blocks. 429 genes within the LD blocks were involved in oil metabolism, 124 candidate genes located close to significant SNPs reported in the present study. LD blocks with significant associations are listed in Table3. The Discussion section of Chapter 6 explains how the study contributes to the genetic characterization of Russian sunflower collections, and how the findings can be implemented in future studies.

The concluding part of the thesis “Conclusions and future perspectives” describes how the obtained results could contribute to the characterization of the genetic diversity of the Russian sunflower collection, showing ways to employ this genetic resources to a broader spectrum of practical applications.

The Bibliography section contains a comprehensive list of 319 references.

Especially worth noting is the large Annex section, containing raw phenotyping data, supporting tables and figures as well as actual experimental results to ensure their validity.

### **The relevance of the topic of dissertation work to its actual content**

The aim of the study was to characterize Russian sunflower germplasm collections at the whole-genome level, to obtain lipid profiles of genotyped lines and use them to search for genotype-phenotype associations in order to identify candidate genetic markers of phenotypic traits and candidate genes involved in phenotypic expression.

The actual content describes results of high-throughput genotyping (GBS) and high-throughput oil lipidome phenotyping (UPLC-MS) of 601 accessions from three Russian sunflower germplasm collections (Vavilov seed bank, VNIIMK Applied Agricultural Institute, and Agroplasma Breeding Company). As the results, the genetic variation of Russian sunflower accessions was compared with an international sunflower collection containing 1374 wild and cultivated accessions. Further, genetic variants for sunflower linked to classic phenotypic traits (pollen fertility restoration), as well as SNPs linked to molecular phenotypes variation (linoleic acid content) were determined. Thus, the actual content of dissertation fits perfectly to the declared topic.

#### **The relevance of the methods used in the dissertation**

A broad spectrum of methods was used in this study. The whole genome association analysis demands high throughput genotyping data, they were obtained using GBS approach, namely the RADseq technology. According to RADseq protocol DNA libraries were constructed using two restriction endonucleases – rarely and frequently cutting enzymes. RADseq is widely used genomic approach for high-throughput SNP discovery and genotyping of non-model organisms, its relevance to the performed study is clear. The large set of NGS output data required computational analysis of sequencing data including SNP calling via mapping of Illumina reads onto the *Helianthus annuus* reference genome HanXRQr1.0. To do that appropriate bioinformatics tools were used.

For sunflower molecular phenotyping the author employed mass-spectrometry analysis of sunflower seed extracts via UPLC-MS (ultra-high-performance liquid chromatography coupled with mass spectrometry) technology to study lipidome variability of diverse cultivated sunflower accessions from three Russian sunflower collections. The protocol of UPLC-MS profiling of sunflower seeds extract is well described in Methods section. Results of UPLC-MS-based oilseed crop fatty acids profiling were experimentally compared with those obtained using alternative gas chromatography-flame ionization detection (GC-FID), the effectiveness of UPLC-MS method was confirmed. Finally, GBS sequencing data and the UPLC-MS profiling data were combined to find SNPs associated with specific lipid phenotypes. To do this a GWAS analysis was performed using the mixed linear model (MLM) approach. All methods used in the dissertation are relevant to the assigned tasks.

#### **The scientific significance of the results obtained and their compliance with the international level and current state of the art**

The results of the study revealed in the first time the genetic and phenotypic diversity of the Russian sunflower germplasm collection using high throughput genotyping which had not been done before. The analysis discovered that the Pustovoit All-Russia Research Institute of Oil Crops, which is a leading Research Breeding Institute in Russia keeps some unique sunflower germplasm, which is not presented in the VIR worldwide collection. In addition, the diversity of Russian sunflower germplasm, assessed using a large set of SNPs, showed that some cultivated lines of sunflower from Russia carry alleles not present in lines collected all over the world. These discoveries may contribute to the development of sunflower breeding programs as well as strategies for the conservation of sunflower genetic resources.

The dissertation is the first completed metabolomic Genome Wide Association Study (mGWAS) in Russia conducted for an agricultural crop, which discovered specific markers and candidate genes associated with the variability of important agricultural traits assessed using the metabolomic approach. Internationally, metabolomics association studies (mGWAS) were implemented early on *Arabidopsis*, maize, rice, tomato, but not on sunflower.

For the sunflower Rf1 locus on chromosome 13 the genome region spanning 7.72 Mb was firstly defined containing candidate genes responsible for pollen fertility restoration in sunflower. This provides significant contribution to the further study of the genetic nature and molecular mechanisms of this important agricultural trait.

It was shown that UPLC-MS approach has great potential to be used in the evaluation of the FA composition of oil crops as highly sensitive and suitable for the individual seed analysis technique which

can be successively applied for the precise identification of FA profiles of oilseed crops. UPLC-MS mass-spectrometry approach allowed quantitative measurements of fatty acids in sunflower oil in minor amounts, which were not previously assessed.

Six large LD blocks containing SNPs significantly associated with FA content variation in sunflower seeds within chromosome 3 were firstly discovered. One of them possibly contains one of the key regulators of sunflower oil FA composition. Furthermore, strong genetic determinants associated with minor fatty acids content in sunflower oil (docosanoic acid content) were identified, indicating that such fatty acids could also be selected as potential breeding targets for marker-assisted selection. Altogether, the research discovered new candidate genes affecting oil content variation and oil composition in sunflower seeds.

### **The quality of publications**

Four solid research papers containing the main results presented in the dissertation have been published in reputable scientific journals, including PeerJ (Q1 in Agricultural and Biological Sciences) and Biomolecules (Q1 in Biochemistry). The main findings of the study were presented in the Fifth International Scientific Conference PlantGen2019 (June 24–29, 2019, Novosibirsk, Russia) and published in the Proceeding of the Conference. One more publications focused on the discovered novel genetic determinants of oil fatty acid content in sunflower is currently under review. I believe that the publications perfectly reflect the content of the dissertation.

### **The summary of questions and comments to be addressed:**

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?

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?

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 ( $r^2$ ) was required for an LD decay to assume that SNPs still belong to the same LD block?

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/IL, 2 µL of CutSmart buffer (NEB, USA), and 16.1 µl of mQ+DNA mix was prepared. 2 µl of another adaptor (5mM) called “common” 30 was added into each well to be ligated to the overhanging ends generated by NlaIII afterward”.*

I was confusing by “adaptor 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 adaptor could be added. Should a common adaptor be added to the pooled DNA sample instead?

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

Provisional Recommendation
<input checked="" type="checkbox"/> <i>I recommend that the candidate should defend the thesis by means of a formal thesis defense</i>
<input type="checkbox"/> <i>I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report</i>
<input type="checkbox"/> <i>The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense</i>