Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genotyping and lipid profiling of 601 cultivated sunflower lines reveals novel genetic determinants of oil fatty acid content

Alina Chernova (alin.chernova@gmail.com)
Skoltech https://orcid.org/0000-0001-7022-2790
Rim Gubaev
Skolkovo Institute of Science and Technology: Skolkovskij institut nauki i tehnologij
Anupam Singh
University of Southern California
Katrina Sherbina
University of Southern California
Svetlana Goryunova
Vavilov Institute of General Genetics
Elena Martynova
Skolkovo Institute of Science and Technology: Skolkovskij institut nauki i tehnologij
Denis Goryunov
A N Belozersky Institute of Physico-Chemical Biology: Moskovskij gosudarstvennyj universitet imeni M V Lomonosova Naucno-issledovatel'skij institut fiziko-himiceskoj biologii imeni A N Belozerskogo
Stepan Boldyrev
Skolkovo Institute of Science and Technology
Anna Vanyushkina
Skolkovo Institute of Science and Technology
Nikolay Anikanov
Skolkovo Institute of Science and Technology
Elena Stekolshchikova
Skolkovo Institute of Science and Technology
Ekaterina Yushina
FSBSI N P Bochkov RCMG: FGBNU Mediko-geneticeskij naucnyj centr imeni akademika N P Bockova
Yakov Demurin
Pustovoit All-Russia Institute of Oilseed crops
Zhanna Mukhina
All-Eussian rice research institute
Vera Gavrilova
VIR

Irina Anisimova
VIR
Yulia Karabitsina
VIR
Natalia Alpatieva
VIR
Peter L Chang
University of Southern California
Philipp Khaitovich
Skolkovo Institute of Science and Technology
Pavel Mazin
Skolkovo Institute of Science and Technology: Skolkovskij institut nauki i tehnologij
Sergey Nuzhdin
University of Southern California

Research article

Keywords: Sunflower, genetic markers, UPLC-MS, GBS, GWAS, fatty acids, triglycerides

Posted Date: November 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-108244/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at BMC Genomics on July 5th, 2021. See the published version at https://doi.org/10.1186/s12864-021-07768-y.

Abstract

Background: Sunflower is an important oilseed crop domesticated in North America approximately 4000 years ago. During the last century, oil content in sunflower was under strong selection. Further improvement of the oil properties achieved by modulating its fatty acid composition is one of the main directions in modern oilseed crop breeding.

Results: We searched for the genetic basis of fatty acid content variation by genotyping 601 inbred sunflower lines and assessing their lipid and fatty acid composition. Our genome-wide association analysis based on the genotypes of 15,483 SNPs and the concentrations of 23 fatty acids, including minor fatty acids, revealed significant genetic associations for eleven of them. Identified genomic regions included novel loci on chromosomes 3 and 14 cumulatively, explaining up to 34.5% of the total variation of docosanoic acid (22:0) in sunflower oil.

Conclusions: This is the first large scale implementation of high-throughput lipidomic profiling on sunflower germplasm characterization. This study contributes to the genetic characterization of Russian collections, which made a substantial contribution to the development of a sunflower as the oilseed crop worldwide and provides new insights into the genetic control of oil composition that can be implemented in future studies.

Background

Sunflower is an important oilseed crop, domesticated from wild populations in North America approximately 4000 years ago [1] and introduced into Europe in the 16th century. Compared to wild sunflower, domesticated lines show significant differences in branching, flowering time, plant height, and various seed traits, including oil content [2]. The cultivation of sunflower as an oilseed crop dates back to the beginning of the 19th century. Russian academician V.S. Pustovoit and his colleagues selected sunflower varieties with higher seed oil content, culminating in developing the Peredovik 11 variety in 1958 with oil content increased to 51% [3]. Russian sunflower varieties with high oil content formed the basis of sunflower breeding worldwide, leading to global sunflower cultivation for oil [4–6].

Today, sunflower is one of the main oilseed plants [7], ranking fourth in vegetable oil's global production after oil palm, soybean, and rapeseed [8]. It is cultivated on 26 million hectares, with an average yield of 1.78 metric tons per hectare [9]. In addition to the food industry, sunflower oil is used for polymer synthesis, as a biofuel source, and as an emulsifier or lubricant [10]. Nutritional properties and industrial use of sunflower oil depend heavily on the fatty acid residue composition of main oil lipids, triacylglycerides (TAGs), and some of the minor lipids [11, 12].

Considering the need for varieties with improved oil properties, developing novel varieties with desired oil composition is one of the main directions in oilseed crop breeding [13]. This work requires careful characterization of oil composition using metabolomics and lipidomics techniques, such as ultra-performance liquid chromatography coupled with mass spectrometry (UPLC-MS). For the past few years,

there has been a significant increase in the number of studies in plants implementing these techniques. For instance, LC-MS-based technologies were successfully implemented for profiling more than 260 polar metabolites and non-polar leaf lipids in *Arabidopsis thaliana* [14], as well as for characterization of polar metabolites and lipids in other plants, such as tobacco [15], potato [16], corn [17], soybean [18], sunflower [19], and others [20].

Genome-wide association studies (GWAS) coupled with high-throughput lipidome phenotyping promise to identify genetic variants associated with fatty acid content in sunflower oil lipids. This knowledge will aid genome-based selection and help sunflower breeding programs speed up selecting sunflower breeds with desired fatty acid composition [10]. Today, the rapid development of high-throughput phenotyping and genotyping approaches, as well as the availability of well-assembled and annotated genomes, significantly facilitated the understanding of the genetic basis of oil composition in major oilseed crops like soybean and rapeseed [21, 22]. Although sunflower is an important oilseed plant and its oil composition is one of the key agricultural traits, most of the current association studies based on high-throughput genotyping techniques are focused on understanding the genetic control of classical agricultural phenotypes. These studies indeed succeeded in the identification of genetic loci associated with flowering time [23, 24], male fertility restoration [25], seedling growth [26], the plasticity of oil yield for combined abiotic stresses [27], basal and apical branching [28], and flower morphology [29] traits. Further, functional analysis of sunflower genome mapped 125 chemical reactions contained in 12 oil biosynthesis pathways to 429 sunflower genes [30].

Studies based on the techniques allowing to obtain genotypes for a limited number of loci lead to identifying the quantitative trait loci (QTLs) associated with the major oil fatty acids: oleic, linoleic, stearic, and palmitic [31, 32]. Other fatty acids composing the sunflower oil lipids, however, were not assessed for association mapping.

Lipidome profiling by UPLC-MS has already been used as an input in GWAS for several plants, for example, maize [33]. However, for sunflower, associated regions have only been identified for several of the most abundant FAs [34]. Likewise, genomic predictions have been done for the general oil content trait, but not for its individual components [35]. The presently documented natural diversity of lipids contained in seeds demonstrates that domesticated oilseed crops like sunflower can serve as a source of rare FAs. This highlights the value of high-throughput lipid phenotyping, which, in combination with the genotype data, establishes a significant potential for oil improvement by customizing its content [36].

In this study, we combined high-resolution lipidome phenotyping and genome-wide genotyping of 601 inbred sunflower lines to perform GWAS to identify genetic variants associated with the fatty acid composition variation among the lines. Furthermore, this work has substantially widened the analyzed sunflower genetic diversity by incorporating the genotypes from Vavilov germplasm collections [37], which nearly doubles the number of inbred lines ever used in sunflower GWAS [29]. Our analysis yields genetic variants and candidate genes associated with eleven fatty acids, including five minor ones, which have not been previously analyzed.

Results

GBS sequencing and SNP calling

We extracted DNA from inbred sunflower lines from the Vavilov seed bank, VNIIMK Applied Agricultural Institute, and Agroplasma Breeding Company collections (Table S1). Two to three technical replicates of each sample were sequenced on the Illumina HiSeq 4000 platform using a GBS protocol (see *Experimental Procedures*), resulting in 1490 genotypes. Reads were mapped onto the *Helianthus annuus* reference genome (HanXRQr1.0), with the mapping rates varying between 75 and 90%. Variant calling identified 2,360,111 single nucleotide polymorphisms (SNPs) spanning all 17 chromosomes. Homozygosity and Principal Component Analysis (PCA) showed no obvious bias concerning plate batch or seed bank variables (Figure S1)

Population structure, relative kinship, and linkage disequilibrium

We assessed the population structure of the genotypes used in GWAS analysis using the ADMIXTURE package. No visible clusters were observed for cases of K = 1:10 (Fig. 1A). However, visualization of genetic variation using the first two principal components of PCA revealed a distinct group of genotypes derived from the Agroplasma collection, which clustered separately (Fig. 1B). While the average genotype correlation (r^2) dropped to half of its maximum value at 0.7 Mb, the linkage disequilibrium (LD) decay varied among the 17 chromosomes (Fig. 1C, D; Figure S2). Notably, some chromosomes, such as chromosome 3, demonstrated extended LD within the 1–3 Mb interval (ANOVA, $p < 0.0001$).

Genotypes variability and relation to other sunflower germplasms

To place analyzed cultivars on a broader map of sunflower genotype variation, we compared our genotypes to previously sequenced 1065 wild sunflower varieties, 20 landraces, and 289 cultivated sunflower lines [38]. Principal component analysis just on cultivated lines and landraces based on 2345 SNPs shared between the datasets showed that cultivated sunflower lines from the Russian dataset are distinguishable from those collected worldwide by the third principal component (Fig. 2A, 2B). The analysis further reaffirmed the broad genetic difference between cultivated and wild material (Figure S3A). However, it has to be noted that such analysis that is confined to the positions polymorphic in both datasets could, therefore, underestimate the differences between the datasets. The third principle component further separated some of the *Helianthus* species (Figure S3B). Notably, the landraces present in the Hübner dataset mostly situated between the cultivated lines from foreign and Russian collections and the wild sunflower varieties (Figure S3A, S3B).

Oil lipidome quantification

We extracted the total lipid fraction from sunflower seeds of same sunflower lines used in the genotype analysis. We then divided the lipid extracts into two fractions and analyzed them independently using UPLC-MS technology. The first fraction was kept intact, while the second was hydrolyzed before the analysis. The hydrolyzed fraction contained fatty acid residues of all oil lipids and a minor fraction of free fatty acids present in intact samples before hydrolysis (FAs). Mass spectrometry analysis yielded

826 computationally annotated lipid peaks and 27 post-hydrolysis fatty acids. We focused on a specific lipid class, the most important among sunflower oil lipids, the triacylglycerides (TAGs), in our further analysis of intact lipids. To optimize the detection of both high and low abundancy FAs and TAGs, we conducted the UPLC-MS measurements at two extract dilutions (see *Experimental Procedures*).

Quantification of genetic and environmental effects on oil lipidome composition

To assess FA and TAG data environmental and biological reproducibility, we grew plants from six sunflower inbreed lines (1 conventional and 5 high oleic) originating from the VNIIMK collection for three years with five biological replicates per year yielding a total of 89 accessions (Table S1). We conducted genotyping using the same GBS protocol and UPLC-MS measurements at extract dilutions to ensure quantitative coverage of the entire concentration range. We then tested the effects of the genotype-environment interaction using ANOVA with the following model: G + E + G:E (where G is genotype and E is environment, *i.e.* year). All FAs and TAGs measured in both dilutions displayed significant differences between genotypes after BH-correction ($p < 0.05$, Figs. 3A, S4A, S5A, S6, S7, S8, Table S2). Further, most FAs (11 out of 15) and TAGs (32 out of 42 and 43 out of 59 for 1:25 and 1:3 dilutions, respectively) also showed significant G:E interaction. However, the interaction effect, although statistically significant, had a much smaller amplitude than the effect of the genotype (Figure S9). Biological replicates of the same genotype collected in different years displayed greater similarity than plants of different lines collected in the same year The strongest variation among genotypes was observed for oleic, linoleic and palmitic acids, the major fatty acids in sunflower oil (Fig. 3B-E), as well as for the following TAGs: 50:2, 51:3, 54:3, 54:4, 54:6 (Figure S4 B-E and S5 B-E). Analysis of Nei's genetic distances between genotypes obtained for the same seeds used for lipidomic analysis showed the reproducibility of genotypes between biological replicates (Table S3).

Oil lipidome variation analysis

Computational annotation of intact lipidome of the oil samples extracted from 601 sunflower lines yielded 687 lipids falling into seven lipid classes: glycerolipids (GL), glycophospholipids (GP), free fatty acids (FA), sterols (ST), prenols (PR), polyketides (PK), and saccharolipids (SP) (Fig. 4A). A subclass of glycerolipids, TAGs, occupied 87% of lipid intensities of uniquely identified compounds (Fig. 4B, Table S4). The most present TAGs were 54:6, 54:5, 54:4, 54:3, 52:4, 52:3, and 52:2 (Fig. 5A). Among computationally annotated 27 FAs, the highest abundance FAs were 18:1, 18:2, 16:0, and 18:0 (Fig. 5B). The statistics on each fatty acid are presented in Table S5.

Association analysis

Of the 601 sunflower lines taken into the study, we obtained both genotype and lipid intensity data for a total of 543 accessions. We conducted GWAS analysis using the mixed linear model (MLM) approach to test for genetic determinants of FAs variation based on these data. The analysis included 15068 SNPs that passed the filtering criteria (missing calls rate < 0.3, DP > 4, MAF > 0.01) for oleic and linoleic acids and 12528 SNPs for other fatty acids (missing calls rate < 0.3, DP > 4, MAF > 0.03). From 27 detected FAs 23, satisfying the criteria for GWAS were selected. Of 23 analyzed FAs, we detected significant associations for eleven: stearic acid (18:0), oleic acid (18:1), linoleic acid (18:2), nonadecanoic acid

(19:0), eicosanoic acid (20:0), docosanoic acid (22:0), tetracosanoic acid (24:0), tetracosenoic acid (24:1), hexadecadienoic acid (16:2) and such rare fatty acids as 17:2 and 19:2 (MLM, Bonferroni-corrected $p <$ 0.00001; Fig. 6; Figure S10A-F). We further performed GWAS for the oleic-linoleic acid ratio yielding six significant SNPs as response variables in MLM (Bonferroni-corrected $p < 0.00001$; Figure S10G-I). Altogether, we identified 140 trait-associated SNPs (MLM, Bonferroni-corrected $p <$ 0. 0.00001; Fig. 6A). Among them, docosanoic acid (22:0) abundance variation showed the strongest association with 53 genotype variants located on chromosomes 3 and 14 (Table S6). These genetic variants cumulatively explain up to 35.4% of quantitative variation of docosanoic acid abundance among sunflower lines.

SNP annotation and candidate gene identification

To annotate genes potentially linked to genetic variants significantly associated with FAs quantitative variation, we determined the boundaries of the corresponding LD blocks (Figure S11, Table 1). We then retrieved annotation of all genes located within these LD blocks and checked their intersection with the genes annotated to be involved in sunflower oil metabolism [30] (Table S7). From 44144. sunflower genes it was reported 429 genes involved in oil metabolism [30]. Among 124 candidate genes located close to significant SNPs reported in the current study and stayed in LD with this SNPs, four genes coincide with genes from the Badouin list, which is significantly more than expected by chance (Fisher exact test, p = 0.03, odds ratio = 3.4). According to the sunflower genome annotation, these genes encode putative beta-hydroxyacyl-(acyl-carrier-protein) dehydratase FabZ, HotDog domain protein, probable phosphatidic acid phosphatase (PAP2) family protein, and putative MYB-CC type transcription factor, LHEQLE-containing domain and are located on Chr3 and Chr14 (Fig. 6C, D).

Table 1
LD blocks with significant associations

Phenotype	Chromosome	LD block Location		Length (kb)
		Start position	End position	
Oleic Acid (18:1)	6	64066219	64889534	823
	9	168736699	169306761	570
	13	116940760	117370881	430
	15	38043597	38078709	35
Linoleic Acid (18:2)	3	66733584	68666170	932
	5	37199838	37569381	396
	11	5004818	50619247	414
	11	95051157	92468132	416
Linolenic acid (18:3)	11	43846946	44328722	481
Oleic/Linoleic ratio	3	66733584	68666170	932
	12	121534492	121906701	372
Nonadecanoic acid (19:0)	2	179620148	179872251	252
	14	53394600	53480813	86
	14	59829070	60503626	664
Docosanoic acid (22:0)	3	32332262	32562669	230
	3	42596595	43078214	481
	3	44696624	46188263	1491
	3	48304030	49705352	1401
	3	53949047	54230339	281
	3	57635146	57714809	79
	14	91496885	91547710	50
	14	96632645	97927934	614
	16	176846705	176869659	22
Tetracosanoic acid (24:0)	2	56777868	56880436	102

Phenotype	Chromosome	LD block Location		Length (kb)
		Start position	End position	
	2	73398255	74229960	831
	3	102040303	102070280	29
Nervonic acid (24:1)	3	44696624	46188263	1491
	3	57635146	57714809	79

Discussion

Our results broaden the list of candidate genes and genetic variants associated with sunflower oil lipids' fatty acid composition. Our analysis, based on UPLC-MS mass-spectrometry included the quantitative measurements of fatty acids present in sunflower oil in minor amounts, which were not previously assessed. Our results indicate that there are genetic loci with a substantial effect on the quantitative phenotype for at least some of these minor fatty acids, such as docosanoic acid (22:0). It requires further research, but there is a possibility in the production of sunflower breeds with elevated levels of minor fatty acids in the future.

The reason for the lack of significant signals for 12 of 23 analyzed fatty acids, as well as relatively weak genetic signals for many of the eleven fatty acids with identified associations, could lay in the selection of the analyzed lines. The 543 lines used in the GWAS analysis, as well as all the 601 lines used in our study, were not preselected to contain contrasted phenotypes for fatty acid content, with the exception for oleic and linoleic acids, because until now, only these fatty acids together with stearic and palmitic acids were considered in breeding [39]. Nonetheless, our approach involving a large number of diverse cultivated lines yielded enough variability for nine more sunflower fatty acids to produce significant genetic associations. Further work involving lines specifically selected to vary in terms of fatty acid content is required to determine the full scope of genetic associations underlying sunflower oil composition.

We have identified six large LD blocks containing SNPs significantly associated with FA content variation within chromosome 3 (Table 1). Furthermore, among the reported candidate genes predicted to affect oil quality, three genes associated with lipid metabolism localized within the large 1,491 kb LD block of chromosome 3 (Figure S11, Table 1). These genes encode the putative phospholipase A2 (this protein releases FAs from the phospholipid), putative CRAL-TRIO lipid-binding domain-containing protein, and putative ethanolamine-phosphate cytidylyltransferase. Predicted functions of these genes, although not yet assessed experimentally in sunflower, single out this genomic region as one of the key regulators of sunflower oil FA composition (Fig. 6B, Table S7). Further, among the genes located within the

chromosome 14 region associated with FA 22:0 variation, two were annotated as membrane-bound proteins: putative membrane-bound transcription factor site-2 protease, and putative membrane-bound O-acyl transferase (MBOAT). This finding agrees with the fact that very-long-chain fatty acids in sunflower are synthesized by membrane-bound enzymes [40].

Among the genes important for fatty acid metabolism according to [1] and located within the LD blocks linked to FA variation, one of the most interesting is the gene encoding a putative FabZ dehydratase, the protein responsible for FA elongation (Fig. 6C). It has to be mentioned, however, that the genomic resolution of our study is limited to LD blocks, which typically include multiple genes. Thus, further work is needed to map associations to specific genes and causative genetic variants.

Genetic variants (SNPs) linked to the oleic-linoleic acids ratio also map to a chromosome 3 region (302 kb region; Figures S10I, S11). This LD block overlaps with the one carrying SNPs significantly associated with linoleic acid content (Figures S10H, S11). This finding supports the notion that genomic regions underlying linoleic acid content should also be involved in oleic-linoleic acids ratio determination. Unfortunately, there were no annotated genes known to be directly related to fatty acid biosynthesis or modification in this region. Previous studies demonstrated that genes encoding desaturases, the major enzymes responsible for the oleic-linoleic acids ratio, are located on the chromosomes 1, 12, 14 [41]. Nonetheless, loci potentially associated with oleic and linoleic acid contents were previously identified on chromosome 3 by means of QTL mapping [32, 34], as well as by computational predictions [30]. These loci, however, did not overlap with the locus obtained in the current study. A number of reasons could cause the fact that the previously reported regions potentially related to oleic-linoleic acids ratio were not identified by association mapping in the present study. First, the SNP coverage for these regions might not have been dense enough in our study. Second, previously identified associations might play lesser roles in determining linoleic and oleic-linoleic acid ratio under Russia's environmental conditions. Third, lack of overlap could be related to the specific genetic features of the studied cohort that was restricted to the lines from the Russian collections.

In addition to genetic variants linked to oleic-linoleic acids ratio, we have identified nine and 22 SNPs significantly associated with individual oleic and linoleic acid content. These SNPs localized on chromosomes 9, 13, 15 for oleic acid and 3, 11, 12, 14, 15, and 17 for linoleic acid. Previously, a study reported QTLs identified by means of ORS markers for oleic acid content on chromosomes 8 and 9, and linoleic acid content on chromosomes 8 and 14 [42]. We also identified significant associations and putative candidate genes on these chromosomes for linoleic acid and on chromosome 9 for oleic acid. However, our chromosome 9 LD block did not overlap with the QTL associated with oleate reported by [30].

Interestingly, we have additionally identified a putative FAO1 gene on chromosome 9 as a candidate gene for docosanoic acid abundance. This is a long-chain fatty alcohol oxidase involved in the omega-oxidation pathway of lipid degradation [43]. For minor FAs, we identified a large LD block on chromosome

14 containing the associations with docosanoic and noncosanoic acids, in line with the computational predictions of [30].

Conclusion

This is the first such a large-scale study on Russian sunflower germplasm, which made a significant contribution to sunflower development as the oilseed crop worldwide. Comparison of the Russian sunflower lines with the data on the cultivated and wild sunflower published by Hübner et al. 2019 [38] showed that Russian sunflower germplasm contains unique variation which is not presented in international collections.

Due to climate change, sunflower can become the leading plant in oil production because of its ability to grow under different environmental conditions [44]. In this view, sunflower varieties with oil properties customized for specific applications may become in-demand in the future. Our study makes a step in this direction by identifying the genetic associations both for major and for minor FAs represented in sunflower oil. Genetic markers for minor FAs, such as docosanoic and noncosanoic acids, have not been previously studied. We hope that future sunflower breeding programs will benefit from understanding the genetic bases governing the proportions of these oil components important for industrial applications.

Experimental Procedures

Samples

Sunflower samples were provided by N.I. Vavilov Institute of Plant Genetic Resources (VIR, St. Petersburg, Russia), V.S. Pustovoit All-Russian Research Institute of Oilseed Crops (VNIIMK), and Agroplasma Seed and Breeding Company (Krasnodar, Russia). Samples are accessible upon request.

292 (255 were sequenced) inbred lines from N.I. Vavilov Institute of Plant Genetic Resources (VIR, St. Petersburg, Russia) are mostly conventional lines in terms of fatty acid composition (18:2 range from 36–79%). 3 middle – oleic (18:1 > 50%), 1 high-oleic line (18:1 > 80%).

199 inbred lines from V.S. Pustovoit All-Russian Research Institute of Oilseed Crops (VNIIMK) (Krasnodar, Russia). Fatty acid composition is known for 99 lines: 2 lines high-oleic (18:1 > 80%), 7 middle – oleic (18:1 > 50%). Other lines with 18:2 range between 36–70%).

147 oil-producing sunflower lines provided by Agroplasma Seed and Breeding Company (Krasnodar, Russia). The fatty acid composition is unknown.

All seeds were grown and collected in the Krasnodar region in Russia. For UPLC-MS analysis the seeds themselves were used. For DNA extraction seeds were germinated in the lab. All the reagents can be found in Methods S1

Plants were grown in field in the middle part of the Krasnodar Region.

Soils of the leached black earth soil type. Sunflower was sowed following the preceding crop, fall wheat, at the seeding rate of 40,000 plants per hectare.

Sowing was carried out according to the following sowing system: 70 × 35 cm, a single plant per planting pit. Farming techniques, as commonly used for sunflower.

Each line was grown on the plot with an area of 9.1 m^2.

Details on the dataset can be found in Appendix S1.

DNA Extraction

DNA was extracted from chlorophyll-free sprouts after 1 week of germination without light. 100 mg of tissue for each sample was grounded to powder using FastPrep-96™ Automated Homogenizer (MP Biomedicals). Total DNA was extracted according to the CTAB protocol using the NucleoSpin® Plant II plant DNA extraction kit (Macherey-Nagel, Germany) and stored at -20 ˚C until needed. The purified DNA samples' quality and concentration were determined by gel electrophoresis and in the Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA).

GBS Library Preparation

Illumina libraries were constructed using two restriction endonucleases – HindIII (rarely cutting enzyme, A/AGCTT) and NlaIII (frequently cutting, CATG/) according to the protocol described by [45] with minor modifications. Details of the method are provided in Methods S2.

GBS Sequencing And Primary Data Analysis

Each 96-multiplexed library was sequenced across three lanes in Illumina HiSeq 4000 (San Diego, CA, USA) at the Skoltech Genomics Core Facility as either 150 bp or 75 bp paired-end reads. The sequencing dataset can be found in NCBI repository: https://www.ncbi.nlm.nih.gov/bioproject/620114 [46]. Illumina reads were mapped onto the *Helianthus annuus* reference genome HanXRQr1.0 [47] using BWA MEM 0.7.9a-r786[48] with consideration for uniquely mapped reads whose PE ends mapped within 1K of each other. Variants were called using the GATK pipeline, which considers indel realignment and base quality score recalibration and calls variants across all samples simultaneously through the HaplotypeCaller program in GATK. Variants were filtered using hard filtering parameters: MQ > 36, QD > 24, and MQRankSum < 2, ensuring that the reads were mapped to a unique place in the reference with high quality (MQ), that the reads carrying both alleles were comparable in terms of mapping quality (MQRankSum), and that the actual variants were called with high quality (QD), filters that were not applied by default by GATK's HaplotypeCaller, resulting in the 2.3M SNP calls. To retain SNPs for population and GWAS analyses for oleic and linoleic acids missing calls rate < 0.3, DP > 4, MAF > 0.01 were applied, resulting in 15068 SNPs, for GWAS for other fatty acids we used more strict MAF > 0.03 resulting in 12528 SNPs.

VCF file with SNP variants provided in supporting information (missing calls rate < 0.1, DP > 4, MAF > 0.01).

Validation of SNP calls, was performed using the MALDI-TOF MS technology (Agena Bioscience's MALDI-TOF-based scalable MassARRAY). We selected 75 lines and re-genotyped 11 SNPs, which were significantly associated with at least one studied trait and with MAF above 0.05 among the 75 selected lines (Table S8). For six SNPs, genotypes identified by two technologies were completely identical across all lines studied. For the remaining five lines, the proportion of lines with identical genotypes varied from 0.91 to 0.95. For eight SNPs, the agreement between methods was statistically significant (Permutation test, BH-corrected $p < 0.05$).

Population Structure

Genetic diversity among analyzed lines was estimated using PCA with the aid of PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) [49] based on 15068 SNPs with minor allele frequency (MAF) > 0.01 called on all 17 chromosomes. Population structure was analyzed using ADMIXTURE v1.3.0 [50] with the number of clusters varying from 1 to 10.

Linkage Disequilibrium

LD was estimated across the sunflower genome using VCFtools [51] to calculate frequency correlation (r^2) between 25431 biallelic SNPs with MAF > 0.03 whose genotypes were supported by at least 4 reads called in at least 60% of individuals.

Lipid Extraction

For lipid extraction, 10 mg (for each line) of sunflower seeds (1 sample-1 seed) with 400 uL of methanol/methyl tert-butyl ether mixture (1:3 v:v) were homogenized in Precellys evolution (Bertin corp. USA) (6800 rpm, 3* 20 sec, pause 30 sec) coupled with Cryolis filled with dry ice with 6 2.8 mm zirconium oxide beads (Bertin corp. USA) at the temperature not higher than 10 degrees. Then, extraction was performed using methanol/methyl tert-butyl ether mixture, according to [52] with minor modifications (Methods S3).

For FAs analysis, the extracts obtained in the previous steps were hydrolyzed using the protocol adopted from [53] (Methods S3).

UPLC-MS Profiling

Samples were processed using mass spectrometry (UPLC-MS) coupled with reversed-phase ultra-performance liquid chromatography (ACQUITY UPLC System; Waters, USA) in positive and negative ionization modes in Q-TOF Maxis Impact II, Bruker Daltonik, Germany. Settings: Ion Polarity: positive/negative, Scan mode: MS, Mass range: 50 -1200m/z, Spectra rate: 2 Hz.

UPLC separation was performed on the C8 Acquity Beh column (2.1 mm X 100 mm, 1.7-µm particle size; Waters) and the Acquity BEH C8 1.7 µm Vanguard precolumn (Waters) at 60 °C.

The detailed information can be found in Methods S3.

Previously we have validated the extraction and profiling technique for Sunflower FAs [54] and TAGs [55].

Lipidomic Data Analysis And Annotation

For data processing, optimal parameters were generated using the Bioconductor IPO package. The subsequent peak peaking, chromatogram alignment, chemical noise subtraction and intensity thresholding were performed using the XCMS 3.1 package (https://bioconductor.riken.jp/packages/3.1/bioc/html/xcms.html) [56]. The output was a list of peaks, with retention time, m/z, and intensity for each sample. To exclude possible contaminants, mean intensities of all sunflower peaks were compared to mean intensities in blank samples. (Figure S12). Only lipids with sample intensity at least two times higher than blank intensities were used in the analysis.

To annotate FAs and TAGs, formulas for the possible lipids (irrespective to isomers) of these classes were generated. For FAs, chain lengths from C10:0 to C28:0 with not more than 6 double bonds were considered. For TAGs, the total chain length varied between 30 and 85 carbon atoms, and the number of double bonds varied from 0 to 12. Then, masses of generated lipids were compared to m/z of detected peaks. For FAs just one adduct ($-H^+$) was considered; for TAGs four adducts (H^+, Na^+, K^+, and NH_4^+) were considered. All peaks with ppm ($ppm = abs(m1 - m2)/max(m1, m2) \times 10^6$), where m1 and m2 are masses of lipid and m/z of the peak, respectively) below 10 were considered as the possible lipids of the given class. Then, for each of the two lipid classes and for each adduct, the peaks were manually filtered based on the expectation that correct FAs and TAGs should form a net-like pattern on RT-m/z scatter plot. To annotate non-TAG lipids measured in the positive mode lipid Befdatabase (The LIPID MAPS® Lipidomics Gateway, https://www.lipidmaps.org/) was used. First, all isomers were collapsed, then m/z of all non-TAG peaks were compared with the masses of all lipids from the lipidmap. Same adducts as were used for TAGs were considered. All lipid-peak pairs with ppm < 10 were considered as a valid annotation. All peaks annotated with lipids of just one category were assigned to this category; peaks annotated with lipids of more than one category were considered as ambiguously annotated.

Reproducibility experiments with three years of replicates and the main dataset were measured and processed separately.

In the reproducibility experiment, only TAGs with NH_4^+ adduct and FAs were considered. For both dilutions of TAGs and for FAs, the intensity of individual lipids was divided by the total intensity of all TAGs/FAs in the given sample and multiplied by 100. To assess the role of genetic and environmental factors ANOVA with the following model was used:

Lipid_concentration ~ line + year + line:year.

MDS analysis for two dimensions was performed based on one minus Spearman correlation coefficient distance.

Association Analysis And Annotation

Since the sunflower genotypes were from three different seed banks, it was important to account for their genetic relatedness and hence the mixed linear model was implemented (MLM: Y = SNP + PCs + Kinship + e, where Y – phenotype, SNP and PCs – fixed effects, Kinship – random effect, e - error). In addition, internal standards intensity and LS-MS batch numbers were used as co-factors to account for the batch effect and sample weight in the model.

Before GWAS, FAs distribution between the samples were estimated (Figure S13). For GWAS, all the samples with 10% and more missing data (for phenotypes) were excluded from the analysis. GWAS was performed using TASSEL 5 [57]. SNPs for the analysis were filtered out using the following criteria: missing calls rate < 0.3, DP < 4, and minor allele frequency (MAF) < 0.01 for such traits as oleic and linoleic acids and MAF < 0.03 for other traits. Filtering was performed using VCFtools. A mixed linear model was used where the SNP effect and population structure estimated by PCA were treated as fixed effects and kinship was included in the model as a random effect. The genetic relatedness analysis was performed with the relative kinship coefficients (K-matrix) calculation using the TASSEL software (Centered IBSmethod). The collection and the batch number were also used as factors and sample weight and internal standards intensity as covariates. To estimate the mixed linear model performance, quantile-quantile plots (q-q plots) were used. Observed p-values were plotted against the expected probability of their distribution. To represent GWAS results, Manhattan plots were used where p-values were plotted for all sunflower linkage groups one by one. GWAS results were visualized with the help of the qqman R package (version 0.1.4).

To determine the significance of observed hits, 0.05/5000 p-value threshold was used. This is a a Bonferroni correction based on average number of LD blocks. The total number of SNPs used in GWAS was divided by 5000 - the number of LD blocks estimated from LD analysis. LD block analysis was performed using Haploview software [58]. Gene annotation within each LD block was performed using the sunflower genome browser (https://sunflowergenome.org).

To estimate the variance in docosanoic concentration explained by identified 53 SNPs, the linear model with same covariates as was used in GWAS analysis and with all 53 SNPs was used. Results show that these SNPs could explain 35.4% of the variance.

Abbreviations

GBS Genotyping-by-sequencing

TAG Triacylglyceride

FA Fatty acid

UPLC -MS Ultra-performance liquid chromatography coupled with Mass-Spectrometry

GWAS Genome-wide association study

QTL Quantitative trait locus

VIR N.I.Vavilov Research Institute of Plant Industry

VNIIMK Pustovoit All-Russia Research Institute of Oil Crops

PCA Principal component analysis

LPA Lysophosphatidic acid

PC Phosphatidylcholine

PE Phosphatidylethanolamine

PI Phosphatidylinositol

PA Phosphatidic acid

LD Linkage disequilibrium

MLM Mixed linear model

MAF Minor allele frequency

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable

Availability of data and materials

The authors confirm that the data supporting this study's findings are available within the article, its supplementary materials, and public available sources. Sequencing dataset can be found in NCBI repository: https://www.ncbi.nlm.nih.gov/bioproject/620114

Competing interests

The authors declare no conflict of interest.

Funding

This study was carried out using the resources of the Skoltech Genomics Core Facility. The funders had no role in study design, data collection, and analysis, decision to publish, or manuscript preparation. This work was supported by the Ministry of Science and Higher Education of the Russian Federation, Grant Number: 14.609.21.0099, Identification No. RFMEFI60916X0099

Author Contributions

Concept and design: A.C, S.G, S.N, P.K; Lipidomic profiling methodology: A.V, E.S; Genotyping methodology: S.G, S.N, E.M; Seed samples preparation: E.Y,A.C, S.B; Lipid extraction: N.A, A.C; UPLC-MS analysis: A.C, E.S; Plant growing: A.C, S.B; DNA extraction: A.C, E.M, Y.K; GBS library preparation: A.C, Y. K; Analysis and interpretation of genomic data: P.C, A.S, A.C, R.G, P.M, D.G; Analysis and interpretation of lipidomic data: P.M, K.S, A.C, R.G; Figure construction: P.M, R.G, A.C; Wrote the manuscript drafting: A.C, P.C, A.S, K.S. P.M, R.G, A.C, S.N, P.K; Genetic resources: Y.D, Z.M, V.G, I.A; Supervision: S.N, P.K.

All the authors read and approved the manuscript

Acknowledgments

We want to extend our sincere gratitude to Loren Rieseberg, Jean-Sébastien Légaré, and Gregory Owens, who had shared the data with us. Also, we want to mention Lauren McIntyre, who helped us with useful advice according MS data processing.

References

1. Crites GD. Domesticated Sunflower in Fifth Millennium B.P. Temporal Context: New Evidence from Middle Tennessee. Am Antiq. 1993;58:146–8.
2. Burke JM, Tang S, Knapp SJ, Rieseberg LH. Genetic analysis of sunflower domestication. Genetics. 2002;161:1257–67.
3. Martínez Force E, editor. Sunflower: chemistry, production, processing, and utilization. Urbana, Illinois: AOCS Press; 2015.
4. Friedt W. Present state and future prospects of biotechnology in sunflower breeding. Field Crops Res. 1992;30:425–42.

5. Seiler GJ, Rieseberg LH. Systematics, Origin, and Germplasm Resources of the Wild and Domesticated Sunflower. Sunflower Technol Prod. 1997; 21–65.
6. Terzić S, Boniface M-C, Marek L, Alvarez D, Baumann K, Gavrilova V, et al. Gene banks for wild and cultivated sunflower genetic resources. OCL. 2020;27:9.
7. Dimitrijevic A, Horn R. Sunflower Hybrid Breeding: From Markers to Genomic Selection. Front Plant Sci. 2018;8.
8. Rauf S, Jamil N, Tariq SA, Khan M, Kausar M, Kaya Y. Progress in modification of sunflower oil to expand its industrial value. J Sci Food Agric. 2017;97:1997–2006.
9. Konyalı S. Sunflower Production and Agricultural Policies in Turkey. Sos Bilim Araşt Derg. 2017;6:11–19.
10. Dimitrijević A, Imerovski I, Miladinović D, Cvejić S, Jocić S, Zeremski T, et al. Oleic acid variation and marker-assisted detection of Pervenets mutation in high- and low-oleic sunflower cross. Crop Breed Appl Biotechnol. 2017;17:235–41.
11. Velasco L, Ruiz-Méndez MV. Sunflower Oil Minor Constituents. In: Sunflower. Elsevier; 2015. p. 297–329.
12. Venegas-Calerón M, Troncoso-Ponce MA, Martínez-Force E. Sunflower Oil and Lipids Biosynthesis. In: Sunflower. Elsevier; 2015. p. 259–95.
13. Jocic' S, Miladinovic' D, Kaya Y. Breeding and Genetics of Sunflower. In: Sunflower. Elsevier; 2015. p. 1–25.
14. Hummel J, Segu S, Li Y, Irgang S, Jueppner J, Giavalisco P. Ultra performance liquid chromatography and high resolution mass spectrometry for the analysis of plant lipids. Front Plant Sci. 2011;2:54.
15. Li L, Lu X, Zhao J, Zhang J, Zhao Y, Zhao C, et al. Lipidome and metabolome analysis of fresh tobacco leaves in different geographical regions using liquid chromatography–mass spectrometry. Anal Bioanal Chem. 2015;407:5009–20.
16. Cenzano AM, Cantoro R, Teresa Hernandez-Sotomayor SM, Abdala GI, Racagni GE. Lipid profiling by electrospray ionization tandem mass spectrometry and the identification of lipid phosphorylation by kinases in potato stolons. J Agric Food Chem. 2012;60:418–26.
17. Sugawara T, Duan J, Aida K, Tsuduki T, Hirata T. Identification of Glucosylceramides Containing Sphingatrienine in Maize and Rice Using Ion Trap Mass Spectrometry. Lipids. 2010;45:451–5.
18. Li M, Butka E, Wang X. Comprehensive Quantification of Triacylglycerols in Soybean Seeds by Electrospray Ionization Mass Spectrometry with Multiple Neutral Loss Scans. Sci Rep. 2014;4.
19. Boukhchina S, Sebai K, Cherif A, Kallel H, Mayer PM. Identification of glycerophospholipids in rapeseed, olive, almond, and sunflower oils by LC□MS and LC□MS□MS. Can J Chem. 2004;82:1210–5.
20. Gao B, Luo Y, Lu W, Liu J, Zhang Y, Yu L (Lucy). Triacylglycerol compositions of sunflower, corn and soybean oils examined with supercritical CO 2 ultra-performance convergence chromatography combined with quadrupole time-of-flight mass spectrometry. Food Chem. 2017;218:569–74.

21. Leamy LJ, Zhang H, Li C, Chen CY, Song B-H. A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC Genomics. 2017;18:18.
22. Qu C, Jia L, Fu F, Zhao H, Lu K, Wei L, et al. Genome-wide association mapping and Identification of candidate genes for fatty acid composition in Brassica napus L. using SNP markers. BMC Genomics. 2017;18.
23. Bonnafous F, Fievet G, Blanchet N, Boniface M-C, Carrère S, Gouzy J, et al. Comparison of GWAS models to identify non-additive genetic control of flowering time in sunflower hybrids. TAG Theor Appl Genet Theor Angew Genet. 2018;131:319–32.
24. Mandel JR, Nambeesan S, Bowers JE, Marek LF, Ebert D, Rieseberg LH, et al. Association Mapping and the Genomic Consequences of Selection in Sunflower. PLOS Genet. 2013;9:e1003378.
25. Goryunov DV, Anisimova IN, Gavrilova VA, Chernova AI, Sotnikova EA, Martynova EU, et al. Association Mapping of Fertility Restorer Gene for CMS PET1 in Sunflower. Agronomy. 2019;9:49.
26. Masalia RR, Temme AA, Torralba N de L, Burke JM. Multiple genomic regions influence root morphology and seedling growth in cultivated sunflower (Helianthus annuus L.) under well-watered and water-limited conditions. PloS One. 2018;13:e0204279.
27. Mangin B, Casadebaig P, Cadic E, Blanchet N, Boniface M-C, Carrère S, et al. Genetic control of plasticity of oil yield for combined abiotic stresses using a joint approach of crop modelling and genome-wide association. Plant Cell Environ. 2017;40:2276–91.
28. Nambeesan SU, Mandel JR, Bowers JE, Marek LF, Ebert D, Corbi J, et al. Association mapping in sunflower (Helianthus annuus L.) reveals independent control of apical vs. basal branching. BMC Plant Biol. 2015;15:84.
29. Dowell JA, Reynolds EC, Pliakas TP, Mandel JR, Burke JM, Donovan LA, et al. Genome-Wide Association Mapping of Floral Traits in Cultivated Sunflower (Helianthus annuus). J Hered. 2019;110:275–86.
30. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 2017;546:148–52.
31. Al-Chaarani GR, Gentzbittel L, Huang XQ, Sarrafı A. Genotypic variation and identification of QTLs for agronomic traits, using AFLP and SSR markers in RILs of sunflower (Helianthus annuus L.). TAG Theor Appl Genet Theor Angew Genet. 2004;109:1353–60.
32. Pérez-Vich B, Fernández-Martínez JM, Grondona M, Knapp SJ, Berry ST. Stearoyl-ACP and oleoyl-PC desaturase genes cosegregate with quantitative trait loci underlying high stearic and high oleic acid mutant phenotypes in sunflower. Theor Appl Genet. 2002;104:338–49.
33. Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, et al. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci. 2012;109:8872–7.
34. Ebrahimi A, Maury P, Berger M, Kiani SP, Nabipour A, Shariati F, et al. QTL mapping of seed-quality traits in sunflower recombinant inbred lines under different water regimes. Genome. 2008;51:599–615.

35. Mangin B, Bonnafous F, Blanchet N, Boniface M-C, Bret-Mestries E, Carrère S, et al. Genomic Prediction of Sunflower Hybrids Oil Content. Front Plant Sci. 2017;8.
36. Voelker TA, Kinney AJ. Variations in the Biosynthesis of Seed-storage Lipids. Annu Rev Plant Physiol Plant Mol Biol. 2001;52:335–61.
37. Gavrilova VA, Rozhkova VT, Anisimova IN. Sunflower Genetic Collection at the Vavilov Institute of Plant Industry. Helia. 2014;37:1–16.
38. Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat Plants. 2019;5:54–62.
39. Radanović A, Miladinović D, Cvejić S, Jocković M, Jocić S. Sunflower Genetics from Ancestors to Modern Hybrids—A Review. Genes. 2018;9.
40. Salas JJ, Martínez-Force E, Garcés R. Very Long Chain Fatty Acid Synthesis in Sunflower Kernels. J Agric Food Chem. 2005;53:2710–6.
41. Martínez-Rivas JM, Sperling P, Lühs W, Heinz E. Spatial and temporal regulation of three different microsomal oleate desaturase genes (FAD2) from normal-type and high-oleic varieties of sunflower (Helianthus annuus L.). Mol Breed. 2001;8:159–68.
42. Premnath A, Narayana M, Ramakrishnan C, Kuppusamy S, Chockalingam V. Mapping quantitative trait loci controlling oil content, oleic acid and linoleic acid content in sunflower (Helianthus annuus L.). Mol Breed. 2016;36.
43. Vanhanen S, West M, Kroon JTM, Lindner N, Casey J, Cheng Q, et al. A Consensus Sequence for Long-chain Fatty-acid Alcohol Oxidases from Candida Identifies a Family of Genes Involved in Lipid ω-Oxidation in Yeast with Homologues in Plants and Bacteria. J Biol Chem. 2000;275:4445–52.
44. Miladinović D, Hladni N, Radanović A, Jocić S, Cvejić S. Sunflower and Climate Change: Possibilities of Adaptation Through Breeding and Genomic Selection. In: Kole C, editor. Genomic Designing of Climate-Smart Oilseed Crops. Cham: Springer International Publishing; 2019. p. 173–238.
45. Zhigunov AV, Ulianich PS, Lebedeva MV, Chang PL, Nuzhdin SV, Potokina EK. Development of F1 hybrid population and the high-density linkage map for European aspen (Populus tremula L.) using RADseq technology. BMC Plant Biol. 2017;17:180.
46. Helianthus annuus (ID 620114) - BioProject - NCBI. https://www.ncbi.nlm.nih.gov/bioproject/620114.
47. HanXRQr1.0 - Genome - Assembly - NCBI. https://www.ncbi.nlm.nih.gov/assembly/GCF_002127325.1.
48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009;25:1754–60.
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

50. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
51. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
52. Giavalisco P, Li Y, Matthes A, Eckhardt A, Hubberten H-M, Hesse H, et al. Elemental formula annotation of polar and lipophilic metabolites using 13C, 15N and 34S isotope labelling, in combination with high-resolution mass spectrometry. Plant J. 2011;68.
53. Bromke MA, Hochmuth A, Tohge T, Fernie AR, Giavalisco P, Burgos A, et al. Liquid chromatography high-resolution mass spectrometry for fatty acid profiling. Plant J. 2015;81:529–36.
54. Chernova A, Mazin P, Goryunova S, Goryunov D, Demurin Y, Gorlova L, et al. Ultra-performance liquid chromatography-mass spectrometry for precise fatty acid profiling of oilseed crops. PeerJ. 2019;7:e6547.
55. Chernova A, Gubaev R, Mazin P, Goryunova S, Demurin Y, Gorlova L, et al. UPLC⁻MS Triglyceride Profiling in Sunflower and Rapeseed Seeds. Biomolecules. 2018;9.
56. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78:779–87.
57. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
58. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinforma Oxf Engl. 2005;21:263–5.

Figures

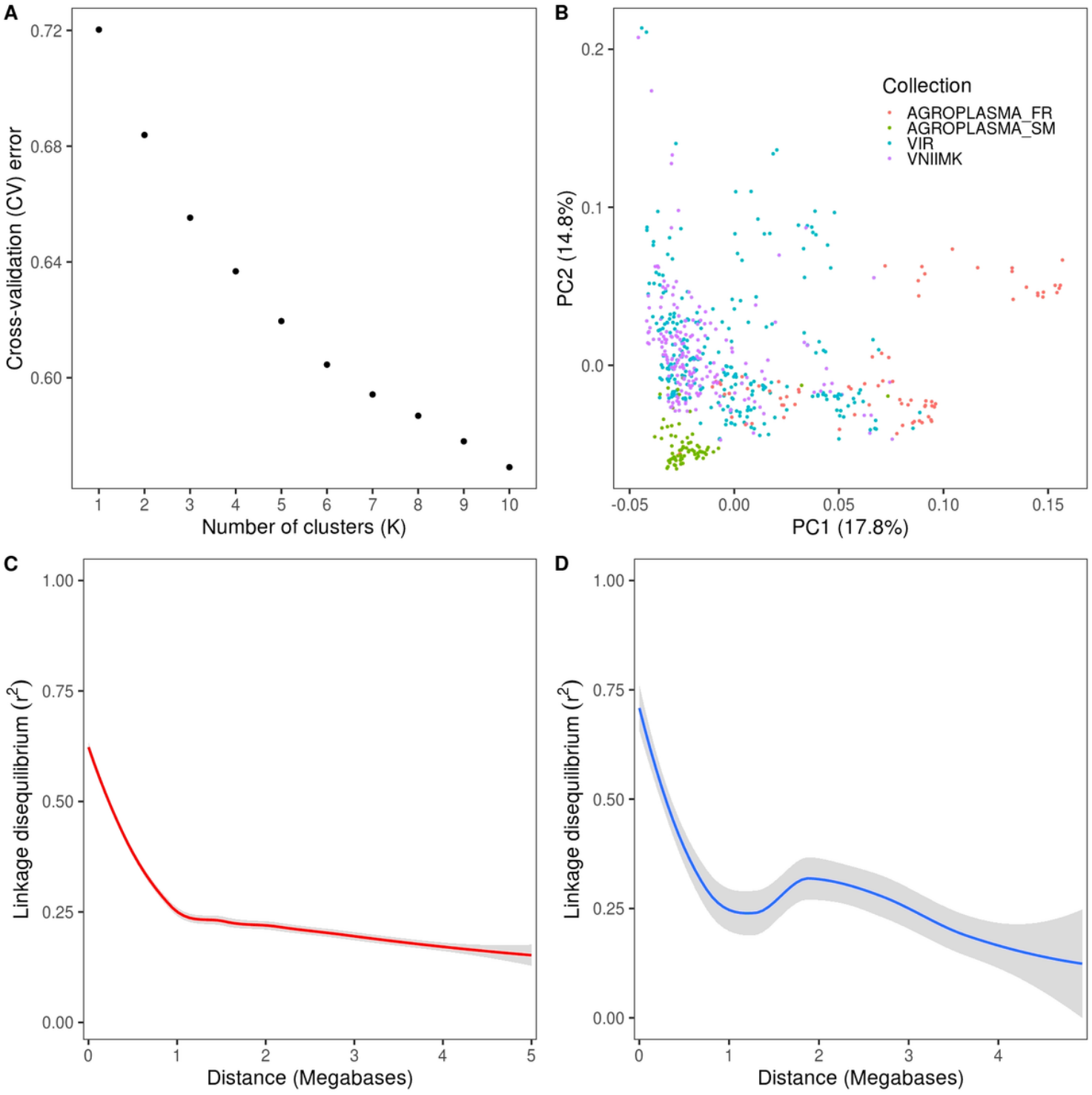

Figure 1

Population structure of germplasm and linkage disequilibrium (LD) values. (A) Estimated cross-validation error value for possible cluster number from 1 to 10. (B) Subpopulations were assessed using Principal Component Analysis. Each dot corresponds to a sunflower accession used in the study. Color corresponds to sunflower lines from different collections. Agroplasma_SM indicates sterility maintaining lines from Agroplasma; Agroplasma_FR indicates fertility restorer lines. (C-D) Genome-wide (C) and per-chromosome 3 (D) LD-decay. Lines correspond to loess curves.

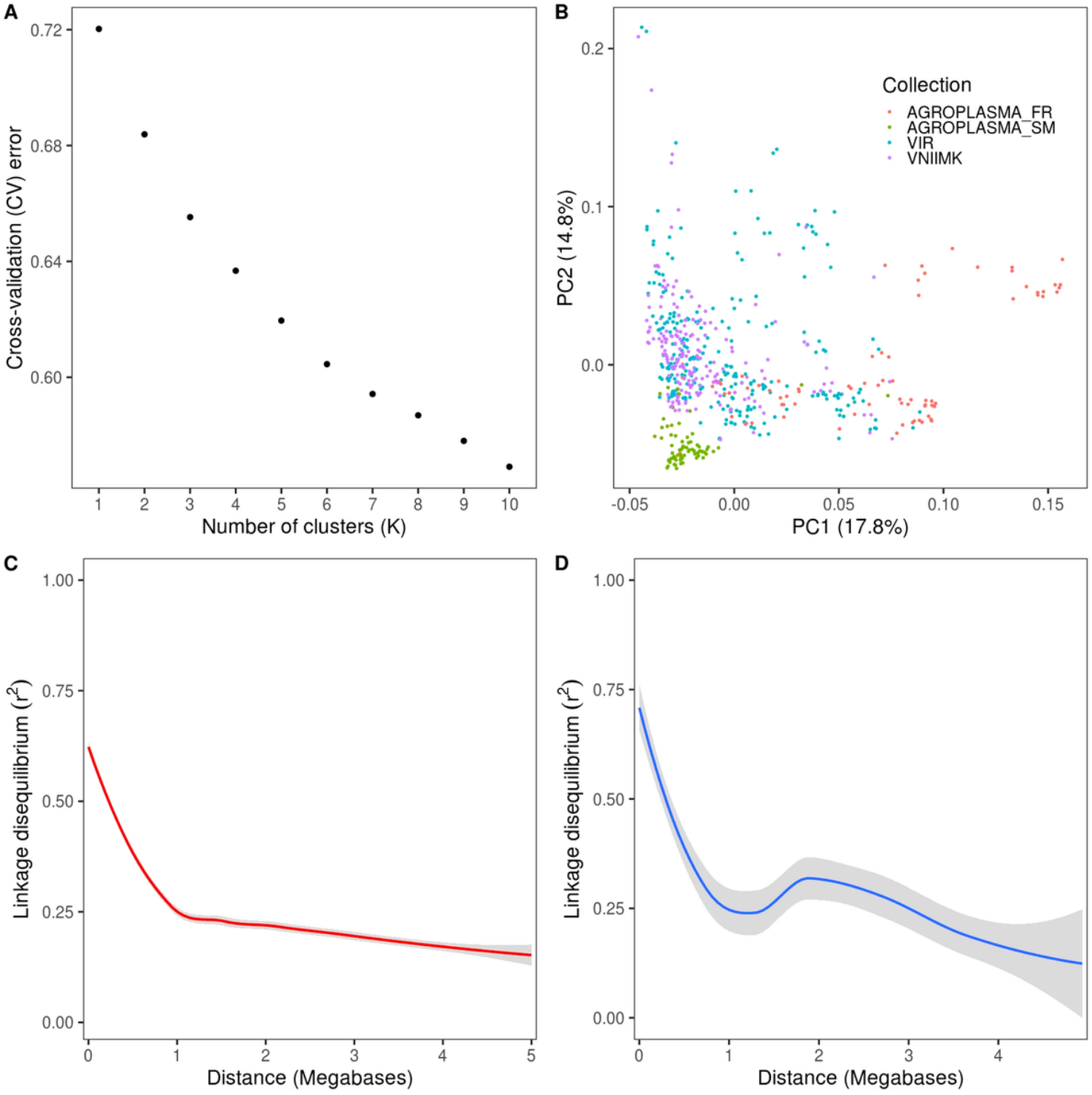

Figure 1

Population structure of germplasm and linkage disequilibrium (LD) values. (A) Estimated cross-validation error value for possible cluster number from 1 to 10. (B) Subpopulations were assessed using Principal Component Analysis. Each dot corresponds to a sunflower accession used in the study. Color corresponds to sunflower lines from different collections. Agroplasma_SM indicates sterility maintaining lines from Agroplasma; Agroplasma_FR indicates fertility restorer lines. (C-D) Genome-wide (C) and per-chromosome 3 (D) LD-decay. Lines correspond to loess curves.

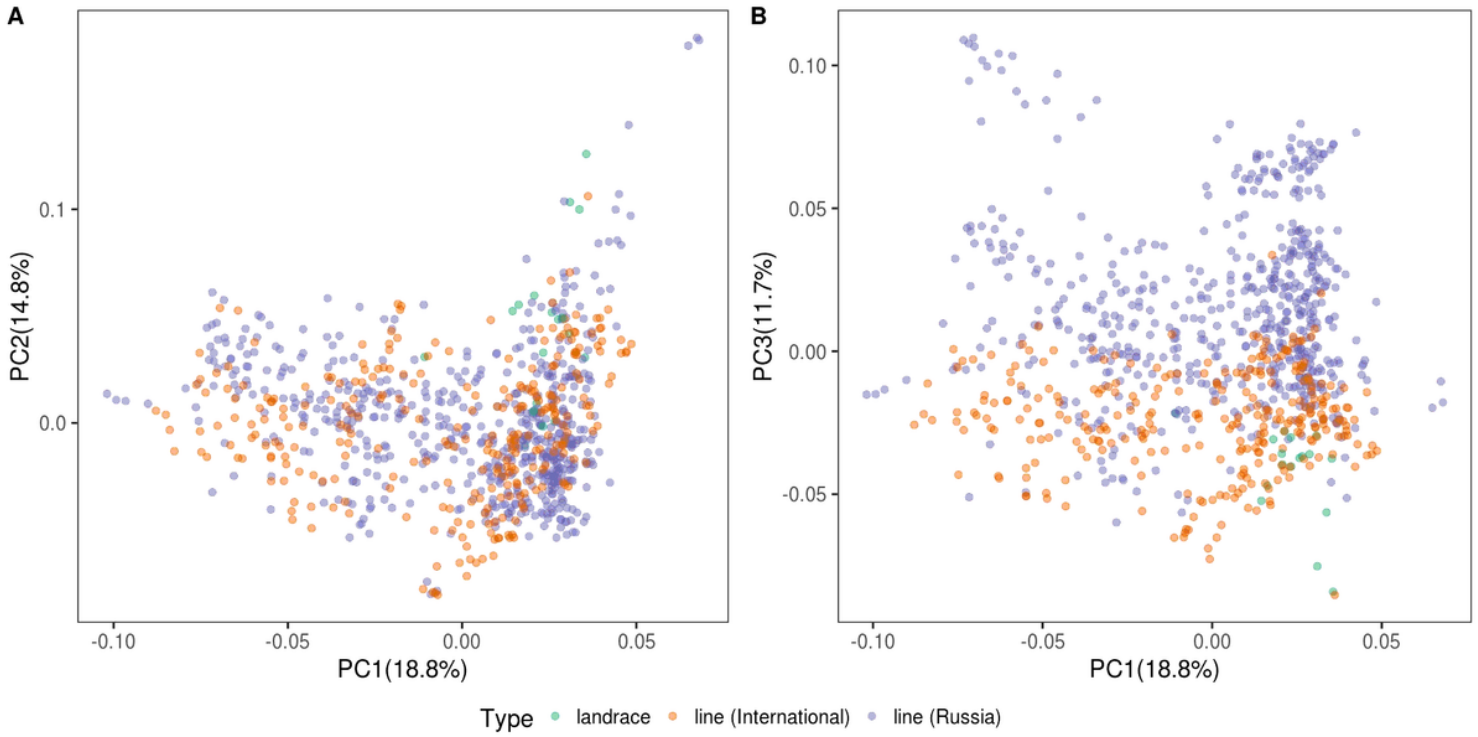

Figure 2

The relationship between sunflower germplasm of different origins estimated based on 2345 SNPs shared between this and the Hübner (2019) studies. (A) The first and the second components of the PCA. (B) The first and the third components of the PCA. Each dot corresponds to a plant accession. Colors indicate the origin.

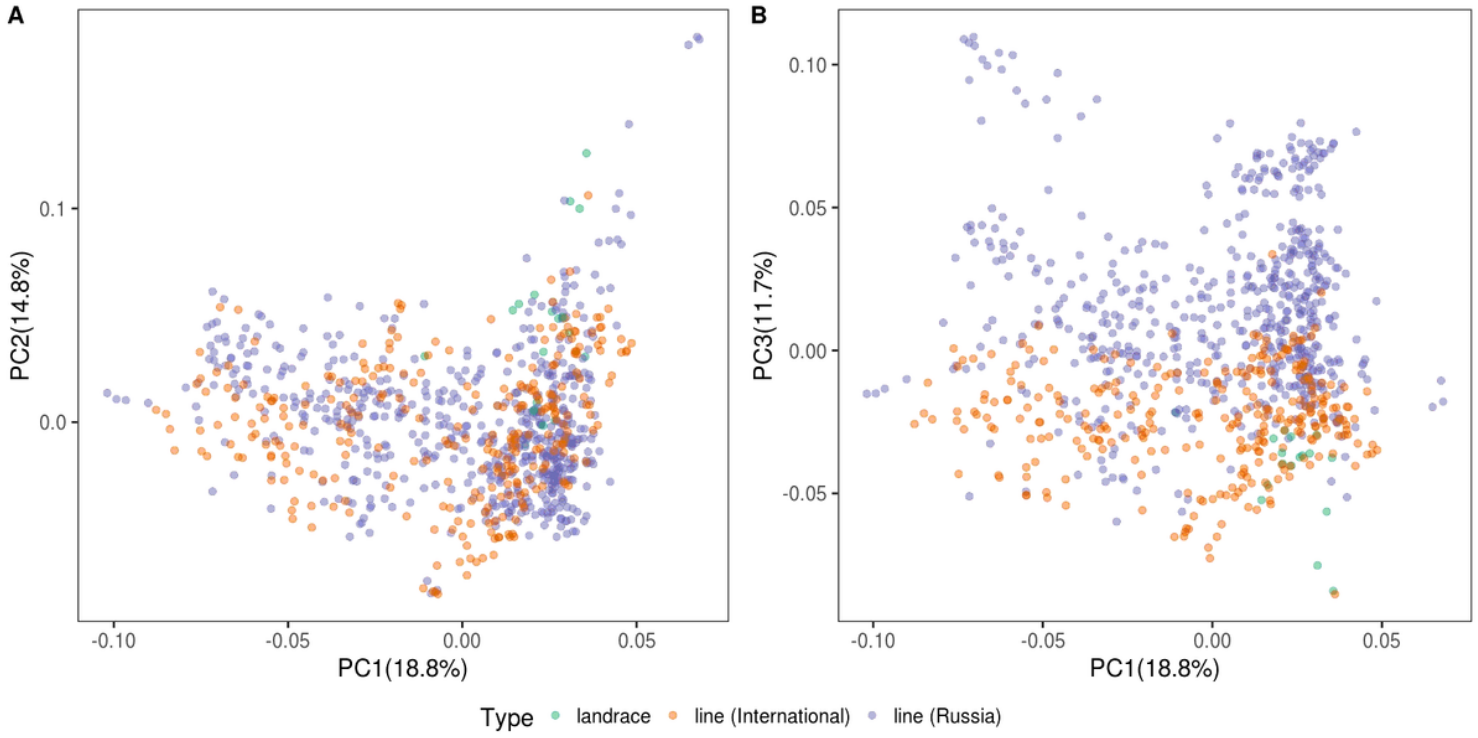

Figure 2

The relationship between sunflower germplasm of different origins estimated based on 2345 SNPs shared between this and the Hübner (2019) studies. (A) The first and the second components of the PCA. (B) The first and the third components of the PCA. Each dot corresponds to a plant accession. Colors indicate the origin.

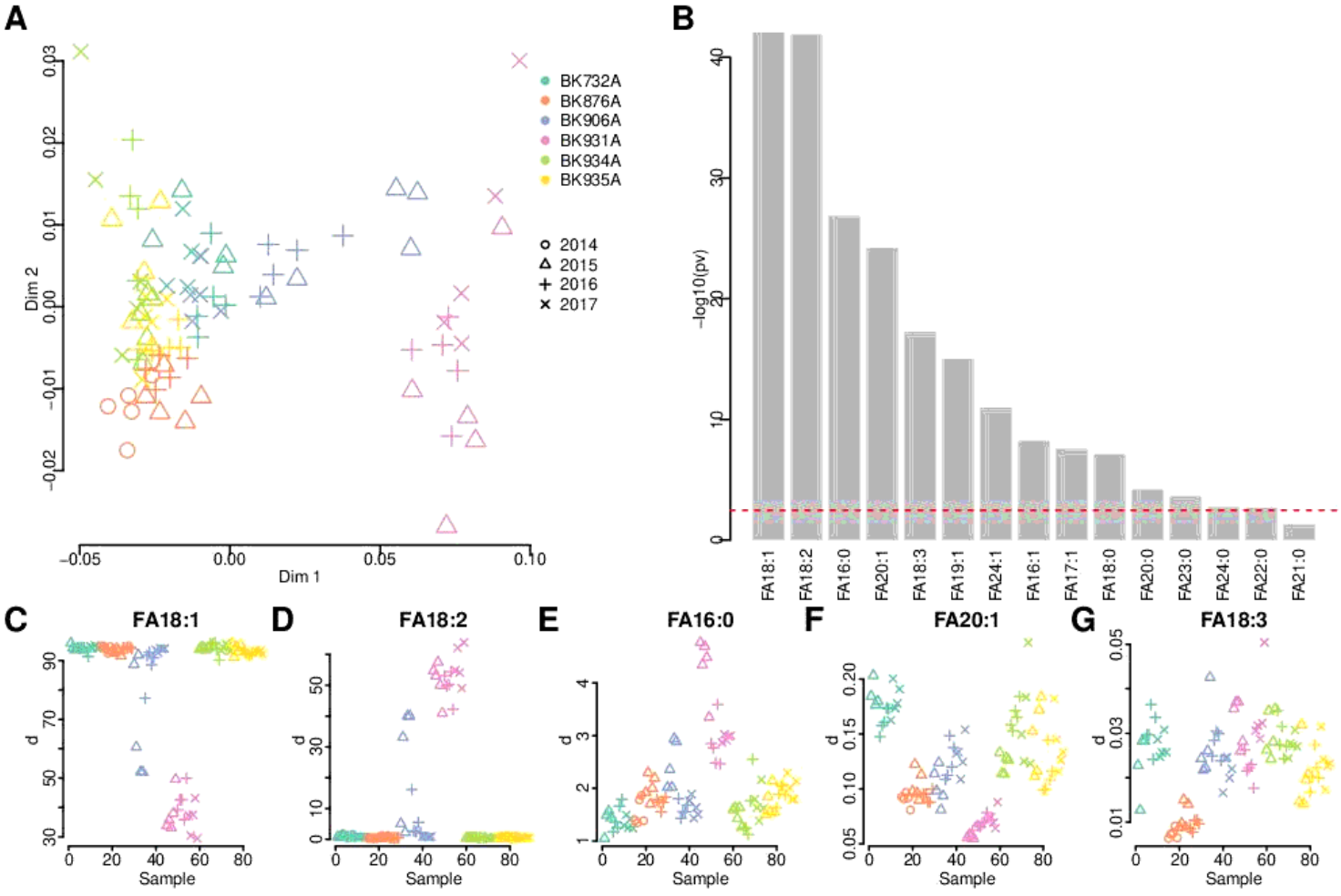

Figure 3

FAs concentrations in replication experiments. (A) MDS plot (two dimensions, 1 - Spearman correlation coefficient between FAs concentrations was used as distance). One sample is shown by one point; lines are shown by different colors; different years are shown by points of different shapes. (B) Minus log10 p-values for the differences between lines (ANOVA) are shown. Bonferroni adjusted 0.05 significance level is shown by red line; (C) Linoleic acid (18:2); (D) Oleic acid (18:1); (E) Palmitic acid (16:0), (F) Eicosenoic acid (20:1), (G) Linolenic acid (18:3). Each point represents 1 sample, point shapes, and colors as in (A).

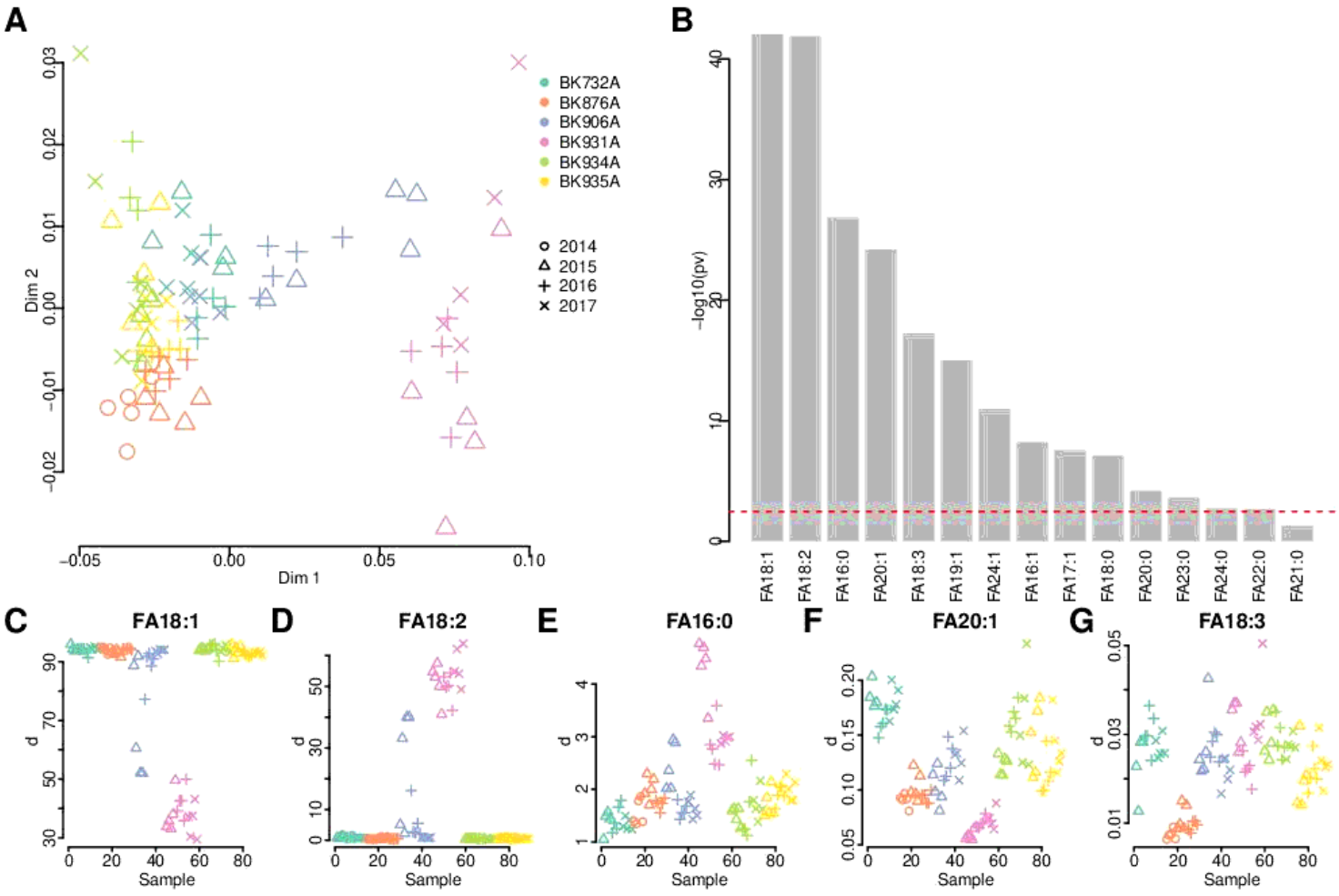

Figure 3

FAs concentrations in replication experiments. (A) MDS plot (two dimensions, 1 - Spearman correlation coefficient between FAs concentrations was used as distance). One sample is shown by one point; lines are shown by different colors; different years are shown by points of different shapes. (B) Minus log10 p-values for the differences between lines (ANOVA) are shown. Bonferroni adjusted 0.05 significance level is shown by red line; (C) Linoleic acid (18:2); (D) Oleic acid (18:1); (E) Palmitic acid (16:0), (F) Eicosenoic acid (20:1), (G) Linolenic acid (18:3). Each point represents 1 sample, point shapes, and colors as in (A).

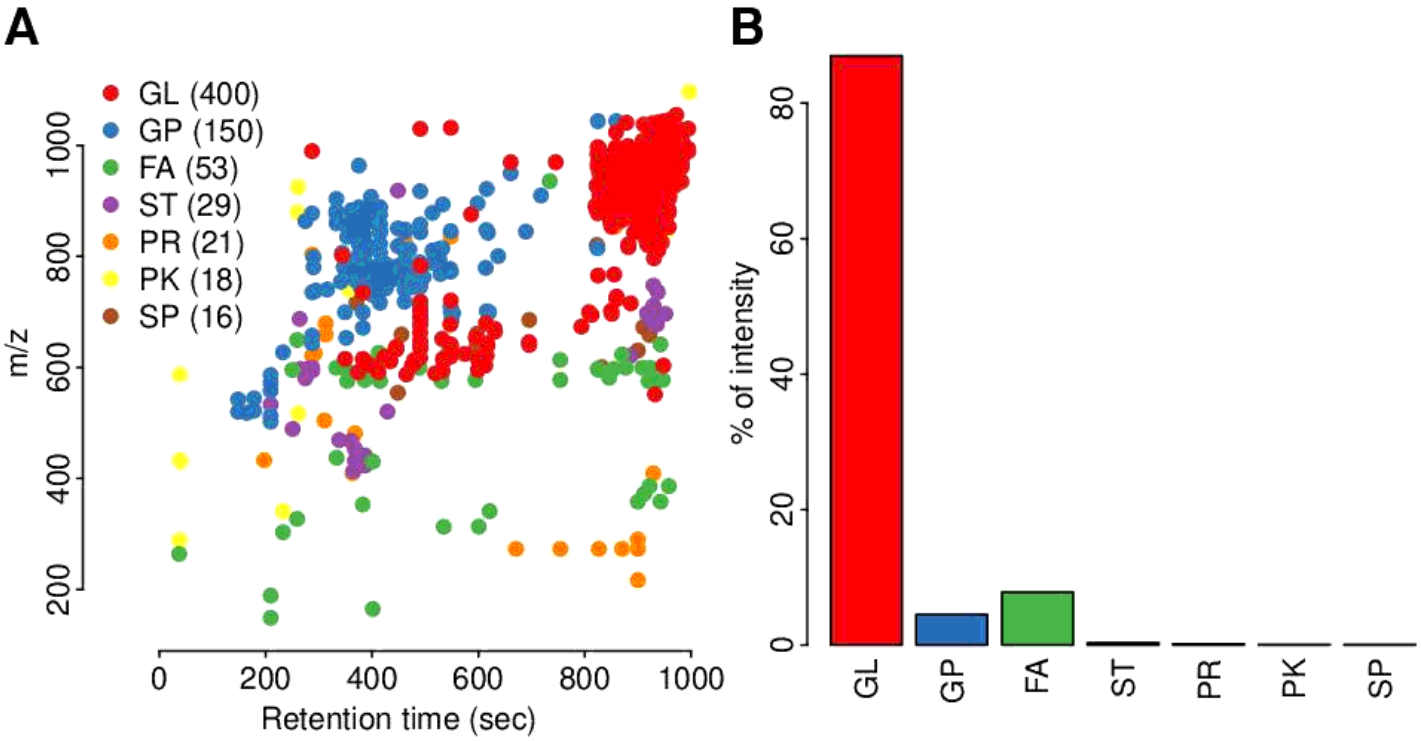

Figure 4

Lipid annotation (A) Mz/rt plot. One point represents one peak; different lipid categories are shown in different colors. Only peaks with sample intensities at least two times higher than blank intensities are shown. (B) Relative intensities of all lipid categories. The intensity of the given category was calculated as the sum of intensities of all lipids of the category. GL- glycerolipids, GP- glycophospholipids , FA -fatty acids, ST- sterols, PR- prenols, PK- polyketides and SP-saccharolipids.

A
GL (400)
GP (150)
FA (53)
ST (29)
PR (21)
PK (18)
SP (16)
m/z
200
400
600
800
1000
0
200
400
600
800
1000
Retention time (sec)
B
% of intensity
0
20
40
60
80
GL
GP
FA
ST
PR
PK
SP

Figure 4

Lipid annotation (A) Mz/rt plot. One point represents one peak; different lipid categories are shown in different colors. Only peaks with sample intensities at least two times higher than blank intensities are shown. (B) Relative intensities of all lipid categories. The intensity of the given category was calculated as the sum of intensities of all lipids of the category. GL- glycerolipids, GP- glycophospholipids , FA -fatty acids, ST- sterols, PR- prenols, PK- polyketides and SP-saccharolipids.

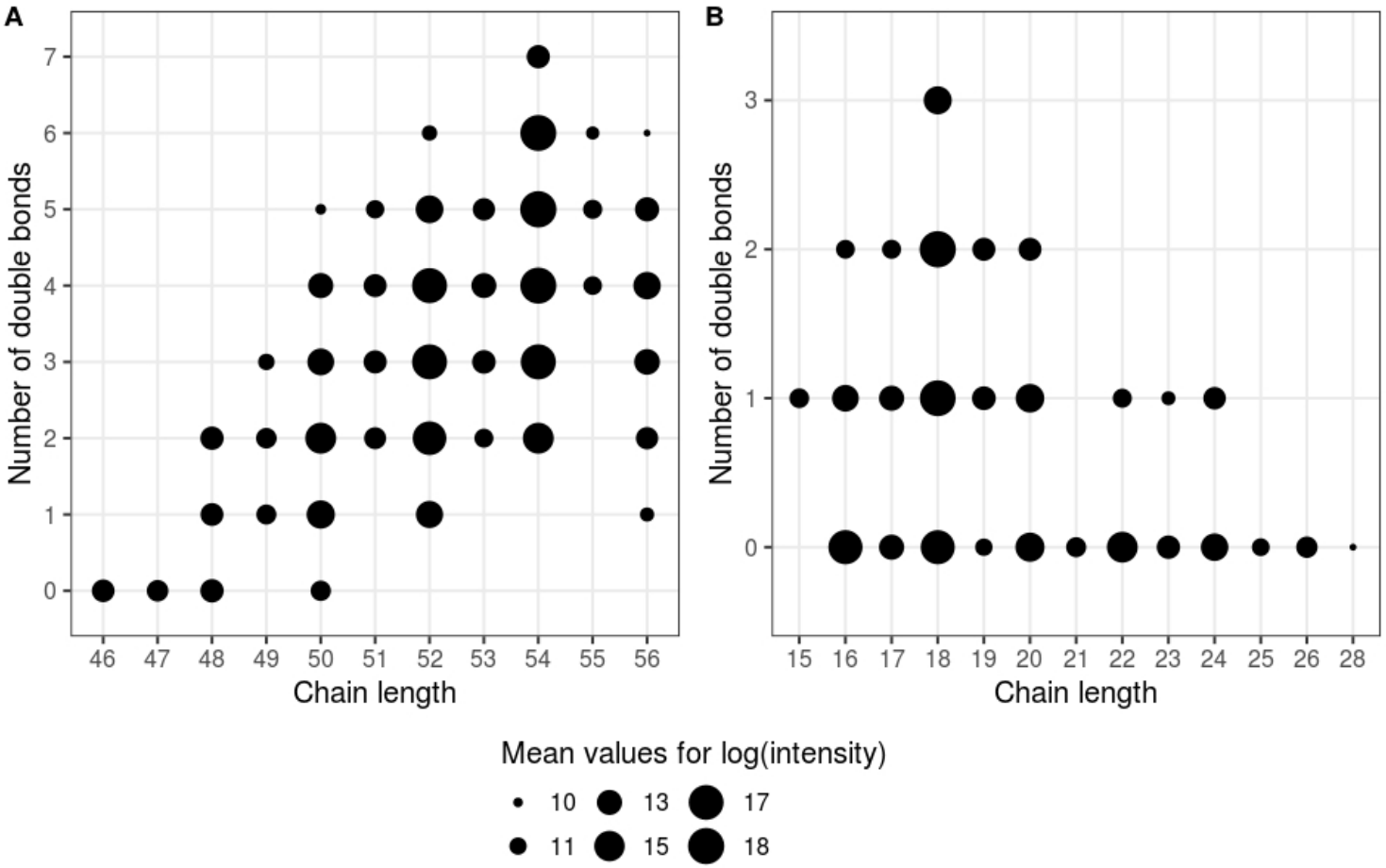

Figure 5

Schematic representation of fatty acid properties (fatty acid chain length and degree of saturation) for detected lipids. (A) Cumulative chain length and double bonds number of three fatty acid residues composing detected TAG molecules. (B) Chain length and double bonds number of fatty acid (FAs) released after lipid hydrolysis. Each circle corresponds to a FA or a TAG. The circles' size corresponds to the mean relative amount of this molecule in a sample (log-transformed MS peak intensity).

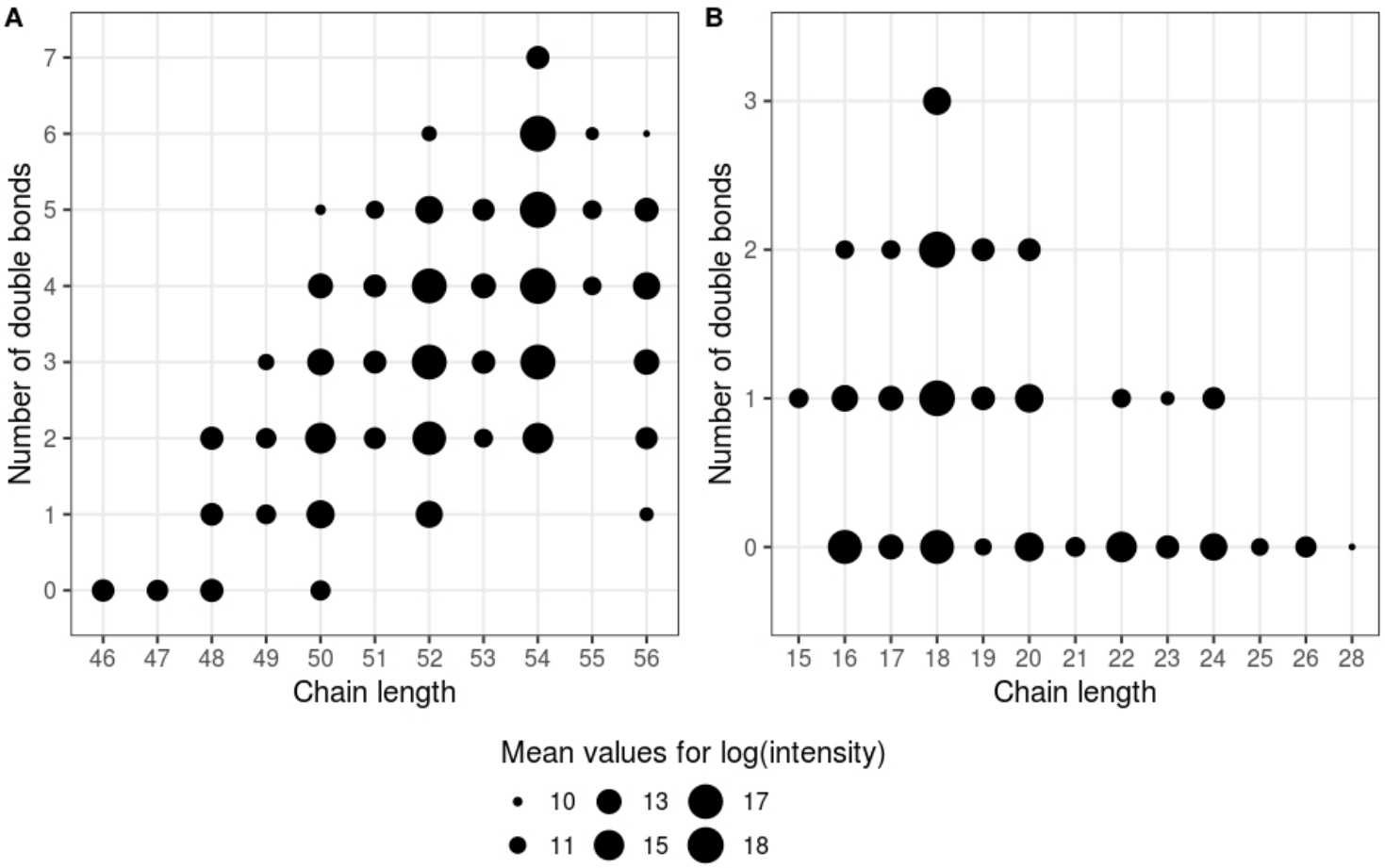

Figure 5

Schematic representation of fatty acid properties (fatty acid chain length and degree of saturation) for detected lipids. (A) Cumulative chain length and double bonds number of three fatty acid residues composing detected TAG molecules. (B) Chain length and double bonds number of fatty acid (FAs) released after lipid hydrolysis. Each circle corresponds to a FA or a TAG. The circles' size corresponds to the mean relative amount of this molecule in a sample (log-transformed MS peak intensity).

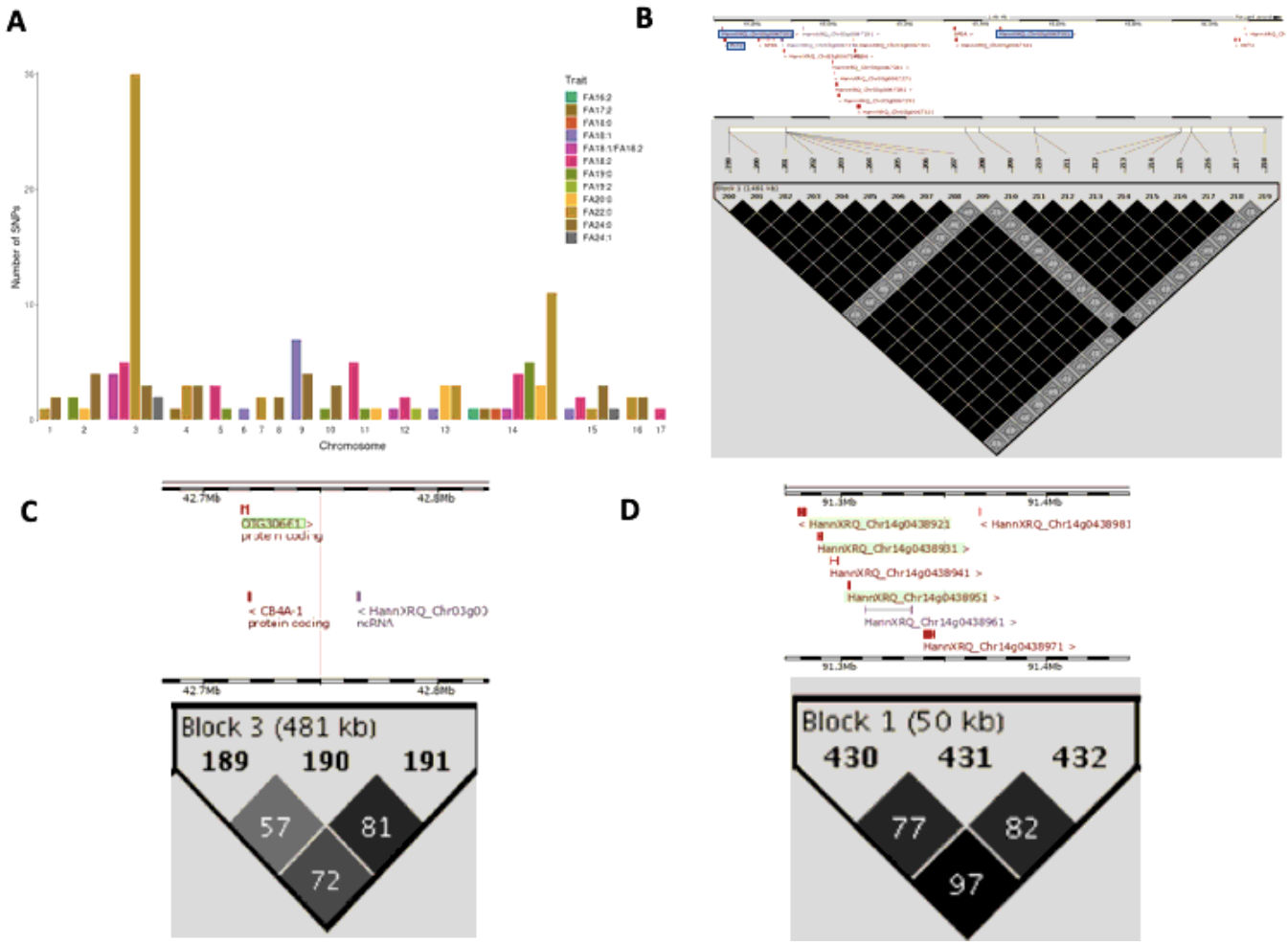

Figure 6

GWAS results for FAs in Sunflower lines and candidate genes for docosanoic acid improvement. (A) Cumulative plot representing the number of significant associations for each of all traits. Traits represented by colors. Chromosome number and number of SNPs are presented on the X and Y-axes respectively. (B) LD block in Chr3 (Location 44696624 - 46188263). (C) LD block in Chr3 (Location 42596595 - 43078214). (D) LD block in Chr14 (Location 91496885 -91547710). Candidate genes in blue associated with lipid metabolism, Candidate genes in green associated with lipid metabolism described by Badouin et.al 2017 [30].

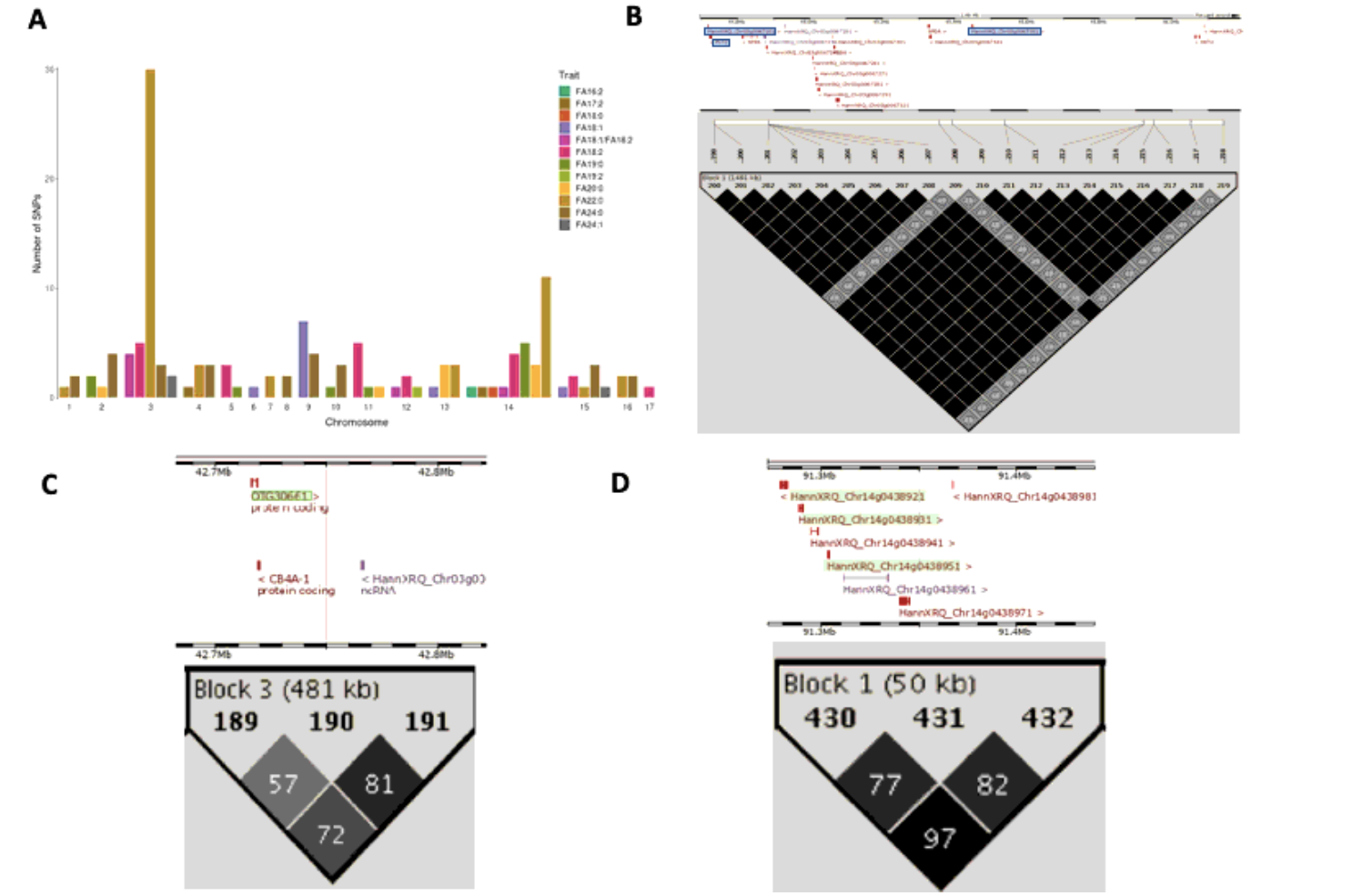

Figure 6

GWAS results for FAs in Sunflower lines and candidate genes for docosanoic acid improvement. (A) Cumulative plot representing the number of significant associations for each of all traits. Traits represented by colors. Chromosome number and number of SNPs are presented on the X and Y-axes respectively. (B) LD block in Chr3 (Location 44696624 - 46188263). (C) LD block in Chr3 (Location 42596595 - 43078214). (D) LD block in Chr14 (Location 91496885 -91547710). Candidate genes in blue associated with lipid metabolism, Candidate genes in green associated with lipid metabolism described by Badouin et.al 2017 [30].

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.pdf
- FigureS1.pdf
- FigureS2.pdf
- FigureS2.pdf

- FigureS3.pdf
- FigureS3.pdf
- FigureS4.pdf
- FigureS4.pdf
- FigureS5.pdf
- FigureS5.pdf
- FigureS6.pdf
- FigureS6.pdf
- FigureS7.pdf
- FigureS7.pdf
- FigureS8.pdf
- FigureS8.pdf
- FigureS9.pdf
- FigureS9.pdf
- FigureS10.pdf
- FigureS10.pdf
- FigureS11.pdf
- FigureS11.pdf
- FigureS12.pdf
- FigureS12.pdf
- FigureS13.pdf
- FigureS13.pdf
- TableS1.xlsx
- TableS1.xlsx
- TableS2.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS6.xlsx

- TableS7.xlsx
- TableS7.xlsx
- TableS8.xlsx
- TableS8.xlsx
- Supportingexperimentalprocedures.docx
- Supportingexperimentalprocedures.docx
- AppendixS1.docx
- AppendixS1.docx
- maxmiss.7.maf0.01.vcf
- maxmiss.7.maf0.01.vcf