Construction and Genetic Analysis of a Worldwide Core Collection of Rapeseed

Yiyi Guo

Zhejiang University

Ying Xu

Zhejiang University

Tao Yan

Hunan Agricultural University

Lixi Jiang

Zhejiang University

Jie Dong

Zhejiang University

Dezhi Wu (✉ wudezhi230@163.com)

Zhejiang University <https://orcid.org/0000-0002-7900-0542>

Liuhui Kuang

Zhejiang University

Research Article

Keywords: Brassica napus, Core collections, Genetic diversity, GWAS, Flowering time

Posted Date: December 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1059860/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Rapeseed (*Brassica napus*) is an important oilseed crop, which is widely planted in the world. In a previous study, we collected 991 accessions of rapeseed from the worldwide germplasm and revealed genetic polymorphisms within these germplasm by whole-genome resequencing. However, management of such a large amount of accessions is time-consuming, laborious and costly. Therefore, we constructed a core collection of rapeseed consisting of 300 worldwide accessions based on their genetic diversity. Compared with 991 accessions, the worldwide core collection showed similar geographic distribution, the proportion of three ecotypes, nucleotide diversity and the associated SNPs of flowering time. Besides, we identified *FT* ortholog (*BnaA02g12130D*) and *FLC* ortholog (*BnaA10g22080D*) responsible for flowering time and ecotype differentiation through selective sweep analysis and genome-wide association analysis (GWAS) of flowering time using the rapeseed core collection. *FT* and *FLC* are two well-known genes regulating flowering time in *Arabidopsis*. These results indicate that the worldwide core collection can represent the genetic diversity of 991 worldwide accessions, which could be more efficiently used for phenotypic and genetic studies in rapeseed.

Introduction

Rapeseed (*Brassica napus* L., AACC, 2n=38) is one of the most economically important oilseed crops in the world (Wang et al., 2014). It originated about 7,500 years ago by the hybridization between *B. rapa* (AA, 2n=18) and *B. oleracea* (CC, 2n=18) (Chalhoub et al., 2014). With the industrial development of society, rapeseed was widely planted in the late nineteenth century due to the high level of erucic acid (Wang et al., 2014). However, excessive erucic acid and glucosinolates in rapeseed are undesirable components for edible oil and harmful to health (Li et al., 2012). Consequently, the single-low (low erucic acid) and double-low (single-low plus low glucosinolate content) cultivars were generated (Li et al., 2012). After that, these cultivars and their derivatives were spread into different countries to develop new canola cultivars adapted to the local climate and environment (Li et al., 2012). To meet the demand of modern agriculture, many new cultivars are released, which have improved their tolerance to diverse biotic and abiotic stresses (Rygulla et al., 2007; Wang et al., 2005).

In a previous study, we collected 991 accessions of rapeseed from 39 countries (Wu et al., 2019). However, management of such large amounts of rapeseed accessions is time-consuming, laborious and costly. To improve the efficiency, a worldwide core collection of rapeseed was urgently needed to be constructed. The concept of a core collection is proposed that in which a limited set of accessions are selected to represent the whole group's genetic diversity (Frankel et al., 1984). In crops, the core collection has been constructed in many species, such as rice (*Oryza sativa*), cotton (*Gossypium barbadense*) and soybean (*Glycine max*) (Li et al., 2003; Xu et al., 2006; Wang et al., 2006). The establishment of a core collection is an efficient way for maintenance and evaluation of the germplasm collections (Holbrook et al., 2000).

Construction of a core collection generally needs three steps. Firstly, the size of the core collection should be decided (Van Hintum et al., 2000). Then, the accessions are clustered into non-overlapping groups based on the obtained information, such as the origin of the accessions, morphological and agronomic characters, molecular markers and so on (Van Hintum et al., 2000; Odong et al., 2013). And finally, the accessions from each cluster were selected using a certain allocation strategy (Van Hintum et al., 2000; Brown et al., 1989). With the enrichment of the theory and practice of core collections, many strategies have been used to determine the number of entries per group, such as the constant (C) strategy, the proportional (P) strategy, the logarithmic (L) strategy and the maximum (M) allelic richness strategy (Van Hintum et al., 2000). Among them, the C-, P-, and L-strategies are

mainly based on morphological and agronomic characters, such as grain yield, leaf area, plant height, etc (Van Hintum et al., 2000; Odong et al., 2013). The M-strategy is based on marker diversity such as simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) (Van Hintum et al., 2000; Odong et al., 2013). With the development of bioinformatic methods, many softwares for selecting core collections were developed and released based on the above strategies, such as MSTRAT, PowerCore, and Core Hunter3 (De Beukelaer et al., 2018; Liu et al., 2018; Gouesnard et al., 2001; Kim et al., 2007). The Core Hunter3 software is widely used to construct core collections in many species (De Beukelaer et al., 2018; Corak et al., 2019; Niu et al., 2020; Milner et al., 2019; Lv et al., 2020).

In this study, we established a worldwide core collection of rapeseed using the Core Hunter3 software, and compared the genetic diversity between 300 core collection and 991 accessions through the genome-wide approaches including principal component analysis (PCA), patterns of linkage-disequilibrium (LD) decay, genome-wide scans of nucleotide diversity, selective sweep analysis and genome-wide association studies (GWAS). The present results demonstrated that the worldwide core collection of rapeseed could represent the genetic diversity within 991 accessions. The construction of this core collection will improve the efficiency of management and utilization of rapeseed resources in future.

Materials And Methods

Construction of a worldwide core collection of rapeseed

Based on the nucleotide diversity of our re-sequenced 991 worldwide accessions of rapeseed (Wu et al., 2019). The Core Hunter3 software was used to construct a core collection with consisting 300 accessions (De Beukelaer et al., 2018). Totally, 4,124,743 SNPs from the 300 accessions were extracted from the VCF file of 991 accessions using VCFTOOLS software (<http://vcftools.sourceforge.net/>; v1.17). Furthermore, 2,291,926 high-quality SNPs were obtained with minor allele frequency (MAF) > 5% and missing data rate ≤ 50% using PLINK software (<https://www.cog-genomics.org/plink2>; v1.9). These high-quality SNPs were used in the following analysis.

Population structure, phylogenetic and principal component analysis

To minimize the interference of extensive strong LD, we filtered SNPs with LD decay (r^2) > 0.2 (sliding window of 100 kb, steps of 10 kb) using the PLINK software (<https://www.cog-genomics.org/plink2>; v1.9) and 311,716 high-confidence SNPs were obtained. These high-confidence SNPs were used to analyze population structures within the 300 core collections using a software package ADMIXTURE (v1.30) in haploid mode with five different random seeds for each K (different numbers of clusters) (K = 2-10) (Alexander et al., 2009). The 'delta K' was calculated for each K value and used to assess the optimal number of populations (Evanno et al., 2005). The Q matrix with the same K value was merged using CLUMPP software (<https://web.stanford.edu/group/rosenberglab/clumpp.html>) and bar plots were drawn using R package Pophelper (<http://royfrancis.github.io/pophelper>).

The PCA was conducted in the smartPCA program from the EIGENSOFT package (<https://github.com/DReichLab/EIG>; v.6.0.1). The PC1 and PC2 were used to distinguish samples from different groups. The maximum-likelihood (ML) phylogenetic tree was constructed using MEGA7 (http://www.megasoftware.net/download_form) and painted by web application EVOLVIEW (<https://evolgenius.info/evolview-v2/#mytrees/300-2/300-2>).

Linkage disequilibrium and nucleotide diversity analysis

Linkage disequilibrium decay was analyzed by calculating correlation coefficients (r^2) between SNPs within 1000 kb using the PLINK software (<https://www.cog-genomics.org/plink2>; v1.9). The parameter as following: -ld-window-r2 0 -ld-window 99999 -ld-window-kb 1000. To determine the degree of variability between the 300 core collection and 991 accessions of rapeseed, nucleotide diversity (p_i) was calculated using the software VCFTOOLS (<http://vcftools.sourceforge.net/>; v1.17) with the following parameters: -window-pi 100000 -window-pi-step 10000.

Selective sweep analysis and genome-wide association studies

The fixation statistics (F_{st}) between different rapeseed ecotypes were calculated using VCFTOOLS (<http://vcftools.sourceforge.net/>; v1.17) with 100 kb sliding and 10 kb step size. Sliding windows with the mean of F_{st} greater than 97% were considered as significant windows. The genes located in the significant windows were considered as selective sweep candidate genes.

A total of 2,291,926 high-quality SNPs were used to perform GWAS of flowering time in the 300 core collections. The data of flowering time was recorded in our previous study (Wu et al., 2019). The kinship matrix (K value) was measured in TASSEL (<http://www.maizegenetics.net/tassel>; v5.0). The PCA matrix (Q value) was calculated using samrtPCA software (<https://github.com/DReichLab/EIG>; v.6.0.1). The K- and Q- values were used in the weighted mixed linear model (MLM) using TASSEL (<http://www.maizegenetics.net/tassel>; v5.0). The $-\log_{10}P$ was calculated for each SNPs and SNPs with $-\log_{10}P > 6$ were defined as the significant SNPs. The figures of manhattan plot were draw using the R package CMplot (<https://github.com/YinLiLin/R-CMplot>).

Results

The core collection can represent the genetic diversity of 991 worldwide accessions of rapeseed

The core collection consists of 300 accessions from 29 countries, and includes 55 semi-winter, 58 spring and 187 winter type accessions, which is similar to the proportion of three ecotypes in 991 worldwide accessions (Fig. 1a, b). At the genomic level, the core collection contains 98% of InDels and 97% of SNPs in the 991 accessions (Fig. 1c). Moreover, the core collection represents more than 94% of SNPs in the 991 accessions on each chromosome (Table 1), and also shows a similar distribution of accessions at the PC1 and the PC2 levels (Fig. 1d). It indicates that the 300 core collection can represent the genetic diversity of 991 worldwide accessions of rapeseed.

Table 1
Comparison of SNP number and LD structures between the core collection and 991 worldwide accessions of rapeseed.

Parameters	Chr.	Core (300)	Worldwide (991)	Percentage (%)	Chr.	Core (300)	Worldwide (991)	Percentage (%)
SNP Number	A01	164,957	175,443	94	C01	245,720	251,468	98
	A02	192,667	203,069	95	C02	269,999	276,530	98
	A03	217,851	230,887	94	C03	348,907	358,520	97
	A04	151,527	157,808	96	C04	303,583	310,911	98
	A05	166,543	175,004	95	C05	218,939	226,111	97
	A06	186,640	196,886	95	C06	223,267	229,123	97
	A07	179,274	189,745	94	C07	251,378	258,922	97
	A08	136,421	144,631	94	C08	230,189	236,310	97
	A09	247,999	262,780	94	C09	259,028	266,102	97
	A10	129,854	136,520	95	Total	4,124,743	4,286,770	96
LD	Core (300)			Worldwide (991)				
$r^2 =$ (half_maxR ²)	0.31			0.30				
LD decay (kb; $r^2 =$ half_maxR ²)	15.94			12.10				
LD decay (kb; $r^2 = 0.1$)	584.72			465.50				
LD blocks	456,666			553,398				
< 1 kb	440,555			544,441				
1–10 kb	15,340			8,505				
> 10 kb	771			452				

We further analyzed the LD structures of the genomes in the core collection and the 991 accessions (Fig. 1e; Table 1). The LD decay rate was faster in the A-subgenome than that in the C-subgenome in the two populations (Fig. 1e), suggesting the frequency of recombination in the A-subgenome is higher than that in the C-subgenome in both populations. The LD decay distance in the core collection (15.94 kb, half_max $r^2 = 0.31$) was similar to that in the 991 accessions (12.10 kb, half_max $r^2 = 0.30$) (Table 1). And, 456,666 LD blocks (< 1 kb 440,555; 1-10 kb 15,340; > 10 kb 771) and 553,398 LD blocks (< 1 kb 544,441; 1-10 kb 8,505; > 10 kb 452) were detected in the core collection and in the 991 accessions, respectively (Table 1), indicating a high frequency of genetic recombination in both populations. In short, the LD structures in the core collection shows a strong resemblance to that in the 991 accessions of rapeseed (Fig. 1e; Table 1).

The core collection has similar patterns of nucleotide diversity, geographic distribution and the proportion of three ecotypes with 991 worldwide accessions

In this study, we compared nucleotide diversity between the core collection and the 991 accessions (Fig. 2, Table S1). The statistic π was used to calculate nucleotide diversity for SNPs located on 19 chromosomes. In the A-subgenome, chromosome A04 had the highest π value (1.34×10^{-3} for the core collection and 1.28×10^{-3} for the 991 accessions, hereafter) and A09 showed the lowest π value (9.0×10^{-4} and 8.59×10^{-4}) (Table S1). In the C-subgenome, the highest π value was found on C01 (1.42×10^{-3} and 1.38×10^{-3}), and the lowest π value was detected on C05 (6.62×10^{-4} and 6.51×10^{-4}) (Table S1). The results showed that the core collection has similar nucleotide diversity and can represent that of the 991 worldwide accessions (Fig. 2, Table S1). To investigate the genetic structures of the core collection, we performed population structure analysis, the PCA and phylogenetic analysis. It is found that the core collection can be grouped into three groups ($K = 3$) (Fig. 3), which is similar to the population structures of the 991 accessions (Wu et al., 2019). Group 1, 2 and 3 include 62, 65 and 173 accessions, respectively, and mainly correspond to the semi-winter, spring and winter type (Table S2). In detail, 82% (51/62) of the accessions in the group 1 exhibit semi-winter type, 83% (54/65) of the accessions in the group 2 are categorized as spring type, and 95% (168/173) of the accessions in the group 3 are winter type (Table S2), indicating the population structures ($K = 3$) with the core collection are mainly consistent with their ecotypes, as consistent with the results of the 991 accessions of rapeseed (Wu et al., 2019). The results of the PCA showed that the PC1 distinguished the group 3 from the groups 1 and 2, while the PC2 separated the group 2 from the group 1 (Fig. 3c). It is demonstrated that the group 1, 2 and 3 were genetically different from each other.

Selective regions were identified by selective sweeps analysis in the core collection

To determine genomic regions with strong selective sweep signals by natural or artificial selection pressure in the core collection, the F_{st} method was employed to determine genomic differences between rapeseed ecotypes (Fig. 4a). The value of F_{st} between spring and winter types, semi-winter and spring types, and semi-winter and winter types was ranged from 0.33 to 0.59, 0.27 to 0.52, and 0.36 to 0.66, respectively (Tables S3-5). It is indicated that there are strong selective regions among three ecotypes in the core collection. The candidate genes located in the selective sweep regions are probably associated with the ecotype differentiation. In particular, we identified a *FLOWERING LOCUS T (FT)* ortholog (*BnaA02g12130D*) between spring and winter types and a *FLOWERING LOCUS C (FLC)* ortholog (*BnaA10g22080D*) between semi-winter and winter types (Fig. 4a), which were two well-known genes regulating flowering time in *Arabidopsis*. On the other hand, the same *FT* and *FLC* orthologs were identified in our previous study (Wu et al., 2019).

Two key genes associated with flowering time was identified by GWAS in the core collection

Flowering time is an important trait associated with yield and ecotypes of *B. napus* (Amasino et al., 2010). In this study, we recorded the flowering time in the core collection of rapeseed. It was ranged from 125 to 201 days, as similar to the 991 accessions (125 to 206 days) (Table 2). In addition, statistical parameters and coefficient of variation (CV) of the flowering time in the core collection were also similar to that in the 991 accessions (Table 2). It is indicated that the core collection can represent the 991 accessions in terms of the flowering time.

Table 2
Comparison of the flowering time (days) between the core collection and 991 worldwide accessions of rapeseed.

Flowering time	Core (300)	Worldwide (991)
MIN	125	125
MAX	201	206
Mode	182	182
Median	181	181
Mean	175.8	176.7
SD	11.5	10.5
CV (%)	6.5	5.9

FT ortholog (*BnaA02g12130D*) and *FLC* ortholog (*BnaA02g12130D*) located on A02 and A10 were showed the strongest selection signals (Fig. 4a). Thus, we performed GWAS for flowering time using the core collection. We identified 37 and 111 significant SNPs ($-\log_{10}P \geq 6$) on A02 and A10, respectively (Fig. 4b, Table S6). Interestingly, a *FT* ortholog (*BnaA02g12130D*) was located in the 100 kb region of an SNP with $-\log_{10}P$ of 7.45 (Fig. 4b). Five SNPs were found in the 5 kb upstream region of *FTC* ortholog (*BnaA02g12130D*), which were significantly associated with the trait of flowering time (Fig. 4b). Taken together, *FT* (*BnaA02g12130D*) and *FLC* (*BnaA02g12130D*) identified by GWAS are two key genes associated with flowing time in the core collection of rapeseed.

Discussion

Constructing an excellent core collection can improve the efficiency of management and utilization of accessions stored in germplasm banks (Frankel et al., 1984). In this study, we successfully constructed a worldwide core collection of rapeseed, which include 300 accessions representing 97% of SNPs from 991 worldwide accessions of rapeseed (Fig. 1c). A series of analysis indicates that this core collection is high-quality and useful for phenotypic and genetic studies in the rapeseed community.

The quality of a core collection decides the quality of subsequent researches. Thus, the most important aid is to verify the quality of a core collection (Odong et al., 2013). With the improvment of the theories and practices, a set of criteria for determining the quality of a core collection have been used, such as summary statistics, shannon diversity index (SH) and class coverage (Odong et al., 2013). Different criteria are suitable for different kinds of traits. Summary statistics of the mean, range and CV are widely used to evaluate the quality of core collection based on continuous traits, such as quantitative traits (Hu et al., 2000). SH and class coverage are suitable for evaluating core collections based on categorical data, such as quality traits and genetic markers (Odong et al., 2013). Besides, the PCA can be used to perform the distribution of the SNPs of core collection in the whole collection (Mahajan et al., 2007). In this study, we found that the summary statistics of flowering time in the 300 core collection were similar to those in the 991 accessions (Table 2). In the PCA, the plot of the PC1 and the PC2 indicated that the SNPs of the core collection were evenly distributed in the 991 accessions (Fig. 1d). Moreover, the results of geographic distribution, the proportion of ecotypes (spring, semi-winter and winter), the number of indels

and SNPs, nucleotide diversity and LD decay showed that the core collection can represent the genetic diversity of the 991 accessions of rapeseed (Fig. 1, 2; Table 2).

The major objective of creating a core collection is efficient use of the genetic resources (Mackay et al., 1995). Recently, many important genes were found in crop species using core collections (Liu et al., 2021; Wang et al., 2016; Ma et al., 2018). For example, 110 accessions of rice from the United States Department of Agriculture Rice Mini-Core Collection were used to perform a GWAS for the tillering, and *OsTCP19* was identified as a major gene responding to rice tillering and nitrogen use (Liu et al., 2021). In other study, 203 rice varieties from rice mini-core collection were used to do GWAS for amylose content, seed length and pericarp color (Wang et al., 2016). For the traits of amylose content and seed length, the causal SNPs associated with these traits were identified (Wang et al., 2016). Recently, a core collection of upland cotton (*Gossypium hirsutum*) comprising 419 accessions was conducted the GWAS for 13 fiber-related traits and identified several important genes for fiber length, fiber strength and flowering time (Ma et al., 2018). These studies implied that a high-quality core collection can be used for exploiting genetic resources in crops.

Flowering time is an important trait, which plays a crucial role in response to environmental cues and endogenous pathways in plants (Amasino et al., 2010). Rapeseed is the second largest oil crop, which is spread into different countries under different climate and environment condition (Chalhoub et al., 2014). To adapt to various growth conditions, rapeseed forms three ecotypes (winter, semi-winter and spring) through regulating its flowering time (Wang et al., 2016). To detect natural or artificial selection between three ecotypes and candidate genes responsible for flowering time, we investigated genomic divergence among three ecotypes in the core collection by calculating pairwise fixation index (F_{st}) values and performing GWAS for flowering time. As expected, the results of the selective-sweep and the GWAS showed that *FT* ortholog (*BnaA02g12130D*) and *FLC* ortholog (*BnaA02g12130D*) were identified in the core collection (Fig. 4). *FT* and *FLC* are two key genes regulating the flowering time and ecotype differentiation in *Arabidopsis* (Michaels et al., 1999; Corbesier et al., 2007).

Undoubtedly, the construction of a worldwide core collection greatly improved the efficiency and was successfully used to identify important genes (Wang et al., 2021; Xuan et al., 2019). For example, *patain-like lipase* (*BnaC07g30920D*) was identified to regulate seed oil content gene through GWAS using the 300 worldwide core collection of rapeseed (Wang et al., 2021). *BnaA.GL1.a* (*BnaA06g31780D*), *BnaC.SWEET4.a* (*BnaC07g24860D*), *BnaC.WAT1.a* (*BnaC07g24950D*) and *BnaC.WAT1.b* (*BnaC07g24960D*) were identified to be associated with the divergence of hairy- and glabrous-leaf genotypes of *B. napus* through GWAS and transcriptomic studies (Xuan et al., 2019). These results demonstrate that our core collection has great value for identification of important candidate genes in rapeseed.

Conclusions

In this study, we constructed a high-quality core collection of rapeseed based on the genetic diversity, which includes 300 worldwide accessions from 29 countries. Through the analysis of the PCA, LD and nucleotide diversity, we confirmed the core collection can represent the genetic diversity of the 991 worldwide accessions. Furthermore, we identified *FT* ortholog (*BnaA02g12130D*) and *FLC* ortholog (*BnaA10g22080D*) responsible for the flowering time of rapeseed through selective sweep analysis and GWAS methods. These results indicate that the worldwide core collection of rapeseed improves the efficiency of the phenotypic and genetic experiments, which is useful for researches and rapeseed breeding.

Declarations

Acknowledgements

We are grateful to Dr. J.M. Xu (Zhejiang University, China) for the technical support. This research was supported by Jiangsu Collaborative Innovation Center for Modern Crop Production.

Authors' contributions

DZ Wu and LX Jiang designed the research. YY Guo, Y Xu, Y Tao and J Dong performed the research. YY Guo and LH Kuang analyzed the data. YY Guo, LH Kuang and DZ Wu wrote the article. The authors declare no conflict of interest.

Founding information

This research was supported by the Zhejiang Provincial Key Research Project (2021C02057) and Jiangsu Collaborative Innovation Center for Modern Crop Production.

Compliance with ethical standards

Conflict of interest

The authors declared no conflict of interest.

References

1. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664. <https://doi.org/10.1101/gr.094052.109>
2. Amasino RM, Michaels SD (2010) The timing of flowering. *Plant Physiol* 154(2):516–520. <https://doi.org/10.1104/pp.110.161653>
3. Brown AHD (1989) Core collections - a practical approach to genetic-resources management. *Genome* 31(2):818–824. <https://doi.org/10.1139/g89-144>
4. Corak KE, Ellison SL, Simon PW, Spooner DM, Dawson JC (2019) Comparison of representative and custom methods of generating core subsets of a carrot germplasm collection. *Crop Sci* 59(3):1107–1121. <https://doi.org/10.2135/cropsci2018.09.0602>
5. Corbesier L, Vincent C, Jang SH, Fornara F, Fan QZ, Searle I, Giakountis A, Farrona S, Gissot L, Turnbull C, Coupland G (2007) FT protein movement contributes to long-distance signaling in floral induction of *Arabidopsis*. *Science* 316(5827):1030–1033. <https://doi.org/10.1126/science.1141752>
6. Chalhoub B, Denoeud F, Liu SY, Parkin IAP, Tang HB, Wang XJ, Chiquet J, Belcram H, Tong CB, Samans B, Correa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao MX, Edger PP, Chelaifa H, Tack D, Lassalle G, Mestiri I, Schnel N, Le Paslier MC, Fan GY, Renault V, Bayer PE, Golicz AA, Manoli S, Lee TH, Thi VHD, Chalabi S, Hu Q, Fan CC, Tollenaere R, Lu YH, Battail C, Shen JX, Sidebottom CHD, Wang XF, Canaguier A, Chauveau A, Berard A, Deniot G, Guan M, Liu ZS, Sun FM, Lim YP, Lyons E, Town CD, Bancroft I, Wang XW, Meng JL, Ma JX, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury JM, Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou YM, Hua W, Sharpe AG, Paterson AH, Guan CY, Wincker P (2014) Early allopolyploid evolution

- in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345(6199):950–953.
<https://doi.org/10.1126/science.1253435>
7. De Beukelaer H, Davenport GF, Fack V (2018) Core Hunter 3: flexible core subset selection. *BMC Bioinformatics* 19(203). <https://doi.org/10.1186/s12859-018-2209-z>
 8. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14(8):2611–2620. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>
 9. Frankel OH (1984) Genetic perspectives of germplasm conservation. In: W. Arber (ed) *Genetic manipulation: impact on man and society*, Cambridge, England, pp 161–170
 10. Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: An algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *J Hered* 92(1):93–94. <https://doi.org/10.1093/jhered/92.1.93>
 11. Holbrook CC, Timper P, Xue HQ (2000) Evaluation of the core collection approach for identifying resistance to *Meloidogyne arenaria* in peanut. *Crop Sci* 40(4):1172–1175. <https://doi.org/10.2135/cropsci2000.4041172x>
 12. Hu J, Zhu J, Xu HM (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor Appl Genet* 101(1–2):264–268. <https://doi.org/10.1007/s001220051478>
 13. Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG, Park YJ (2007) PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23(16):2155–2162. <https://doi.org/10.1093/bioinformatics/btm313>
 14. Li W, Jiang W, Zhao HX, Vyvadilova M, Stamm M, Hu SW (2012) Genetic diversity of rapeseed accessions from different geographic locations revealed by expressed sequence tag-simple sequence repeat and random amplified polymorphic dna markers. *Crop Sci* 52(1):201–210. <https://doi.org/10.2135/cropsci2011.06.0301>
 15. Liu XB, Li J, Yang ZL (2018) Genetic diversity and structure of core collection of winter mushroom (*Flammulina velutipes*) developed by genomic SSR markers. *Hereditas* 155(1):3. <https://doi.org/10.1186/s41065-017-0038-0>
 16. Li ZC, Zhang HL, Cao YS, Qiu ZE, Wei XH, Tang SY, Yu P, Wang XK (2003) Studies on the sampling strategy for primary core collection of Chinese ingenious rice. *Acta Agronomica Sinica* 29(1):20–24. <https://doi.org/10.3321/j.issn:0496-3490.2003.01.004>
 17. Liu YQ, Wang HR, Jiang ZM, Wang W, Xu RN, Wang QH, Zhang ZH, Li AF, Liang Y, Ou SJ, Liu XJ, Cao SY, Tong HN, Wang YH, Zhou F, Liao H, Hu B, Chu CC (2021) Genomic basis of geographical adaptation to soil nitrogen in rice. *Nature* 590(7847). <https://doi.org/10.1038/s41586-020-03091-w>
 18. Lv JB, Li CR, Zhou CP, Chen JB, Li FG, Weng QJ, Li M, Wang YQ, Chen SK, Chen JC, Gan SM (2020) Genetic diversity analysis of a breeding population of *Eucalyptus cloeziana* F. Muell. (Myrtaceae) and extraction of a core germplasm collection using microsatellite markers. *Ind Crop Prod* 145(112157). <https://doi.org/10.1016/j.indcrop.2020.112157>
 19. Ma ZY, He SP, Wang XF, Sun JL, Zhang Y, Zhang GY, Wu LQ, Li ZK, Liu ZH, Sun GF, Yan YY, Jia YH, Yang J, Pan ZE, Gu QS, Li XY, Sun ZW, Dai PH, Liu ZW, Gong WF, Wu JH, Wang M, Liu HW, Feng KY, Ke HF, Wang JD, Lan HY, Wang GN, Peng J, Wang N, Wang LR, Pang BY, Peng Z, Li RQ, Tian SL, Du XM (2018) Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat Genet* 50(6):803–813. <https://doi.org/10.1038/s41588-018-0119-7>

20. Mackay MC (1995) One core collection or many? In: Hodgkin T, Brown AHD, Van Hintum TJL, Morales EAV (eds) Core collections of plant genetic resources. John Wiley & Sons Ltd., Chichester, pp 199–210
21. Mahajan RK, Bisht IS, Dhillon BS (2007) Establishment of a core collection of world sesame (*Sesamum indicum* L.) germplasm accessions. SABRAO J Breed Genet 39(1):53–64. <https://doi.org/10.3724/sp.j.1006.2009.01936>
22. Michaels SD, Amasino RM (1999) *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11(5):949–956. <https://doi.org/10.1105/tpc.11.5.949>
23. Milner SG, Jost M, Taketa S, Mazon ER, Himmelbach A, Oppermann M, Weise S, Knuepffer H, Basterrechea M, Koenig P, Schueler D, Sharma R, Pasam RK, Rutten T, Guo GG, Xu DD, Zhang J, Herren G, Mueller T, Krattinger SG, Keller B, Jiang Y, Gonzalez MY, Zhao YS, Habekuss A, Faerber S, Ordon F, Lange M, Boerner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N (2019) Genebank genomics highlights the diversity of a global barley collection. Nat Genet 51(2):319. <https://doi.org/10.1038/s41588-018-0266-x>
24. Niu SZ, Koiwa H, Song QF, Qiao DH, Chen J, Zhao DG, Chen ZW, Wang Y, Zhang TY (2020) Development of core-collections for Guizhou tea genetic resources and GWAS of leaf size using SNP developed by genotyping-by-sequencing. PEERJ 8:e8572. <https://doi.org/10.7717/peerj.8572>
25. Odong TL, Jansen J, Van Eeuwijk FA, Van Hintum TJL (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor Appl Genet 126(2):289–305. <https://doi.org/10.1007/s00122-012-1971-y>
26. Rygulla W, Snowdon RJ, Eynck C, Koopmann B, von Tiedemann A, Luehs W, Friedt W (2007) Broadening the genetic basis of *Verticillium longisporum* resistance in *Brassica napus* by interspecific hybridization. Phytopathology 97(11):1391–1396. <https://doi.org/10.1094/PHYTO-97-11-1391>
27. Van Hintum ThJL, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic resources. Rome, Italy
28. Wang HR, Xu X, Vieira FG, Xiao YH, Li ZK, Wang J, Nielsen R, Chu CC (2016) The power of inbreeding: NGS-based GWAS of rice reveals convergent evolution during rice domestication. Mol Plant 9(7):975–985. <https://doi.org/10.1016/j.molp.2016.04.018>
29. Wang HY, Wang Q, Pak H, Yan T, Chen MX, Chen XY, Wu DZ, Jiang LX (2021) Genome-wide association study reveals a patatin-like lipase relating to the reduction of seed oil content in *Brassica napus*. BMC Plant Biol 21(1). <https://doi.org/10.1186/s12870-020-02774-w>
30. Wang JX, Chen ZL, Du JZ, Sun Y, Liang AH (2005) Novel insect resistance in *Brassica napus* developed by transformation of chitinase and scorpion toxin genes. Plant Cell Rep 24(9):549–555. <https://doi.org/10.1007/s00299-005-0967-3>
31. Wang LX, Guan Y, Guan RX, Li YH, Ma YS, Dong ZM, Liu X, Zhang HY, Zhang YQ, Liu ZX, Chang RZ, Xu HM, Li LH, Lin FY, Luan WJ, Yan Z, Ning XC, Zhu L, Cui YH, Piao R, Liu Y, Chen PY, Qiu LJ (2006) Establishment of Chinese soybean *Glycine max* core collections with agronomic traits and SSR markers. Euphytica 151(2):215–223. <https://doi.org/10.1007/s10681-006-9142-3>
32. Wang N, Li F, Chen BY, Xu K, Yan GX, Qiao JW, Li J, Gao GZ, Bancroft I, Meng JL, King GJ, Wu XM (2014) Genome-wide investigation of genetic changes during modern breeding of *Brassica napus*. Theor Appl Genet 127(8):1817–1829. <https://doi.org/10.1007/s00122-014-2343-6>
33. Wang N, Chen BY, Xu K, Gao GZ, Li F, Qiao JW, Yan GX, Li J, Li H, Wu XM (2016) Association mapping of flowering time QTLs and insight into their contributions to rapeseed growth habits. Front Plant Sci 7.

34. Wu DZ, Liang Z, Yan T, Xu Y, Xuan LJ, Tang J, Zhou G, Lohwasser U, Hua SJ, Wang HY, Chen XY, Wang Q, Zhu L, Maodzeka A, Hussain N, Li ZL, Li XM, Shamsi IH, Jilani G, Wu LD, Zheng HK, Zhang GP, Chalhoub B, Shen LS, Yu H, Jiang LX (2019) Whole-genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. *Mol Plant* 12(1):30–43. <https://doi.org/10.1016/j.molp.2018.11.007>
35. Xu HM, Mei YJ, Hu J, Zhu J, Gong P (2006) Sampling a core collection of island cotton (*Gossypium barbadense* L.) based on the genotypic values of fiber traits. *Genet Resour Crop Ev* 53(3):515–521. <https://doi.org/10.1007/s10722-004-2032-4>
36. Xuan LJ, Yan T, Lu LZ, Zhao XZ, Wu DZ, Hua SJ, Jiang LX (2019) Genome-wide association study reveals new genes involved in leaf trichome formation in polyploid oilseed rape (*Brassica napus* L.). *Plant Cell Environ* 43(3):675–691. <https://doi.org/10.1111/pce.13694>

Figures

Figure 1

Comparison of geographic distribution, ecotypes and genetic diversity between the core collection and 991 worldwide accession of rapeseed. **a** The geographic distribution of the core collection (300 accessions) and 991 worldwide rapeseed accessions. **b** The proportion of spring, semi-winter and winter type in the core collection and the 991 accessions, respectively. **c** The number of indels and SNPs in the core collection, which accounts for 98%

and 97% of that in the 991 accessions. **d** The PCA plot of the accessions in the core and the whole accessions. The blue and red points indicate the accessions in the whole and core collections, respectively. **e** LD decay of the A- and C- subgenomes in the core and the whole collections. The blue and red lines respectively represent the rate of LD decay in the A and C subgenomes. The straight and dashed lines represent the core collection and the 991 accessions, respectively.

Figure 2

Comparison of nucleotide diversity (π) between the core collection and 991 accessions of rapeseed. **a** The nucleotide diversity (π) of A subgenome. **b** The nucleotide diversity (π) of C subgenome. Blue and red lines indicate the π values in the 991 accessions and the core collections.

Figure 3

Population structure, PCA and phylogenetic tree of the core collection of rapeseed. **a** Model-based clustering analysis of the core collection (300 accessions) with the different groups ($K = 2-10$) using ADMIXTURE. The x-axis represents different accessions, and the y axis quantifies cluster membership. Different colors mean different clusters. **b** The ΔK was estimated for population structure analysis. **c** The PCA plot of the core collection by the PC1 and PC2. **d** Phylogenetic tree of the core collection.

Figure 4

Selective sweep regions and GWAS analysis of the core collection. **a** Genome-wide selective-sweep signals through the comparisons between three ecotypes (spring, winter and semi-winter). The red dashed lines show the thresholds (top 3% of F_{st} values). **b** Manhattan plots of GWAS results for the flowering time. The identified SNPs were located in the regions of 4-9 Mb on A02 (left) and the region of 13-17 Mb on A10 (right). The red dashed lines indicate the threshold ($-\log_{10}P = 6.0$).

Supplementary Files

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

- [TablesS16.xlsx](#)