

# Diversity analysis and genomic prediction of *Sclerotinia* resistance in sunflower using a new 25 K SNP genotyping array

Maren Livaja<sup>1</sup> · Sandra Unterseer<sup>1</sup> · Wiltrud Erath<sup>1</sup> · Christina Lehermeier<sup>1</sup> · Ralf Wieseke<sup>2</sup> · Jörg Plieske<sup>2</sup> · Andreas Polley<sup>2</sup> · Hartmut Luerßen<sup>2</sup> · Silke Wieckhorst<sup>3</sup> · Martin Mascher<sup>5</sup> · Volker Hahn<sup>4</sup> · Milena Ouzunova<sup>3</sup> · Chris-Carolin Schön<sup>1</sup> · Martin W. Ganal<sup>2</sup>

Received: 1 July 2015 / Accepted: 23 October 2015 / Published online: 4 November 2015  
© Springer-Verlag Berlin Heidelberg 2015

## Abstract

**Key message** We have developed a SNP array for sunflower containing more than 25 K markers, representing single loci mostly in or near transcribed regions of the genome. The array was successfully applied to genotype a diversity panel of lines, hybrids, and mapping populations and represented well the genetic diversity of cultivated sunflower. Results of PCoA and population substructure analysis underlined the complexity of the genetic composition of current elite breeding material. The performance of this genotyping platform for genome-based prediction of phenotypes and detection of QTL with improved resolution could be demonstrated based on the re-evaluation of a population segregating

for resistance to *Sclerotinia* midstalk rot. Given our results, the newly developed 25 K SNP array is expected to be of great utility for the most important applications in genome-based sunflower breeding and research.

**Abstract** Genotyping with a large number of molecular markers is a prerequisite to conduct genome-based genetic analyses with high precision. Here, we report the design and performance of a 25 K SNP genotyping array for sunflower (*Helianthus annuus* L.). SNPs were discovered based on variant calling in de novo assembled, UniGene-based contigs of sunflower derived from whole genome sequencing and amplicon sequences originating from four and 48 inbred lines, respectively. After inclusion of publicly available transcriptome-derived SNPs, *in silico* design of the Illumina<sup>®</sup> Infinium iSelect HD BeadChip yielded successful assays for 22,299 predominantly haplotype-specific SNPs. The array was validated in a sunflower diversity panel including inbred lines, open-pollinated varieties, introgression lines, landraces, recombinant inbred lines, and F<sub>2</sub> populations. Validation provided 20,502 high-quality bi-allelic SNPs with stable cluster performance whereby each SNP marker represents a single locus mostly in or near transcribed regions of the sunflower genome. Analyses of population structure and quantitative resistance to *Sclerotinia* midstalk rot demonstrate that this array represents a significant improvement over currently available genomic tools for genetic diversity analyses, genome-wide marker-trait association studies, and genetic mapping in sunflower.

Communicated by B. Hulke.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-015-2629-3) contains supplementary material, which is available to authorized users.

✉ Chris-Carolin Schön  
chris.schoen@tum.de

<sup>1</sup> Department of Plant Sciences, Plant Breeding, Technische Universität München, Liesel-Beckmann-Strasse 2, 85354 Freising, Germany

<sup>2</sup> TraitGenetics GmbH, Am Schwabeplan 1b, 06466 Gatersleben, Germany

<sup>3</sup> KWS SAAT SE, Grimsehlstrasse 31, 37555 Einbeck, Germany

<sup>4</sup> State Plant Breeding Institute, Universität Hohenheim, Fruwirthstrasse 21, 70599 Stuttgart, Germany

<sup>5</sup> Research Group Domestication Genomics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Stadt Seeland, Germany

## Introduction

Many sophisticated genomic tools have been established within the last decade and led to a potentiation in resource development of major crop species that are useful for

plant breeding and genetic studies. Sequencing of whole genomes/transcriptomes enabled the genome/transcriptome-wide discovery of single nucleotide polymorphism (SNP) markers amenable for high-throughput genotyping platforms. Genotyping arrays are now used for many purposes, such as genetic diversity analysis, high density genetic mapping, fine mapping of quantitative trait loci (QTL), and detection of marker-trait associations suitable for the application of marker-assisted selection (MAS) in breeding programs.

Cultivated sunflower (*Helianthus annuus* L.) is an oil-seed crop of great economic importance with a worldwide production of approximately 44.7 million tons per annum (FAOSTAT 2013). Sunflower is considered a model species for other large-genome members of the Compositae family, especially with regards to evolutionary and ecological questions. The whole genome (~3.6 Gbp) of inbred line HA412HO is target of an on-going sequencing project (Gill et al. 2014; Grassa et al. 2015; Kane et al. 2011; Natali et al. 2013) and will represent the reference genome sequence for this species. Once assembled, it will dramatically facilitate the discovery of SNP markers by resequencing of other sunflower lines.

The development of molecular markers is advanced in sunflower and, over the years, different marker types have been generated. Restriction Fragment Length Polymorphisms (RFLP) fingerprinting was the first molecular marker technique available in sunflower (Berry et al. 1994; Gentzbittel et al. 1992, 1994). Several detailed genetic linkage maps have been developed based on simple sequence repeat (SSR), sequence-tagged-site (STS) markers, and EST-derived SNP markers (Pérez Vich and Berry 2010) for both cultivated sunflower (Al-Chaarani et al. 2004; Berrios et al. 2000; Lai et al. 2005a; Tang et al. 2002; Yu et al. 2003), as well as some wild relatives (Barb et al. 2014; Burke et al. 2004; Heesacker et al. 2009; Lai et al. 2005b; Rieseberg et al. 2003). Based on transcriptome sequencing, a medium density SNP array (10 K) for sunflower was successfully developed (Bachlava et al. 2012). It was used to construct an integrated high-resolution genetic linkage map of *H. annuus* L. (Bowers et al. 2012a) and for association mapping (Mandel et al. 2013). Using a custom Affymetrix Expression GeneChip, an ultra-dense genetic map for sunflower was developed by placing 67,486 short features representing 22,481 unigenes (Bowers et al. 2012b). Sunflower SNP resources were extended by applying a restriction site-associated DNA sequencing (RAD-Seq) approach that finally resulted in an Illumina Infinium array with 8723 SNPs suitable for genotyping and genetic mapping of three populations (5019 mapped markers) (Pegadaraju et al. 2013; Talukder et al. 2014). To meet the demand for an integrated dense genetic map based on publically available SNP resources and genetic maps developed by two

SNP marker consortia (Bowers et al. 2012a; Talukder et al. 2014) an *in silico* approach was used. Resequencing of a mapping population and alignment of resulting contigs and of known marker flanking sequences on draft genome scaffolds allowed to determine the genetic positions of more than 10,000 markers in an unified map (Hulke et al. 2015). To analyze and predict complex agronomic traits, genome-wide approaches as genome-wide association studies (GWAS) or whole genome-based prediction (Meuwissen et al. 2001) have become popular genomic tools. However, compared to QTL mapping in bi- or multiparental crosses, a larger set of markers is required in GWAS or genomic prediction to ensure that linkage disequilibrium (LD) between markers and QTL is preserved (Goddard and Hayes 2007; Mammadov et al. 2012). The aim of the current study was to develop a genotyping array based on the Illumina® Infinium assay, with a high number of predominantly haplotype-specific SNP markers located mostly in or near genes that can be used for a better understanding of the genetic regulation of complex agronomic traits in sunflower.

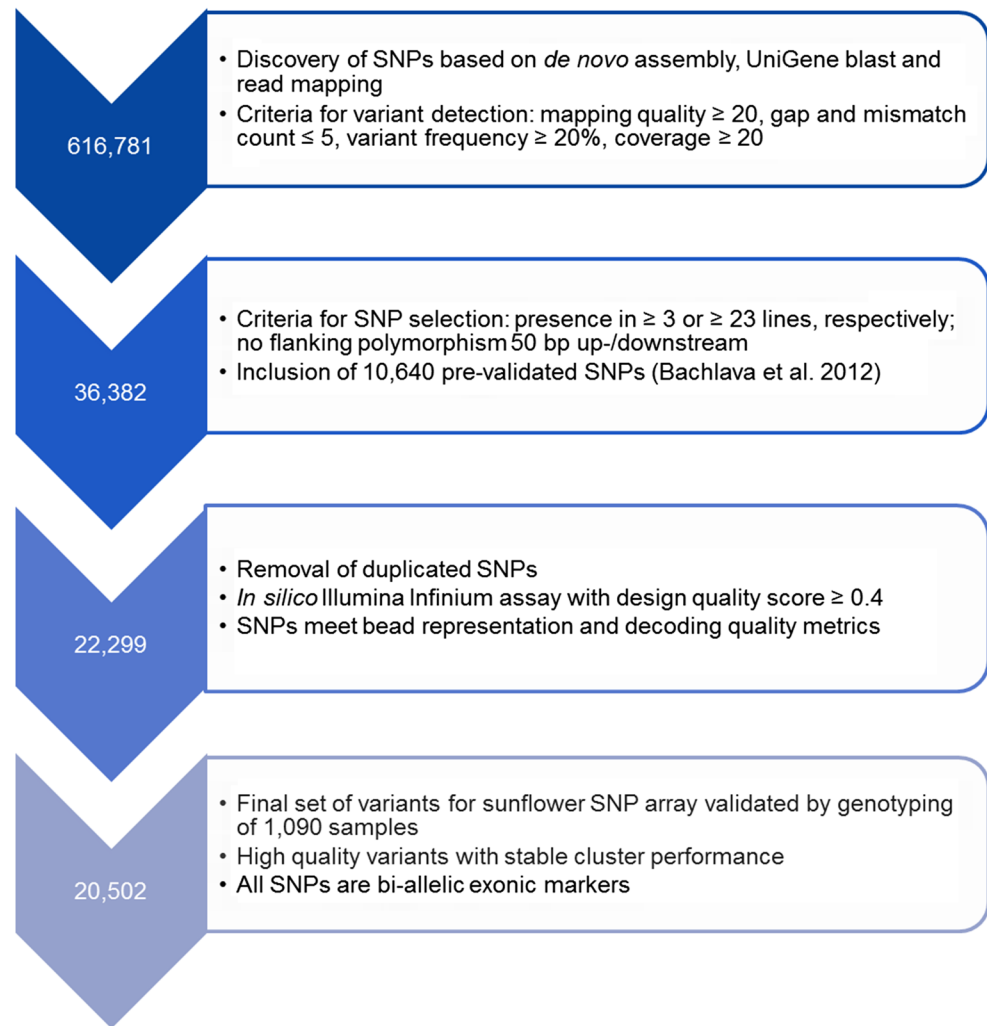
## Materials and methods

### Whole genome and amplicon sequencing

Four sunflower inbred lines representing two main groups of the sunflower gene pool (two restorer; SUN48-0003, SUN48-0006 and two maintainer lines; SUN48-0025, SUN48-0026) were selected for whole genome sequencing (WGS). For high molecular weight DNA extraction, a protocol for cell nuclei isolation (Murray and Thompson 1980) was applied that minimizes mitochondrial and chloroplast DNA contamination. Quality of DNA was checked electrophoretically, and 5 µg was subjected to standard 350 bp library preparation according manufacturer's protocols (Illumina, San Diego, USA).

For amplicon sequencing, 48 maintainer and restorer lines representative of current elite breeding material were selected. Based on 5955 EST sequences of the *Helianthus annuus* UniGene EST Set (<http://www.ncbi.nlm.nih.gov/genbank/dbest>), primers were designed using Primer 3.06 software (Untergasser et al. 2012). The primer pairs were tested for amplification from eight genotypes using following PCR conditions: 5 min at 94 °C; 40 cycles of 1 min at 94 °C, 1 min at 60 °C, and 2 min at 72 °C; and a final extension step of 10 min at 72 °C. The reaction volume was set to 25 µl containing 20 ng DNA, 1 × GoTaq® buffer and 1 unit GoTaq® polymerase (Promega, Madison, USA), 0.5 µl of each primer (10 µmol/l), 1.5 µl dNTPs (25 µmol/l). More than 56 % (3356) of the deduced primer pairs showed distinct bands on agarose gel after PCR. Amplification was

**Fig. 1** Flow diagram describing steps and major criteria of the SNP selection process during the development of the sunflower 25 K genotyping array



carried out using these 3356 primer pairs (Table S1) for further 40 genotypes. Amplicons of each inbred line were pooled and each 10  $\mu\text{g}$  was used to prepare 200 bp insert libraries as recommended by Illumina (Illumina Inc., San Diego, CA, USA). Pool-specific bar codes were added by ligation of a six-base index sequence-containing adapter allowing for filtering of reads after sequencing. WGS and amplicon sequencing was performed on an Illumina HiSeq 2000 instrument using the  $2 \times 100$  bp paired-end sequencing strategy (Aros Applied Biotechnology A/S, Aarhus, Denmark).

### SNP detection and final choice

A multi-step selection procedure (Fig. 1) was followed to obtain high-confidence bi-allelic SNP markers with stable cluster performance. For *de novo* assembly, we employed the CLC Assembly Cell *de novo* assembler (version 3.2.2, CLC bio, Aarhus, Denmark). First, raw sequence reads were quality-trimmed with the program “quality\_trim”.

In order to create a reference sequence against which SNP calling could be carried out, trimmed reads of all WGS and amplicon sequencing were assembled using the parameter settings “-p fb ss 250 450”. Contigs generated from the genomic sequences were used for blast analyses to detect genomic targets from the UniGene set (Build # 11, *Helianthus annuus*, NCBI). UniGene (<http://www.ncbi.nlm.nih.gov/UniGene/>) is a largely automated analytical system that produces an organized view of a species specific transcriptome. Subsequently, all reads were mapped to the generated *de novo* reference sequence (identified contigs representing unigenes) whereby reads that matched more than once were ignored. CLC Genomics Workbench 5.01 was used for further mapping reads with the following CLC parameters applied for quality-based variant detection: maximum expected variations (ploidy) = 2; maximum gap and mismatch count = 5; minimum average quality = 15; minimum central quality = 20; minimum coverage = 20; minimum variant frequency (%) = 20.0; window length = 11.

SNPs were called from genomic sequences if the sequences were present in  $\geq 3$  sunflower lines. SNPs derived from amplicon sequences were included if sequences at that position were available from at least 23 lines. SNPs with nearby polymorphisms within the 50 bp left and right were eliminated. Further, SNPs with more than two alleles were discarded. SNP markers that represented the same haplotype over the entire contig were also reduced to one entry.

Additional sequences of pre-validated SNPs available from the sunflower 10,640 SNP genotyping array (Bachlava et al. 2012) were included in the selection process. SNP selection was continued with the removal of duplicated SNPs. All remaining SNPs were submitted to the Illumina Assay Design Tool (Illumina, San Diego, CA), and only SNPs that matched the Illumina<sup>®</sup> Infinium assay quality requirements (final score  $\geq 0.4$ ) were finally used for the array design.

### Plant material and genotyping

The resulting sunflower Infinium iSelect HD BeadChip (Illumina<sup>®</sup>, San Diego, USA) was used to genotype a diversity panel of lines, hybrids, and mapping populations, altogether 1090 genotypes. Among them was a sunflower collection of 287 accessions (Table S2) representing 243 inbred lines, 19 open-pollinated varieties (OPVs), 5 landraces, and 20 lines with recent introgressions from wild *Helianthus* relatives that are referred to as introgression lines. This set of accessions captures nearly 90 % of the allelic diversity present within the gene pool of cultivated sunflower (Mandel et al. 2011) and originates from collections of the USDA North Central Regional Plant Introduction Station (NCRPIS) and the French National Institute for Agricultural Research (INRA). Pedigree information was available for about half of the accessions (USDA-ARS 2014; USDA 2006). Information on the designation into the categories maintainer, restorer, nonoil, and oil was available for almost all USDA inbred lines (USDA 2006). INRA-derived accessions could not be assigned to the nonoil or oil class. However, they could be distinguished in terms of breeding history into maintainer and restorer class (INRA 2014; Mandel et al. 2011, 2013).

Further, a subset of the population NDBLOS<sub>sel</sub> × CM625 consisting of 159 recombinant inbred lines (RILs) which was previously used for QTL mapping of resistance to *Sclerotinia* midstalk rot (Micic et al. 2005) was investigated. Prior to DNA isolation and genotyping, the ninth generation of each RIL was generated by selfing the previous generation ( $F_8$ ) through single seed descent.

22,299 SNPs were analyzed with respect to their genotype clustering using GenomeStudio software (v2011.1, Illumina, San Diego, USA). In order to create three

high-quality clusters to represent the three possible genotypes at each locus, SNP marker quality was assessed by visual inspection of the cluster distribution and by subsequent adjustment of the cluster calling for each marker, exemplified in Figure S1. SNP markers for which two or more polymorphic loci were scored simultaneously (i.e. SNPs that created more than three clusters) were excluded.

### Use of SNP array for analysis of population structure

To assess the utility of the 25 K SNP array in detecting population structure within the set of inbred lines, Principal Coordinate Analysis [PCoA; (Gower 1966)] and population substructure analysis using ADMIXTURE (Alexander et al. 2009) were performed using genotypic data of 243 inbred lines. SNP markers with  $\geq 5$  % missing data were excluded. Remaining missing data were imputed using Beagle (Browning and Browning 2009) via the R package “synbreed” (Wimmer et al. 2012) using R version 3.0.1 (<http://www.R-project.org/>). PCoA was calculated based on Rogers’ distances using R with the packages “synbreed” (Wimmer et al. 2012), “adegenet” (Jombart 2008), and “ape” (Paradis et al. 2004). File conversion was done via Plink version 1.07 (Purcell et al. 2007) and analysis of population substructure was calculated using ADMIXTURE version 1.23 (Alexander et al. 2009) running with default settings for  $K = 1$  to  $K = 20$ . Nucleotide diversity was calculated per SNP according to Tajima (1983) and the differentiation index  $F_{ST}$  according to Weir and Cockerham (1984). In case of landraces and OPVs, SNPs were tested for Hardy–Weinberg equilibrium (HWE) using an exact test following Wigginton et al. (2005) with  $p \leq 0.001$ .  $F_{ST}$  calculation and HWE tests were performed with the software Plink (version 1.09) with default parameter settings (Chang et al. 2015).

### Genetic mapping

A genetic map of the 159 RILs derived from the cross NDBLOS<sub>sel</sub> × CM625 was constructed by using JoinMap 4.0 (van Ooijen 2006) with default parameter settings. Graphical genotypes representing the calculated linkage groups (LGs) were visually inspected, and doubtful genotyping results such as low quality data and suspicious double cross-overs were eliminated from the dataset. MapManager QTXb20 version 0.3 (Manly et al. 2001) was used to recalculate the map positions. Distances between SNP markers were estimated using the Kosambi mapping function (Kosambi 1944). Linkage group assignment according to Tang et al. (2002) was based on the overlap of SNPs mapped by Bowers et al. (2012a) and our marker set.

## Use of SNP array for analysis of quantitative traits

The suitability of our SNP genotyping platform was investigated for QTL mapping and genome-based prediction of sunflower midstalk rot resistance caused by the pathogen *Sclerotinia sclerotiorum*. From the study of Micic et al. (2005) we extracted phenotypic data for three resistance traits, stem lesion length (SLL), speed of fungal growth (SFG), and leaf lesion length (LLL) as well as for the morphological trait leaf length with petiole (LLP) for 113 RILs derived from the cross NDBLOS<sub>sel</sub> × CM625. Based on the genotyping data and genetic linkage map described above and on adjusted entry means from field trials across two locations, QTL mapping was performed for each trait using composite interval mapping (CIM) implemented in the software package PLABQTL 1.2 (Utz and Melchinger 2006). A conservative LOD threshold corresponding to an experiment-wise type I error rate of  $\alpha = 0.05$  was chosen. This threshold was determined using 1000 permutations as described by Churchill and Doerge (1994). The support interval of a putative QTL was defined as the chromosomal region surrounding a QTL peak with a LOD fall off of 1.0. The additive effect as well as the phenotypic variance explained ( $R^2$ ) by each QTL was obtained from a multiple regression model fitting all significant QTL simultaneously.

For genomic prediction of *Sclerotinia* resistance traits, a genome-based best linear unbiased prediction (GBLUP) model was used:  $\mathbf{y} = \mathbf{1}_n\mu + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ , where  $\mathbf{y}$  is the  $n$ -dimensional vector of adjusted means across the two locations for the  $n = 113$  RILs,  $\mathbf{1}_n$  is an  $n$ -dimensional vector of ones,  $\mu$  is an overall mean,  $\mathbf{Z}$  is an  $n \times n$  matrix assigning genotypes to phenotypes. The  $n$ -dimensional vector  $\mathbf{u}$  of genotypic effects is assumed to be normally distributed with  $\mathbf{u} \sim N(\mathbf{0}, \mathbf{U}\sigma_g^2)$ , where  $\mathbf{U}$  is a marker-derived relationship matrix calculated according to Habier et al. (2007) and  $\sigma_g^2$  is the genotypic variance. The  $n$ -dimensional vector of residuals is assumed to be normally distributed with  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$ , where  $\mathbf{I}$  is an  $n \times n$  dimensional identity matrix and  $\sigma_\varepsilon^2$  is the residual variance. Genotypic and residual variances were estimated by restricted maximum likelihood using ASREML (Gilmour et al. 2009). To assess the prediction performance of the model, ten times replicated fivefold cross-validation with random sampling in estimation and test set was performed as described in Albrecht et al. (2011). Predictive ability of the model was estimated as Pearson's correlation between predicted and observed phenotypes of lines in the test set. Further, prediction accuracy indicating the correlation between predicted and observed genotypes was approximated by the mean predictive ability divided by the square-root of the trait heritability (Dekkers 2007). Analyses were performed using the "synbreed" R package (Wimmer et al. 2012).

## Results

### SNP array development

Sequencing of the four sunflower lines on Illumina HiSeq 2000 resulted in a total of 268 Gb of DNA sequence generated from 100 bp paired-end reads. On average, a >20-fold coverage was reached for each line. The genome was de novo assembled into 142,137 contigs with an average length of 407 bp. Blast analyses were carried out to compare the de novo assembled contigs to the UniGene set, and matching sequences were used as reference contigs. By applying the selection criteria of step 1 (Fig. 1) onto reference contigs, 616,781 SNPs were called from read mapping; of these, 532,613 SNPs were derived from WGS and 84,168 from amplicon sequences. 25,742 SNP markers passed all filtering steps and were usable for Illumina<sup>®</sup> Infinium array design: 11,042 from amplicon sequencing and 14,700 from the de novo assemblies. The inclusion of additional 10,640 pre-validated publically available SNPs (Bachlava et al. 2012) resulted in 36,382 markers for further processing. After removal of redundant markers, the SNP pool was reduced to a final size of 25,944 bi-allelic candidate SNPs having high design scores. In total, 3645 SNP markers (14 %) failed to meet bead representation and decoding quality metrics during the Illumina manufacturing process. For genotyping of 1090 sunflower samples and subsequent cluster file construction, 22,299 functional SNPs (Table S4) were used.

Finally, a set of 20,502 high-quality, bi-allelic SNPs was obtained that included 6393 publically available markers (Bachlava et al. 2012) and corresponded to an average density of one SNP per ~176 kb of the genome. In the final marker set, 18,990 (92.6 %) SNPs were polymorphic and 15,535 (75.8 %) SNPs had a minor allele frequency (MAF)  $\geq 10$  % when tested on 243 inbred lines, 5 landraces, 19 OPVs, and 20 introgression lines (Table S5). The lowest number of monomorphic SNPs could be observed for the inbred lines (7.7 %). In addition, 3.4 % of the markers detected rare alleles (MAF < 1 %) within the inbred lines of this sunflower diversity collection. The proportion of heterozygous calls per SNP denoted as observed heterozygosity ranged from 0.07 to 0.59 with inbred lines displaying the lowest and introgression lines the highest values (Table 1). Average nucleotide diversity per SNP was with 0.35 lowest for inbred lines and with 0.38 and 0.39 in a comparable range for landraces, OPVs and introgression lines, respectively. Only 50 SNPs (0.24 %) failed the test for Hardy–Weinberg equilibrium in landraces and OPVs.

**Table 1** Overview of diversity parameters

	Polymorph	MAF	Observed heterozygosity	Nucleotide diversity
Inbred lines	18,928	0.254	0.065	0.348
Landraces, OPVs	17,574	0.291	0.384	0.382
Introgression lines	16,921	0.304	0.587	0.387
All 287 genotypes	18,990	0.258	0.117	0.341

Average minor allele frequencies (MAFs), observed heterozygosity and nucleotide diversity of polymorphic SNPs in the sunflower diversity collection of 287 accessions are shown

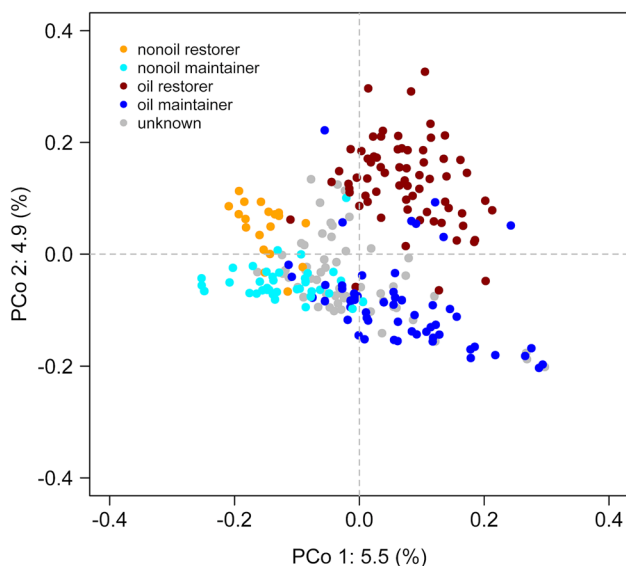
OPVs open-pollinating varieties

### Analysis of population structure and substructure

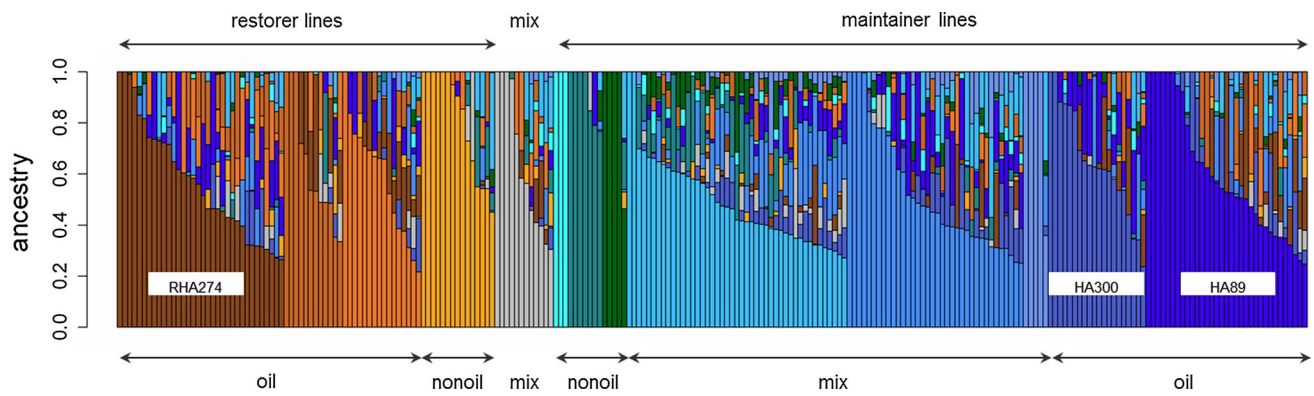
In order to demonstrate the applicability of a genotyping array, a high frequency of polymorphic variants in a representative set of genotypes is crucial. In the present diversity collection of 287 accessions, for a subset of 184 genotypes the affiliation to one of the four categories nonoil restorer ( $n = 19$ ), oil restorer ( $n = 68$ ), nonoil maintainer ( $n = 41$ ) and oil maintainer ( $n = 56$ ) was known. Within this subset of 184 categorized lines, 91.8 % of the SNP variants on the array were polymorphic and only 1677 SNPs were monomorphic indicating a very low false discovery rate during SNP identification. By assigning the genotypic data to the subsets which were separated according agronomic use and breeding history, 95.4 % of the SNPs were identified to be polymorphic for oil maintainer, 95.1 % for oil restorer

and 92.1 % for nonoil maintainer, respectively (Figure S2). Probably due to the small sample size and the higher degree of relationship indicated by the available pedigree information (USDA-ARS 2014; USDA 2006), only 78.3 % of the SNPs were polymorphic within the nonoil restorers.

The determination of population substructure is a key aspect for quantitative genetic or population genetic analyses since population stratification or admixture may affect detection of marker-trait associations, genomic prediction accuracy, or estimation of population genetic parameters. Principal coordinate (PCoA) and subpopulation structure analyses were performed to investigate the potential of the array to resolve population substructure in 243 sunflower inbred lines. When applying PCoA to the dataset, the first axis separated nonoil and oil lines. It explained 5.5 % of the observed variation within the set of inbred lines (Fig. 2) in accordance with a moderate level of differentiation between these two groups ( $F_{ST} = 0.116$ ). The second axis further subdivided restorer and maintainer lines explaining 4.9 % of the total variation (average  $F_{ST} = 0.056$ ). A high number of subgroups was observed for the 243 inbred lines based on the cross-validation errors calculated by ADMIXTURE. Errors were similar for the number of groups  $K = 11$  to 14 (0.696–0.702) with a minimum for  $K = 13$ . The population structure of the set of inbred lines is shown in Fig. 3 for  $K = 13$ . Our analysis separated the 243 lines into four subgroups belonging to the group of restorer lines and eight subgroups known as maintainer. A further subgroup was composed of restorer and maintainer lines. Within the restorer group, three subgroups were found to represent oil restorer. One subgroup contained nonoil restorer. Taking the known pedigrees into account, the largest restorer subgroup could be clearly assigned to progeny derived from line RHA274. In the group of maintainer lines, the majority of lines clustered into three subgroups that contained nonoil as well as oil maintainers. Clearly separated were two oil maintainer subgroups which comprised the offspring of HA300 (Peredovik 301) and HA89, respectively. The majority of inbred lines were strongly admixed highlighting the diversity within the panel. Only genotypes representing nonoil maintainer clustered into subgroups of very closely related lines.

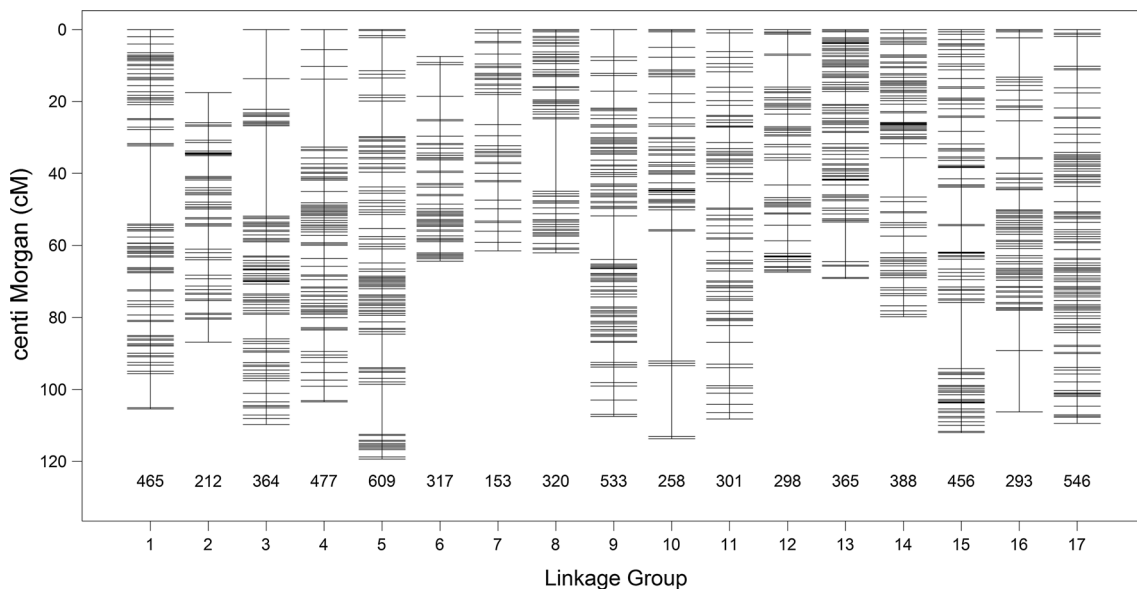


**Fig. 2** Genetic differentiation of 243 sunflower inbred lines. Association of lines as revealed by principal coordinate analysis based on Rogers' distances is presented. Nonoil maintainer, oil maintainer, nonoil restorer, oil restorer, and inbred lines of unknown origin are designated by different colors (color figure online)



**Fig. 3** Population substructure among 243 inbred lines. Identified subgroups are shown as revealed by ADMIXTURE for  $K = 13$ . Individuals are plotted on the x-axis and sorted in descending order

according to their subgroup assignment given at the bottom. Ancestry was plotted on the y-axis



**Fig. 4** Genetic map of the RIL population  $NDBLOS_{sel} \times CM625$  constructed based on 6355 high-quality polymorphic SNP markers. Below each linkage group (LG) the number of markers is presented

### Analysis of $NDBLOS_{sel} \times CM625$ segregating for resistance to *Sclerotinia* midstalk rot

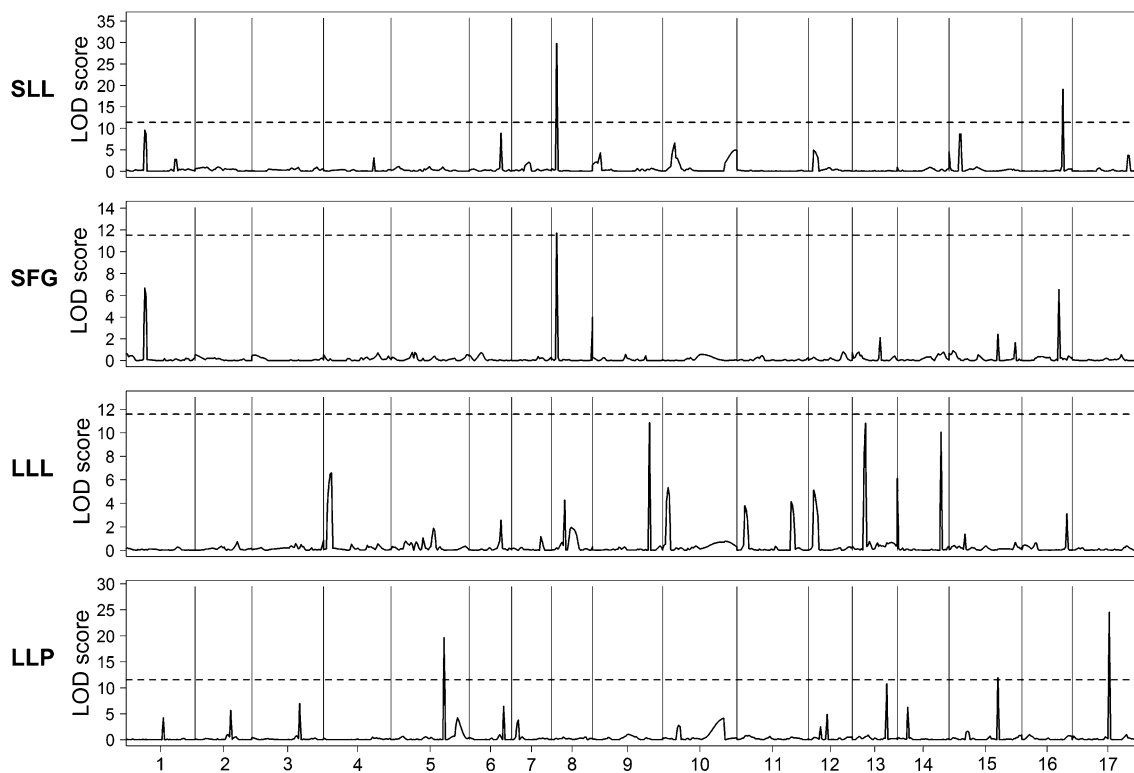
A genetic map of the RIL population  $NDBLOS_{sel} \times CM625$  was developed based on 6355 high-quality polymorphic markers (Table S3). The distribution of the SNPs (Fig. 4) was generally found to be even across the 17 LGs with the exception of the upper half of LG3 and the bottom half of LG10 where only few markers could be placed. This was probably because the genomes of the two parental lines are very similar in the respective regions and thus displayed a lower level of polymorphism. The frequency distributions of adjusted means for the four traits under study are shown in Figure S4. Resistance traits were

significantly correlated (0.40–0.65) with each other (Figure S4). There was no remarkable correlation between the morphological trait leaf length with petiole and the three resistance traits. In the QTL analysis, two, one, and three QTL were identified for stem lesion length, speed of fungal growth, and leaf length with petiole, respectively. Details on putative QTL, including positions in the genome, information of flanking markers, and QTL effects are given in Table 2. LOD scores along the genome for all four traits are shown in Fig. 5. The detected QTL explained 8.1–35.2 % of the phenotypic variance and exhibited small support intervals  $\leq 4$  cM. The largest proportion of phenotypic variance was explained by a QTL on LG8 affecting resistance traits leaf length with petiole and speed of fungal growth.

**Table 2** Characteristics of detected QTL

Trait	LG	Pos (cM)	Left and right flanking markers	Interval	LOD	R <sup>2</sup>	Add effect
SLL	8	8	S48333685-0375 (7.87 cM) SFW0383 (8.520 cM)	6–10	29.76	35.18	4.411
	16	62	HA015328-0033 (60.877 cM) HA013937-0089 (62.904 cM)	60–64	19.13	19.81	3.036
SFG	8	8	S48333685-0375 (7.870 cM) HA015520-0139 (8.520 cM)	6–10	11.71	14.42	0.038
LLP	5	80	SFW8046 (79.394 cM) HA010270-0404 (81.22 cM)	78–82	19.68	18.05	0.604
	15	74	HA011066-0415 (73.567 cM) SFW9147 (74.907 cM)	72–76	11.95	8.07	0.475
	17	56	HA010952-0467 (55.616 cM) HA012049-0209 (56.267 cM)	54–58	24.57	21.49	−0.652

QTL that were detected from composite interval mapping for stem lesion length (SLL), speed of fungal growth (SFG), and leaf lesion length with petiole (LLP). The position of the putative QTL (linkage group, LG; position in cM, Pos), the name of the flanking markers (left and right, including their positions), the corresponding support interval, the LOD of the QTL, the proportion of phenotypic variance explained, and its additive effect are presented. QTL were declared as significant when they exceeded an empirical LOD score threshold corresponding to an experimentwise  $p$  value of 0.05



**Fig. 5** LOD score profile from QTL mapping along the 17 linkage groups for the traits stem lesion length (SLL), speed of fungal growth (SFG), leaf lesion length (LLL), and leaf length with petiole (LLP).

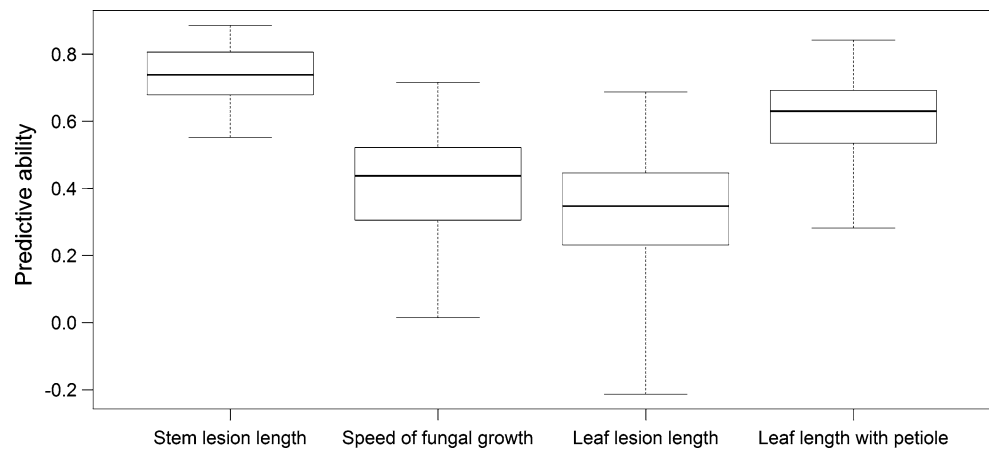
The dashed line represents the LOD threshold obtained by a permutation test according to Churchill and Doerge (1994) corresponding to a type I error rate of 5 %

For all resistance traits, the NDBLOS<sub>sel</sub> allele increased the *Sclerotinia* resistance. The leaf length with petiole increasing allele at the QTL on LG17 originated from NDBLOS<sub>sel</sub> while it was contributed by CM625 at the QTL on LG5 and LG15. Predictive abilities from the GBLUP model are

represented by boxplots for each of the four traits in Fig. 6. High predictive ability was observed for stem lesion length, with on average 0.74 ( $h^2 = 0.79$ ; accuracy = 0.83). For the other traits, predictive abilities were on a medium level with mean predictive ability ranging from 0.31 ( $h^2 = 0.51$ ;



**Fig. 6** Genome-based prediction of phenotypic traits. Box-plots showing the distribution of predictive abilities from ten times replicated fivefold cross-validation within the RIL population NDBLOS<sub>sel</sub> × CM625 obtained with GBLUP for stem lesion length, speed of fungal growth, leaf lesion length, and leaf length with petiole



accuracy = 0.43) for leaf lesion length, 0.41 ( $h^2 = 0.57$ ; accuracy = 0.54) for speed of fungal growth to 0.61 ( $h^2 = 0.63$ ; accuracy = 0.77) for the morphological trait. Predictive abilities for leaf lesion length were highly variable with a standard deviation of 0.18.

## Discussion

This study was aimed at the discovery of a large set of SNP polymorphisms within *Helianthus annuus* L. to generate an Illumina<sup>®</sup> Infinium iSelect HD BeadChip. For this, de novo assembled contigs were filtered for sequences that mapped to the *H. annuus* specific UniGene set prior to variant calling. In order to achieve an unbiased SNP set, two different approaches were used. In the first approach, the focus was on identifying as many SNPs as possible through genome sequencing of four sunflower lines. The drawback of this procedure was that only few lines could be sequenced at high coverage with reasonable costs so that the identified SNPs could not be fully representative for the entire cultivated gene pool. In the amplicon approach, we analyzed 48 lines so that the gene pool was more widely represented in terms of the observed allelic variation, especially since only haplotype-specific markers (one marker per observed haplotype in each amplicon) were selected for the array. The drawback here was that in this way only 3356 genes could be analyzed. In order to allow the data to be related to the draft sunflower genome sequence (Kane et al. 2011), SNP markers recently validated and genetically mapped by other groups were included (Bachlava et al. 2012; Bowers et al. 2012a). Due to the selection procedure, most of the SNPs were located near (in 5' and 3'-flanking sequences) genes or in exons and introns of those. Validation based on genotyping of inbred lines, OPVs, introgression lines, landraces, and RIL and F<sub>2</sub> populations resulted in 20,502 high-quality bi-allelic SNPs, each detecting a single locus in the sunflower genome. This number corresponded to 91.9 %

of the SNPs assayed and was comparable to scoring rates reported for other plant species during Illumina<sup>®</sup> Infinium assay design (Bianco et al. 2014; Dalton-Morgan et al. 2014; Song et al. 2013). The final set included 14,109 new SNP markers and 6393 publically available high-quality SNP markers (Bachlava et al. 2012). With that the developed 25 K array is the largest genotyping array that is currently available for routine sunflower genotyping since the array described by Bowers et al. (2012b) is, as described in their publication, too error-prone and expensive for routine SNP genotyping. The high overall polymorphism rate of 92.6 % depicted the quality of the SNP filtering procedure and is in line with results obtained by other studies regarding genotype array validation in animals and plants (Chen et al. 2014; Ramos et al. 2009; Tosser-Klopp et al. 2014; Unterseer et al. 2014). It further confirmed the utility of the array for applications in a wide range of sunflower germplasm.

Central applications of a genotyping array are the characterization of genetic variation and subpopulation structure in germplasm collections. Here, we found a high level of nucleotide diversity in accordance with a previous report by Mandel et al. (2013) and in line with the history of sunflower breeding. Elite sunflower inbred lines have been developed after passing through at least three major bottlenecks: breeding for oilseed traits, self-compatibility and self-pollination, and for hybrid seed production traits (Hongtrakul 1997). This breeding strategy could have resulted in a considerable decrease of diversity, but migration of OPVs and other exotic germplasm as well as selection of inbred lines from inter-pool crosses counteracted a strong reduction. Furthermore, in the recent past several wild species have become increasingly important as sources of disease resistance, drought tolerance, and other agronomically important traits (Jan and Chandler 1988; Miller 1987; Seiler 1992, 2010). The development of interspecific hybrids is often accompanied by a transfer of large segments of wild species genome into the respective

breeding line (Barb et al. 2014; DuBle et al. 2004; Qi et al. 2012), leading to broadening of genetic diversity but also to undesired linkage drag. Due to its high resolution, the new genotyping array offers the possibility to improve the targeted introgression and reduction of donor segments. In addition, it enables a better representation of the heterozygosity in a set of lines. We observed a three-fold higher level of heterozygosity for inbred lines, introgression lines, as well as landraces and OPVs compared to Mandel et al. (2013) which can be explained by the increased number of SNPs with a MAF below 10 % in our study.

In the last decades of the twentieth century, breeding of sunflower was focused on generating inbred lines and heterotic pools to maximize heterosis and improve traits essential for hybrid breeding (Miller 1987). In the long term, this strategy should result in distinct germplasm groups which exhibit maintenance of cytoplasmic male sterility (cms) in the seed parent pool and fertility restoration in the pollen parent pool. However, when classifying the genetic material into maintainer and restorer lines we observed only a moderate level of differentiation between the two primary sunflower breeding pools in accordance with Mandel et al. (2013). The low level of molecular variation explained by the first two principal coordinates in our study reflected a rather complex genetic composition of the investigated germplasm. Here, the differentiation between nonoil and oil lines was stronger compared to the separation between restorer and maintainer lines. A differentiation between nonoil and oil lines has been observed for restorer lines previously (Mandel et al. 2013), but was indicated by the second coordinate. However, between the two studies only 165 accessions overlapped, corresponding to 61 % (Mandel et al. 2013) and 68 % (this study) of the investigated lines. Moreover, the previous study based on 5.5 K SNPs compared to our set of 18.9 K polymorphic markers. These findings underline the demand for high-density genotypic data to resolve the population structure within the present sunflower collection. Indeed, the fine scale resolution of the array enabled us to uncover the presence of thirteen subgroups. These subgroups could generally be assigned to the maintainer and restorer group and were separated to a large extent regarding agronomic use (oil vs. nonoil), although the majority of inbred lines was characterized by a high degree of admixture. The new SNP array allowed fine scale resolution of ancestry identifying e.g. one clear subgroup formed by descendants of RHA274, a restorer line of the PET1 cms system, which was released in 1973 and represents a prominent parent of the pollen parent pool. Two further subgroups were formed by the progeny of HA89 and HA300, released in 1971 and 1976, respectively that co-founded the seed parent pool (Fick and Miller 1997; Miller 1997). Thus, the array developed here offers the possibility

to depict the allelic diversity of sunflower and to represent the breeding history with high resolution.

Marker-assisted selection based on results from QTL mapping studies or genome-based prediction of genetic values is expected to increase progress in plant breeding. Especially in resistance breeding, phenotyping is often not trivial as it requires the occurrence of pathogens in the field or expensive artificial infection methods. In the QTL analysis of a biparental population with 6355 polymorphic SNPs, we confirmed the majority of QTL detected by Micic et al. (2005). The slightly lower number of QTL identified in our study resulted from the more stringent LOD threshold applied to account for multiple testing which is crucial in large marker datasets. With a maximum of 4 cM, the LOD support intervals were strongly reduced compared to Micic et al. (2005). Thus, genotyping with the 25 K array allowed a better resolution of genomic regions involved in resistance to *Sclerotinia* midstalk rot and should enable successful marker-based selection for this trait. We hypothesize that the high marker density of this new array will be highly beneficial for QTL mapping in advanced mating designs such as multiparental mapping populations (Giraud et al. 2014). Furthermore, the 25 K array will constitute an essential tool for map-based cloning of genes associated with important agronomic traits of sunflower.

An alternative approach to predicting untested phenotypes based on individual QTL is whole genome-based prediction. With the development of high-density marker arrays, genome-based prediction has been successfully applied in a number of crops (Albrecht et al. 2011; Hefner et al. 2011; Hofheinz et al. 2012). So far, whole genome-based prediction studies for resistance traits in sunflower have been lacking, but studies on quantitative fungal or insect resistance conducted in maize (Technow et al. 2013), wheat (Daetwyler et al. 2014; Rutkoski et al. 2012), and barley (Lorenz et al. 2012) have shown its merit and applicability. Here, we assessed the performance of GBLUP to predict *Sclerotinia* midstalk rot in a biparental population genotyped with the newly developed 25 K array. We obtained high predictive abilities especially for the resistance trait stem lesion length (mean predictive ability = 0.74). For leaf lesion length and speed of fungal growth, predictive abilities were lower, but both traits also showed significantly lower trait heritabilities. Predictive abilities obtained in a biparental population need to be considered as an upper bound of what can be achieved in a breeding population. However, we consider the results presented here as a first indication that the potential of genome-based prediction of *Sclerotinia* midstalk rot resistance warrants further investigation.

The high number of SNP markers now available for sunflower opens new avenues for marker-based genetic studies and breeding. It will be particularly useful when

GWAS or genome-based prediction is applied to diverse datasets which require large marker densities to preserve the marker-QTL LD. The same holds true for genome-based prediction of sunflower hybrid performance. A first analysis based on a few hundred AFLP markers (Reif et al. 2013) did not have the power to predict sunflower hybrid performance with high accuracy. It remains to be shown if the availability of the 25 K SNP markers will overcome this limitation.

**Authors Contributions statement** This study was carried out in collaboration between all authors. CCS, MWG, MO, SW, RW, and ML conceived the study. VH and SW provided material. HL, MM, AP, JP, RW, CL, SU, and WE performed analyses. ML, MWG, and CCS drafted the manuscript. All authors read and approved the final version of the manuscript.

**Acknowledgments** This work has been supported by a grant (FKZ: 0315952A-D) from the German Federal Ministry for Education and Research (BMBF). We are grateful to USDA and INRA (Patrick Vincourt) for providing seeds of sunflower accessions used as core collection during this project. We thank Uwe Scholz and the Research group Bioinformatics and Information Technology at IPK Gatersleben for providing computational resources.

#### Compliance with ethical standards

**Conflict of interest** The authors MWG, JP, AP and HL have competing commercial interests as members of TraitGenetics GmbH which is a company that offers marker development and analysis (including this array) for commercial purposes. The authors MO and SW have competing commercial interests as members of KWS SAAT SE which is a sunflower breeding company. This does not alter the authors' adherence to sharing all data and materials. There are no further products in development or marketed products or patents to declare.

## References

- Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, Simianer H, Schön CC (2011) Genome-based prediction of testcross values in maize. *Theor Appl Genet* 123:339–350. doi:10.1007/s00122-011-1587-7
- Al-Chaarani GR, Gentzbittel L, Huang XQ, Sarrafi A (2004) Genotypic variation and identification of QTLs for agronomic traits, using AFLP and SSR markers in RILs of sunflower (*Helianthus annuus* L.). *Theor Appl Genet* 109:1353–1360. doi:10.1007/s00122-004-1770-1
- Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655–1664. doi:10.1101/gr.094052.109
- Bachlava E, Taylor CA, Tang S, Bowers JE, Mandel JR, Burke JM, Knapp SJ (2012) SNP discovery and development of a high-density genotyping array for sunflower. *PLoS ONE* 7:e29814. doi:10.1371/journal.pone.0029814
- Barb JG, Bowers JE, Renaut S, Rey JI, Knapp SJ, Rieseberg LH, Burke JM (2014) Chromosomal evolution and patterns of introgression in *Helianthus*. *Genetics* 197:969–979. doi:10.1534/genetics.114.165548
- Berrios EF, Gentzbittel L, Kayyal H, Alibert G, Sarrafi A (2000) AFLP mapping of QTLs for in vitro organogenesis traits using recombinant inbred lines in sunflower (*Helianthus annuus* L.). *Theor Appl Genet* 101:1299–1306. doi:10.1007/s001220051610
- Berry ST, Allen RJ, Barnes SR, Caligari PDS (1994) Molecular marker analysis of *Helianthus annuus* L. 1. Restriction fragment length polymorphism between inbred lines of cultivated sunflower. *Theor Appl Genet* 89:435–441. doi:10.1007/BF00225378
- Bianco L, Cestaro A, Sargent DJ, Banchi E, Derdak S, Di Guardo M, Salvi S, Jansen J, Viola R, Gut I, Laurens F, Chagne D, Velasco R, van de Weg E, Troggo M (2014) Development and validation of a 20 K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus × domestica* Borkh). *PLoS ONE* 9:e110377. doi:10.1371/journal.pone.0110377
- Bowers JE, Bachlava E, Brunick RL, Rieseberg LH, Knapp SJ, Burke JM (2012a) Development of a 10,000 locus genetic map of the sunflower genome based on multiple crosses. *G3 (Bethesda)* 2:721–729. doi:10.1534/g3.112.002659
- Bowers JE, Nambesan S, Corbi J, Barker MS, Rieseberg LH, Knapp SJ, Burke JM (2012b) Development of an ultra-dense genetic map of the sunflower genome based on single-feature polymorphisms. *PLoS ONE* 7:e51360. doi:10.1371/journal.pone.0051360
- Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Human Genetics* 84:210–223. doi:10.1016/j.ajhg.2009.01.005
- Burke JM, Lai Z, Salmaso M, Nakazato T, Tang S, Heesacker A, Knapp SJ, Rieseberg LH (2004) Comparative mapping and rapid karyotypic evolution in the genus *Helianthus*. *Genetics* 167:449–457. doi:10.1534/genetics.167.1.449
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7. doi:10.1186/s13742-015-0047-8
- Chen HD, Xie WB, He H, Yu HH, Chen W, Li J, Yu RB, Yao Y, Zhang WH, He YQ, Tang XY, Zhou FS, Deng XW, Zhang QF (2014) A high-density SNP genotyping array for rice biology and molecular breeding. *Mol Plant* 7:541–553. doi:10.1093/Mp/Sst135
- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* 138:963–971
- Daetwyler HD, Bansal UK, Bariana HS, Hayden MJ, Hayes BJ (2014) Genomic prediction for rust resistance in diverse wheat landraces. *Theor Appl Genet* 127:1795–1803. doi:10.1007/s00122-014-2341-8
- Dalton-Morgan J, Hayward A, Alamery S, Tollenaere R, Mason AS, Campbell E, Patel D, Lorenc MT, Yi B, Long Y, Meng J, Raman R, Raman H, Lawley C, Edwards D, Batley J (2014) A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. *Funct Integr Genomics* 14:643–655. doi:10.1007/s10142-014-0391-2
- Dekkers JC (2007) Marker-assisted selection for commercial crossbred performance. *J Anim Sci* 85:2104–2114. doi:10.2527/jas.2006-683
- Duñle CM, Hahn V, Knapp SJ, Bauer E (2004)  $P1_{ARG}$  from *Helianthus argophyllus* is unlinked to other known downy mildew resistance genes in sunflower. *Theor Appl Genet* 109:1083–1086. doi:10.1007/s00122-004-1722-9
- FAOSTAT (2013) <http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor>. Accessed 19 Jan 2015
- Fick GN, Miller JF (1997) Sunflower breeding. In: Schneiter AA (ed) *Sunflower technology and production*, vol 35. American Society of Agronomy, Madison, pp 395–439. doi:10.2134/agronmonogr35.c8

- Gentzbittel L, Perrault A, Nicolas P (1992) Molecular phylogeny of the *Helianthus* genus, based on nuclear restriction fragment length polymorphism (RFLP). *Mol Biol Evol* 9:872–892
- Gentzbittel L, Zhang YX, Vear F, Griveau Y, Nicolas P (1994) RFLP studies of genetic relationships among inbred lines of cultivated sunflower, *Helianthus annuus* L: evidence for distinct restorer and maintainer germplasm pools. *Theor Appl Genet* 89:419–425. doi:10.1007/BF00225376
- Gill N, Buti M, Kane N, Bellec A, Helmstetter N, Berges H, Rieseberg LH (2014) Sequence-based analysis of structural organization and composition of the cultivated sunflower (*Helianthus annuus* L.). *Genome Biol (Basel)* 3:295–319. doi:10.3390/biology3020295
- Gilmour AB, Gogel B, Cullis B, Thompson R (2009) ASReml User Guide Release 3.0. VSN International, Hemel Hempstead, UK
- Giraud H, Lehermeier C, Bauer E, Falque M, Segura V, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Schipprack W, Flament P, Melchinger AE, Menz M, Moreno-Gonzalez J, Ouzunova M, Charcosset A, Schön CC, Moreau L (2014) Linkage disequilibrium with linkage analysis of multiline crosses reveals different multi-allelic QTL for hybrid performance in the flint and dent heterotic groups of maize. *Genetics* 198:1717–1734. doi:10.1534/genetics.114.169367
- Goddard ME, Hayes BJ (2007) Genomic selection. *J Anim Breed Genet* 124:323–330. doi:10.1111/j.1439-0388.2007.00702.x
- Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53:325–338. doi:10.2307/2333639
- Grassa CJ, Baute GJ, Moyers B, Owens GL, Kane N, Langlade N, Gouzy J, Gill N, Staton SE, Nguyen TN, Hubner S, Bowers JE, Chaidir NI, Bergès H, King M, Lai Z, Bachlava E, Knapp S, Burke JM, Vincourt P, Rieseberg LH (2015) Reference genome of sunflower, line HA412, Version 1.4: a domesticated *Compilospes*. In: Paper presented at the PAG XXIII, San Diego, CA, January 10–14
- Habier D, Fernando RL, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397. doi:10.1534/genetics.107.081190
- Heesacker AF, Bachlava E, Brunick RL, Burke JM, Rieseberg LH, Knapp SJ (2009) Comparative mapping identifies multiple chromosomal rearrangements and duplications in the common and silverleaf sunflower genomes. *Plant Genome* 2:1–14. doi:10.3835/plantgenome2009.05.0015
- Heffner EL, Jannink JL, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4:65–75. doi:10.3835/plantgenome2010.12.0029
- Hofheinz N, Borchardt D, Weissleder K, Frisch M (2012) Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theor Appl Genet* 125:1639–1645. doi:10.1007/s00122-012-1940-5
- Hongtrakul V (1997) The development and analysis of sequence-based DNA markers in sunflower for DNA fingerprinting and candidate gene analysis. PhD, Oregon State University
- Hulke BS, Grassa CJ, Bowers JE, Burke JM, Qi L, Talukder ZI, Rieseberg LH (2015) A unified single nucleotide polymorphism map of sunflower (*Helianthus annuus* L.) derived from current genomic resources. *Crop Sci* 55:1696–1702. doi:10.2135/cropsci2014.11.0752
- INRA (2014) [http://lipm-helianthus.toulouse.inra.fr/Web/core/Core\\_collections\\_list.html](http://lipm-helianthus.toulouse.inra.fr/Web/core/Core_collections_list.html)
- Jan CC, Chandler JM (1988) Registration of a powdery mildew resistant sunflower germplasm pool, PM1. *Crop Sci* 28:1040. doi:10.2135/cropsci1988.0011183X002800060072x
- Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405. doi:10.1093/bioinformatics/btn129
- Kane NC, Gill N, King MG, Bowers JE, Berges H, Gouzy J, Bachlava E, Langlade NB, Lai Z, Stewart M, Burke JM, Vincourt P, Knapp SJ, Rieseberg LH (2011) Progress towards a reference genome for sunflower. *Botany-Botanique* 89:429–437. doi:10.1139/B11-032
- Kosambi DD (1944) The estimation of map distance from recombination values. *Ann Eugenics* 12:172–175
- Lai Z, Livingstone K, Zou Y, Church SA, Knapp SJ, Andrews J, Rieseberg LH (2005a) Identification and mapping of SNPs from ESTs in sunflower. *Theor Appl Genet* 111:1532–1544. doi:10.1007/s00122-005-0082-4
- Lai Z, Nakazato T, Salmaso M, Burke JM, Tang S, Knapp SJ, Rieseberg LH (2005b) Extensive chromosomal repatterning and the evolution of sterility barriers in hybrid sunflower species. *Genetics* 171:291–303. doi:10.1534/genetics.105.042242
- Lorenz AJ, Smith KP, Jannink JL (2012) Potential and optimization of genomic selection for Fusarium Head Blight resistance in six-row barley. *Crop Sci* 52:1609–1621. doi:10.2135/cropsci2011.09.0503
- Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP markers and their impact on plant breeding. *Int J Plant Genomics* 1–11. doi:10.1155/2012/728398
- Mandel JR, Dechaine JM, Marek LF, Burke JM (2011) Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L. *Theor Appl Genet* 123:693–704. doi:10.1007/s00122-011-1619-3
- Mandel JR, Nambeesan S, Bowers JE, Marek LF, Ebert D, Rieseberg LH, Knapp SJ, Burke JM (2013) Association mapping and the genomic consequences of selection in sunflower. *PLoS Genet* 9:e1003378. doi:10.1371/journal.pgen.1003378
- Manly KF, Cudmore RH Jr, Meer JM (2001) Map Manager QTX, cross-platform software for genetic mapping. *Mamm Genome* 12:930–932. doi:10.1007/s00335-001-1016-3
- Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Micic Z, Hahn V, Bauer E, Schön CC, Melchinger AE (2005) QTL mapping of resistance to *Sclerotinia* midstalk rot in RIL of sunflower population NDBLOS<sub>sel</sub> × CM625. *Theor Appl Genet* 110:1490–1498. doi:10.1007/s00122-005-1984-x
- Miller JF (1987) Sunflower. In: Fehr WR (ed) Principles of cultivar development, vol 2., Crop SpeciesWiley, New York, pp 626–668
- Miller JF (1997) Registration of cmsHA 89 (PEF1) cytoplasmic male-sterile, RPEF1 restorer, and two nuclear male-sterile (NMS 373 and 377) sunflower genetic stocks. *Crop Sci* 37:1984. doi:10.2135/cropsci1997.0011183X003700060063x
- Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res* 8:4321–4325
- Natali L, Cossu RM, Barghini E, Giordani T, Buti M, Mascagni F, Morgante M, Gill N, Kane NC, Rieseberg L, Cavallini A (2013) The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC Genom* 14:686. doi:10.1186/1471-2164-14-686
- Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290. doi:10.1093/bioinformatics/btg412
- Pegadaraju V, Nipper R, Hulke B, Qi L, Schultz Q (2013) *De novo* sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genom* 14:556. doi:10.1186/1471-2164-14-556
- Pérez Vich B, Berry ST (2010) Molecular Breeding. In: Hu J, Seiler G, Kole C (eds) Genetics, genomics and breeding of sunflower. CRC Press, Pullman
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and

- population-based linkage analyses. *Am J Hum Genet* 81:559–575. doi:10.1086/519795
- Qi LL, Seiler GJ, Vick BA, Gulya TJ (2012) Genetics and mapping of the *R<sub>11</sub>* gene conferring resistance to recently emerged rust races, tightly linked to male fertility restoration, in sunflower (*Helianthus annuus* L.). *Theor Appl Genet* 125:921–932. doi:10.1007/s00122-012-1883-x
- Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, Bendixen C, Churcher C, Clark R, Dehais P, Hansen MS, Hedegaard J, Hu ZL, Kerstens HH, Law AS, Megens HJ, Milan D, Nonneman DJ, Rohrer GA, Rothschild MF, Smith TP, Schnabel RD, Van Tassell CP, Taylor JF, Wiedmann RT, Schook LB, Groenen MA (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE* 4:e6524. doi:10.1371/journal.pone.0006524
- Reif JC, Zhao YS, Würschum T, Gowda M, Hahn V (2013) Genomic prediction of sunflower hybrid performance. *Plant Breed* 132:107–114. doi:10.1111/Pbr.12007
- Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science (New York, N Y)* 301:1211–1216. doi:10.1126/science.1086949
- Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink JL, Sorrells M (2012) Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. *Plant Genome* 5:51–61. doi:10.3835/plantgenome2012.02.0001
- Seiler GJ (1992) Utilization of wild sunflower species for the improvement of cultivated sunflower. *Field Crop Res* 30:195–230. doi:10.1016/0378-4290(92)90002-Q
- Seiler G (2010) Utilization of wild *Helianthus* species in breeding for disease resistance. In: International symposium “Sunflower breeding on resistance to diseases”, Krasnodar, Russia, 2010. All-Russia Research Institute of Oil Crops by V. S. Pustovoi (VNIIMK) and The International Sunflower Association (ISA), pp 37–51
- Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB (2013) Development and evaluation of SoySNP50 K, a high-density genotyping array for soybean. *PLoS ONE* 8:e54985. doi:10.1371/journal.pone.0054985
- Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460
- Talukder ZI, Gong L, Hulke BS, Pegadaraju V, Song Q, Schultz Q, Qi L (2014) A high-density SNP Map of sunflower derived from RAD-sequencing facilitating fine-mapping of the rust resistance gene *R12*. *PLoS ONE* 9:e98628. doi:10.1371/journal.pone.0098628
- Tang S, Yu JK, Slabaugh MB, Shintani DK, Knapp SJ (2002) Simple sequence repeat map of the sunflower genome. *Theor Appl Genet* 105:1124–1136. doi:10.1007/s00122-002-0989-y
- Technow F, Burger A, Melchinger AE (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3 (Bethesda)* 3:197–203. doi:10.1534/g3.112.004630
- Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, Donnadiu-Tonon C, Eggen A, Heuven HCM, Jamli S, Jiken AJ, Klopp C, Lawley CT, McEwan J, Martin P, Moreno CR, Mulsant P, Nabihoudine I, Pailhoux E, Palhiere I, Rupp R, Sarry J, Sayre BL, Tircazes A, Wang J, Wang W, Zhang WG, Consortium IGG (2014) Design and characterization of a 52 K SNP chip for goats. *PLoS ONE* 9. doi:10.1371/journal.pone.0086227
- Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3-new capabilities and interfaces. *Nucleic Acids Res* 40:e115. doi:10.1093/nar/gks596
- Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom TM, Fries R, Pausch H, Bertani C, Davassi A, Mayer KF, Schön CC (2014) A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genom* 15:823. doi:10.1186/1471-2164-15-823
- USDA (2006) Sunflower inbred lines. <http://www.ag.ndsu.nodak.edu/aginfo/seedstock/varieties/VH-SUNF.htm>
- USDA-ARS (2014) Germplasm resources information network. [http://www.ars-grin.gov/npgs/acc/acc\\_queries.html](http://www.ars-grin.gov/npgs/acc/acc_queries.html)
- Utz HF, Melchinger AE (2006) PLABQTL: a computer program to map QTL, version 1.2 edn. Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart
- van Ooijen JW (2006) JoinMap 4, software for the calculation of genetic linkage maps in experimental populations. Kyazma, B.V., Wageningen
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370. doi:10.2307/2408641
- Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy–Weinberg equilibrium. *Am J Hum Genet* 76:887–893. doi:10.1086/429864
- Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) Synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28:2086–2087. doi:10.1093/bioinformatics/bts335
- Yu J-K, Tang S, Slabaugh MB, Heesacker A, Cole G, Herring M, Soper J, Han F, Chu W-C, Webb DM, Thompson L, Edwards KJ, Berry S, Leon AJ, Grondona M, Olungu C, Maes N, Knapp SJ (2003) Towards a saturated molecular genetic linkage map for cultivated sunflower. *Crop Sci* 43:367–387. doi:10.2135/cropsci2003.3670