Haplotype-based association mapping complements SNP-based approaches as a powerful tool to analyze the genetic basis of complex traits

Kumulative Dissertation
zur Erlangung des Doktorgrades der Agrarwissenschaften
Doctor agriculturae (Dr. agr.)
der Naturwissenschaftlichen Fakultät III Agrar- und Ernährungswissenschaften,
Geowissenschaften und Informatik
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von
Frau Fang Liu
geboren am 23.11.1991 in Shaodong, Hunan, China

1. Gutachter: Prof. Dr. Jochen Reif
2. Gutachter: Prof. Tobias Würschum

Eingereicht am 16.09.2020
Verteidigt am 26.04.2021

不忘初心，砥砺前行

Don’t Forget Why We Started and Move Forward Bravely

---Is life always this hard, or is it just when you are a kid?

--- Always like this.

《***Léson: The Professional***》

Table of Contents

General Introduction 5

Approaches and mapping populations for dissecting the genetic architecture of traits 6

Basic principle for conducting GWAS in plant populations..8

Limitations of current GWAS approaches using single SNP 9

Advantages of GWAS using haplotypes..10

Approaches for constructing haplotypes 11

Haplotype based on sliding window 11

Haplotype defined based on linkage disequilibrium 12

Further haplotype-based approaches used for GWAS 12

Arabidopsis: a perfect model species for developing statistical methodology 13

The world's second cultivated cereal wheat and the important trait leaf rust 13

Benefits of hybrid mapping population for GWAS in wheat..15

High marker density based on exome capture sequencing in wheat..15

Objectives of this thesis...16

Peer-reviewed scientific articles..17

Selecting closely-linked SNPs based on local epistatic effects for haplotype construction improves power of association mapping. 17

Exome association analysis sheds light onto leaf rust (*Puccinia triticina*) resistance genes currently used in wheat breeding (*Triticum aestivum* L.) 18

Haplotype-based genome-wide association increases the predictability of leaf rust (*Puccinia triticina*) resistance in wheat. 19

General Discussion..20

Genetic variant for GWAS: SNP or haplotype?...20

Functional haplotype and other haplotype based GWAS approaches 23

Data mining considering population structure helps to increase the power of GWAS.............27

Outlook...29

Summary 30

Zusammenfassung...31

Reference...33

Abbreviations 42

Acknowledgements 43

Curriculum vitae..44

Eidesstattliche Erklärung / *Declaration under Oath*..47

Erklärung über bestehende Vorstrafen und anhängige Ermittlungsverfahren / *Declaration concerning Criminal Record and Pending Investigations* 48

General Introduction

In view of the growing world population predicted to reach over 9 billion by 2050, producing sufficient food to meet the demand of the rising population is an unprecedented challenge (Goddard, 2009; Godfray et al., 2010; Ray et al., 2013). In addition, the climate change and decreasing resources, such as water and agricultural land, will potentially restrict crop production. To overcome these challenges, scientists and breeders are developing and implementing tools to breed varieties of crops that provide higher productivity and better adaptability under environmental stress including drought, high temperature, poor soil, diseases and destructive pest (Barbier and Hochard, 2018; Gu et al., 2018; Lesk et al., 2016).

Plant breeding, an economical and environmentally friendly method, has been performed for thousands of years to improve plant-derived products used for human nutrition or feeding of domesticated animals (Hartung and Schiemann, 2014). The traditional plant breeding approaches mainly rely on crossing of crop plants and selecting improved cultivars, which is time-consuming and labor-intensive. Modern breeding methodologies with precise selection and reduced breeding time are considered as efficient approaches to achieve genetic improvement of target traits, such as genomic-aided breeding that will dramatically reduce the time required to identify cultivars, which express the desired trait in a breeding program (Voss-Fels et al., 2019; Wang et al., 2018). Although the production of grain have been boosted in some cropping systems, for instance yields have been increased over sevenfold in maize, the challenge of plant breeding remains regarding not only yield but also quality and tolerance to environmental stress owing to climate change (Breseghello and Coelho, 2013; Wang et al., 2020). Environmental stress will directly affect the growth of plants and finally result in the reduction of grain yield. Taking the influence of pathogens and pests for example, yield loss was estimated at a global level and per hotspot for maize (22.5%; 19.5–41.1%), rice (30.0%; 24.6–40.9%), wheat (21.5%; 10.1–28.1%), potato (17.2%; 8.1–21.0%) and soybean (21.4%; 11.0–32.4%) (Savary et al., 2019). In addition, despite the important discoveries reported by many studies, for most traits the predicted proportion of phenotypic variance explained by genetic scores remains considerably lower than the trait heritability (Kim et al., 2017). In terms of this, a deep understanding of the genetic architecture of important traits would provide knowledge to guide future genetic improvement and close the missing heritability gap.

Approaches and mapping populations for dissecting the genetic architecture of traits

Linkage mapping and association mapping have been widely applied to dissect the genetic architecture of complex traits in plants (Wurschum, 2012). Linkage mapping is a conventional mapping method to identify genomic regions co-segregating with a given trait using biparental populations. In contrast, association mapping, also known as linkage disequilibrium mapping, takes the advantage of historic linkage disequilibrium to associate phenotypes to genotypes utilizing a more diverse population. Both of the two approaches aim to identify quantitative trait locus (QTL) or markers closely linked to QTLs. The key difference between linkage and association mapping resides in whether recombination events occur in biparental populations or not. Thus, the strengths and weaknesses of different mapping approaches are linked with the structure of mapping population.

In general, linkage mapping utilizes biparental mapping populations. The mapping population derived from biparental crosses, such as recombinant inbred line (RIL, Fig.1a), has limited genetic diversity as only two parental lines are used as the population founders (da Silva et al., 2018; Jamann et al., 2015; Semagn et al., 2010). Additionally, the limited recombination events in biparental populations result in low mapping resolution, allowing the localization of QTL to 10-20 cM intervals. An interval of 10 cM in maize could contain 200 or more genes (Haberer et al., 2005; Xu et al., 2017).

Multi-parental mapping populations, including nested association mapping (NAM) and multi-parental advanced generation intercross (MAGIC) populations, were developed to overcome the limitation of biparental populations (Fig. 1b, 1c) (Huang et al., 2015; Meng et al., 2016; Yu et al., 2008). As a combination of RIL populations sharing one common founder line, NAM populations provide high power and resolution for QTL detection, while MAGIC populations present high resolution and genetic diversity owing to the intercrosses of multiple inbred founders. However, the construction of a multi-parental population requires many years and excessive effort.

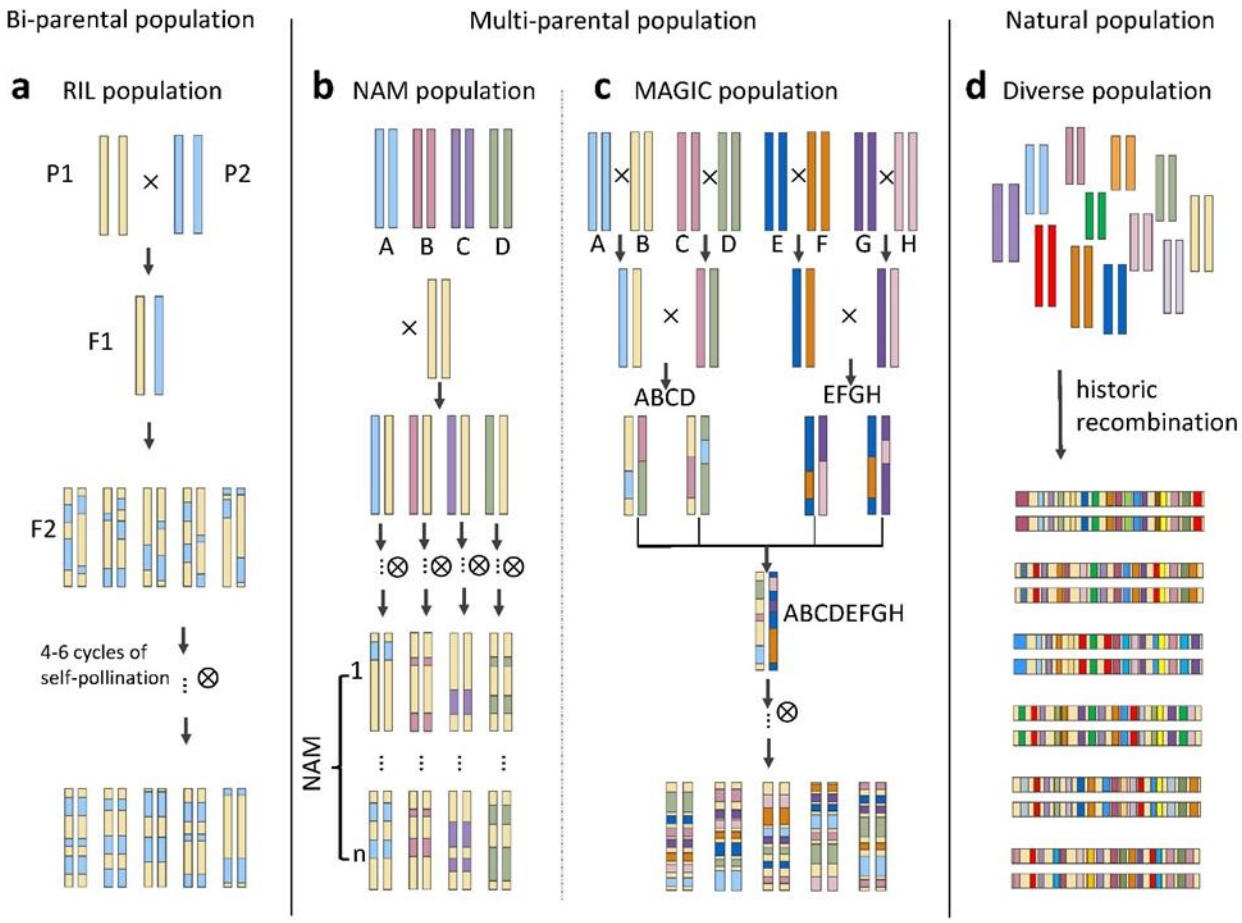


Fig. 1. Different mapping population. (**a**) recombinant inbred line (RIL), (**b**) nested association mapping (NAM), (**c**) multi-parental advanced generation intercross (MAGIC) and (**d**) diverse population.

Natural populations or diverse populations facilitate to take the advantages of abundant historic recombination and are the most prevalently used populations for association mapping (Fig. 1d), which provide due to the broad genetic variation a high mapping resolution. In addition, there is no need of the laborious development of a mapping population, which will save time and energy. As the next generation sequencing techniques advance and sequencing costs are reduced, high-density markers enable association mapping at the genome-wide level with very high resolution, even for direct identification of the underlying candidate genes (Juliana et al., 2018; Sapkota et al., 2019). Therefore, genome-wide association mapping (GWAS) has become a powerful tool for unraveling the molecular genetic basis underlying the phenotypic variation in many crops.

Basic principle for conducting GWAS in plant populations

In general, GWAS mainly consists of three steps (Fig. 2):

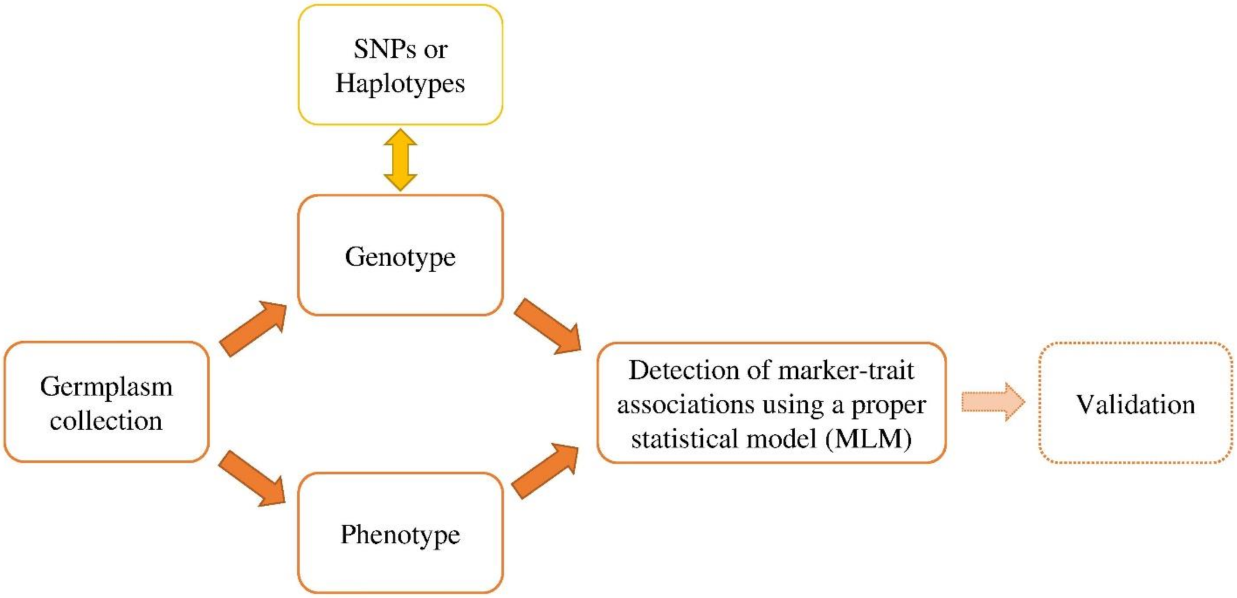

SNPs or Haplotypes
Genotype
Germplasm collection
Phenotype
Detection of marker-trait associations using a proper statistical model (MLM)
Validation

Fig. 2. A flowchart of genome-wide association mapping.

Firstly, a population with diverse genotypes is selected that can represent a range-wide collection of germplasms relative to a specific breeding. It can be a multi-parental population (NAM or MAGIC population) or diverse population as mentioned above.

Secondly, genetic variants are collected and phenotyped for the traits of interest. Currently, single-nucleotide polymorphisms (SNPs) are the most commonly used genetic variants and haplotypes are also used in association studies (this will be described in details later).

Thirdly, statistical analysis is performed to identify genetic variants associated to phenotypic variation. The population structure is prone to result in spurious associations because some neutral markers are significantly correlated with trait difference among subpopulations (Xu et al., 2017). Thus, an appropriate statistical model is critical to control the effect of population structure especially when the samples have very diverse levels of familial relatedness and complex population structure. Currently, the mixed linear model (MLM) suggested by Yu and coauthors

(Yu et al., 2006) is considered as a state-of-the-art method to correct the population structure and family relatedness. MLM can be simply described as follows:

$$y = X\beta + g + e,$$

where y is the vector of observed phenotypic values; β is the vector of fixed effects, such as the common intercept term, the tested marker effects and subpopulation effects, while X stands for the design matrix relating y to β; g is a vector of genotypic effects or polygenic background effects, and e indicates the residual term of the model. g and e are considered as random effects, with $g \sim N(0, \sigma_g^2 K)$ and $e \sim N(0, \sigma_e^2 I)$, where K is a marker-derived kinship matrix, I is the identity matrix, σ_{g}^2 and σ_{e}^2 are the genetic variance and residual variance, respectively.

Although not necessarily included in the standard GWAS procedure, validation is useful for preventing spurious associations so that the results can be generalized to other populations. However, biological or functional validations, which need to generate independent and appropriate population such as NIL populations, require substantial amount of time and are therefore often avoided (Navara and Smith, 2014). Validation using another independent diverse population may be an alternative choice to test whether the discoveries in one population are also effective in other independent populations and to decrease the false positive QTLs to some extent.

Limitations of current GWAS approaches using single SNP

The most commonly used genetic variants to test genotype-phenotype associations in GWAS are SNPs owing to the cost-effective genotyping with SNP arrays. Moreover, as the development in next generation sequencing techniques, the cost of whole genome sequencing (WGS) continues to decrease (Schneeberger, 2014). The utilization of high-density SNPs for GWAS could improve the mapping resolution, while it also results in a stringent level of significance accounting for multiple tests, which may cause false negatives (type II error). Besides, GWAS based on single markers often explain only a small proportion of the genetic variation: in general, single SNP variants contribute to less than 10% of the phenotypic variation for many complex traits (He et al., 2019b; Manolio et al., 2009). Additionally, only a modest proportion of the estimated heritability was explained by the SNP variants detected in GWAS for most complex traits (Manolio et al., 2009; Manolio, 2013). More importantly, complex traits are polygenic and frequently regulated by epistatic interactions (Doust et al., 2014; Jiang et al., 2017). In particular, epistatic effects among markers within small genomic regions are termed as local epistasis. It has long-term impact in plant breeding due to reduced chance to disappear after generations of recombination (Akdemir and Jannink, 2015; He et al., 2017), while it cannot be detected by the separate evaluation of each single SNP in GWAS. Thus, a more efficient method is required.

Advantages of GWAS using haplotypes

Haplotypes, a cluster or block of multiple SNPs, have been proposed to be applied in GWAS. Empirical studies have shown that haplotype-based GWAS were able to detect loci which failed to be identified in single SNP-based GWAS (Pryce et al., 2010; Tregouet et al., 2009). Furthermore, haplotype-based GWAS can dramatically reduce the number of tests because most haplotypes fall into a few classes within the regions of little evidence of recombination (Fig. 3) (Zhao et al., 2007). Importantly, haplotype can capture local epistatic effects between the SNPs, which could increase power and accuracy in dissecting complex traits (Bardel et al., 2005; Jiang et al., 2018). Many other literatures have revealed the advantages of using haplotypes for GWAS in human or animals. For instance, haplotypes provide more polymorphism information content compare to bi-allelic SNPs (N'Diaye et al., 2017); offer more information to estimate whether two alleles are identical by descent (Meuwissen and Goddard, 2000); clades of haplotype alleles capture information from evolutionary history (Templeton et al., 1987) and provide more power than single marker when an allelic series exists (Morris and Kaplan, 2002). These points together suggest that haplotypes provide potential prospect for the development of GWAS.

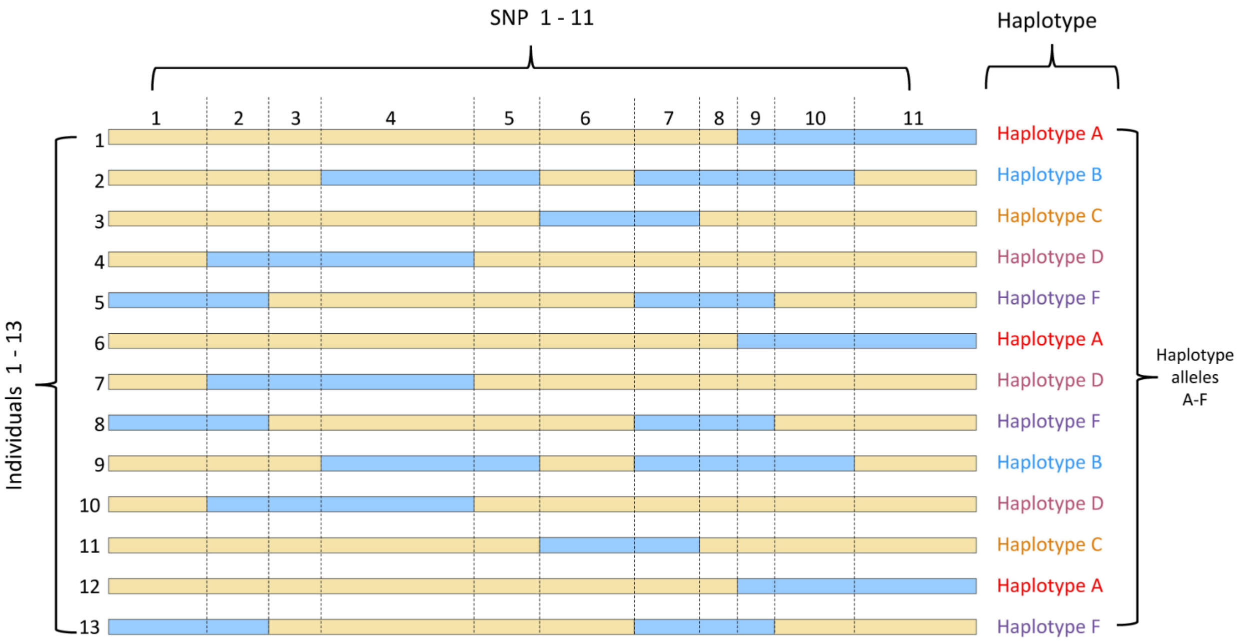


Fig. 3. A simple example to show how the use of haplotypes decreases the number of multiple tests compared to SNPs.

Approaches for constructing haplotypes

Several different approaches have been developed to define or construct haplotype blocks. The most widely used haplotypes for GWAS can be summarized as two types: haplotype based on sliding window or linkage disequilibrium (LD).

Haplotype based on sliding window

The simplest way to define haplotype blocks is the sliding window approach. Generally, the haplotype blocks are constructed by combining adjacent SNPs within a specific window size and the window moves with a certain step that could be smaller or equal to the window size (Huang et al., 2007; Lorenz et al., 2010). The study in barley is an example for the use of overlapping sliding windows with a fixed window size of three SNPs (Lorenz et al., 2010), while other studies attempt different window size ranging from 2 to 10 SNPs (Mathias et al., 2006; Pan et al., 2015). The size-fixed sliding window is easy to be conducted, however, it does not consider the various degrees of LD along the chromosome. Thus, variable-size sliding window approaches are proposed to cope with this problem. Model selection or exhaustive searching approaches have been applied to estimate the optimum window size at each locus (Li et al., 2007; Lin et al., 2004). A previous study has revealed that variable-size sliding windows can increase the power for the detection of disease variants compared to fixed-size sliding windows and single SNP, but with the cost of increased number of multiple tests and computation (Guo et al., 2009).

Haplotype defined based on linkage disequilibrium

Alternatively, haplotype can be constructed based on LD that is the nonrandom association of alleles at different sites, and related to the time of the mutation events and genetic distance. Several studies have found links between the LD pattern and haplotype blocks (Daly et al., 2001; Wall and Pritchard, 2003). Therefore, LD quickly becomes one of the most common approaches to define haplotype blocks. In total, three kind of LD-based approaches have been recommended. The first one is directly based on the measure of LD. The haplotype blocks are defined by grouping closely-linked SNPs that required at least a certain proportion of pairwise LD of those SNP passed a user-determined threshold (Gabriel et al., 2002). The second type is based on haplotype diversity. That is, at least a certain percentage of observed haplotypes must be common haplotypes, which means the frequency of haplotypes must be above a certain threshold. Meanwhile, the number of SNPs that distinguish the haplotypes in each block is required to be minimal. Although definition of measures based on LD and diversity are distinct, the two approaches are highly correlated because regions of low haplotype diversity typically exhibit high LD and vice versa (Cardon and Abecasis, 2003). In the third type of approach, the historical recombination events are considered in the construction of haplotype block. Precisely, SNPs are grouped into a block if there are no historical recombination events measured by the so-called four-gamete test (Wang et al., 2002).

Further haplotype-based approaches used for GWAS

The approaches mentioned above consider only consecutive SNPs, which are often in high LD, when constructing haplotype blocks. As a consequence, the haplotypes usually do not provide much more informative than a single SNP because the SNPs in high LD provide redundant information (Laramie et al., 2007). Another class of approaches that are not limited on adjacent SNPs select the most informative SNPs to generate haplotypes by using stepwise regression (Knuppel et al., 2012; Yang et al., 2008). Despite their potential, this class of approaches are restricted only to candidate gene regions instead of whole genome because of the high computational burden. Currently, most studies developing the haplotype-based GWAS approaches are based on binary traits (case and control) in human diseases, while the situation is different for quantitative traits. Moreover, many studies developing the haplotype-based methods were based on simulations under specific assumptions or ideal situation. However, these cases may not hold in real-world studies owing to much more complex genetic background (Liu et al., 2008).

Arabidopsis: a perfect model species for developing statistical methodology

Arabidopsis (*Arabidopsis thaliana*) as a model species, is one of the first non-human organisms used for GWAS, which demonstrates the feasibility of GWAS in plants (Atwell et al., 2010). Besides, Arabidopsis is considered as the almost ideal organism to conduct GWAS because it maintains inbred lines via continued self-fertilization, thus it is possible to study different phenotypes using genetically identical individuals (Korte and Farlow, 2013b). Over the past years, substantial GWAS studies have been conducted in terms of hundreds of phenotypes including the landmark GWAS study of 107 phenotypes (Atwell et al., 2010) and numerous other traits (Chan et al., 2011; Chao et al., 2012; Filiault and Maloof, 2012; Francisco et al., 2016; Slovak et al., 2015). As the release of abundant high-quality genotypic and phenotypic data, including the 1001 Genomes Project for Arabidopsis (1001 Genomes Consortium, 2016; Weigel and Mott, 2009) and the phenotypic database AraPheno (Seren et al., 2017), Arabidopsis has played a major role in developing appropriate statistical methodologies for GWAS (Korte et al., 2012; Sato et al., 2019; Segura et al., 2012; Song et al., 2018). The developed methodologies serve as a toolbox to unravel the genetic mechanism of important agronomical traits in various crop species.

The world's second cultivated cereal wheat and the important trait leaf rust

Wheat (*Triticum aestivum* L.) is the world's second most cultivated cereal after maize and provides one-fifth of the calorie intake of human population. However, conducting GWAS or genome analysis in wheat is challenging due to its huge hexaploid genome (17Gb, 2n=6x) that originates from three ancestral diploid genomes (A, B and D) (Sukumaran and Yu, 2014). Recently, the published, high-quality and well annotated wheat reference genome (IWGSC et al., 2018) facilitates the implementation and interpretation of GWAS in wheat. Subsequently, the genetic basis of yield-related traits was reported in several GWAS studies (Alqudah et al., 2020; Lujan Basile et al., 2019; Tsai et al., 2020). Besides, scientists and breeders also paid much attention to the disease-related traits due to their severe negative influence on grain yield, such as the resistance of leaf rust, yellow rust, powdery mildew and fusarium head blight. Among them, leaf rust (caused by *Puccinia triticina*) is one of the most common and widespread wheat diseases, which can cause up to 40% loss of wheat yield mainly by reducing kernel weight and decreasing the number of kernels per spike (Khan et al., 2013).

Many scientists and breeders have been studying the genetic architecture of leaf rust resistance in order to efficiently increase the resistance level of cultivars. For instance, about 90 resistance QTLs have been described in wheat (Kassa et al., 2017; Yang and Liu, 2004) and a few leaf rust resistance genes (*Lr*) have been cloned (listed in Table 1). Nevertheless, most of the identified leaf rust resistance QTLs are race-specific, which are only effective against particular pathogen isolates and thus prove to be ineffective after a few years of introduction because of the high mutation rate or virulence dynamics of pathogen populations (Lowe et al., 2011; McCallum et al., 2016). Moreover, most of previous QTL mapping studies rely on bi-parental populations, in which the low mapping resolution prevented researchers from precisely mapping candidate resistance genes (Xu et al., 2017) as well as developing reliable functional markers for marker-assisted selection.

Table 1. Cloned resistance genes of leaf rust.

Genes	Chromosome	Pathogen[1]	Type	Reference
Lr1	5DL	*Pt*	Race-specific	(Cloutier et al., 2007)
Lr10	1AS	*Pt*	Race-specific	(Feuillet et al., 2003)
Lr21	1DL	*Pt*	Race-specific	(Huang et al., 2009)
Lr22a	2DS	*Pt*	Race-specific	(Thind et al., 2017)
Lr34	7DS	*Pt, Pst, Pgt, Bgt*	Non-race-specific	(Krattinger et al., 2011)
Lr67	4DL	*Pt, Pst, Pgt, Bgt*	Non-race-specific	(Moore et al., 2015)

[1]: leaf rust (*Puccinia triticina*; *Pt*), stripe rust (*Puccinia striiformis* f. sp. *tritici*; *Pst*), stem rust (*Puccinia graminis* f. sp. *tritici*; *Pgt*) and powdery mildew (*Blumeria graminis* f. sp. *tritici*; *Bgt*)

Although a few GWAS have been performed in wheat using single SNP-based method for deep genetic analysis of leaf rust resistance (Juliana et al., 2018; Sapkota et al., 2019), the haplotype-based method has not been conducted yet. Moreover, haplotype-based approach can detect the local epistasis that may play a critical role on resistance of leaf rust but cannot be detected by single SNP-based GWAS. Thus, haplotype-based GWAS method should provide an insight into genetic basis of resistance to leaf rust, which will benefit marker-assisted selection of leaf rust resistance and represent promising targets to clone novel resistance genes.

Benefits of hybrid mapping population for GWAS in wheat

In a number of crop species, notably rice and maize, hybrid cultivars present higher grain yield compared to their parents owing to heterosis. For wheat, the studies have shown that hybrids perform better in terms of economic yield and yield stability (Muhleisen et al., 2014; Whitford et al., 2013). However, the mechanisms involved in the hybrid performance cannot be completely revealed by the GWAS based on inbred populations. Thus, in addition to diverse inbred population, hybrid population as a new population type is suitable for GWAS in crops exhibiting hybrid vigor, allowing GWAS to detect not only additive effects but also dominance effects (Mirdita et al., 2015). Recently, numerous traits have been successfully studied using hybrids population in wheat, including male floral traits, anther extrusion and disease resistance (powdery mildew, leaf rust and stripe rust) (Boeven et al., 2016; Gowda et al., 2014; Muqaddasi et al., 2016).

High marker density based on exome capture sequencing in wheat

To achieve high mapping resolution in GWAS, it is necessary to genotype the mapping population with a high-density marker panel. Given the advances in the next generation sequencing techniques and the reduced sequencing costs, WGS has been suggested as a method to characterize the genetic variants of mapping populations (Davey et al., 2011). Nevertheless, the cost of WGS in wheat remains high due to the large genome and its allohexaploid nature, making it necessary to have enough coverage to distinguish homologs and homeologs (Appels et al., 2018). Exome capture sequencing is an alternate solution to dramatically reduce sequencing costs by focusing on gene coding regions (Mo et al., 2018). The potential of using exome capture sequencing has been demonstrated in a pioneering study in wheat where genes underlying wheat improvement and environmental adaptation could be identified (He et al., 2019a).

Objectives of this thesis

Most GWAS studies almost invariably used single point analysis and haplotype approaches are neglected in plant analysis. However, various rationales for testing associations between phenotypes and haplotypes rather than single SNPs have been proposed. Thus, the main goal of the present PhD work was to develop a novel haplotype-based association mapping approach and to use experimental data sets in combination with state-of-the art approaches and study its potential and limits. This work covers the following specific objectives:

1) To develop a new approach of haplotype-based GWAS by considering additive and local epistatic interaction effects (called functional haplotype-based GWAS, FH-based GWAS). To compare the potential of FH-based GWAS to traditional single SNP-based GWAS as well as two other haplotype-based GWAS (haplotype based on sliding window and LD) using the model species Arabidopsis and the complex trait of flowering time. In addition, simulation study was performed to gain insight under which circumstances the FH-based GWAS outperforms SNP-based GWAS (Liu et al., 2019).
2) Because Arabidopsis is a model system with no economic value, the FH-based GWAS is prospected to be used in crops (wheat). SNP-based GWAS, as a standard method of GWAS, always play the role of benchmark for other methods. Thus, to understand the potential of FH-based GWAS in wheat, we firstly attempted to maximize the power of SNP-based GWAS with the important trait of leaf rust in HYWHEAT dataset by considering the influence of subpopulation structure (Liu et al., 2020a).
3) The potential of FH-based GWAS was compared to SNP-based GWAS in the aspect of power and phenotypic variance explained by the significant associations, using the same data set of leaf rust in HYWHEAT population. The validation of traits-marker associations in an independent population to reduce false positive results and reveal stability of different approaches. In addition, to understand whether the haplotype-based approach would benefit in closing the missing heritability gap, the predictability was compared between associated haplotypes and SNPs (Liu et al., 2020b).

Peer-reviewed scientific articles

Selecting closely-linked SNPs based on local epistatic effects for haplotype construction improves power of association mapping.

Published in:

G3: GENES, GENOMES, GENETICS (2019)

DOI: 10.1534/g3.119.400451

Authors: Fang Liu, Renate H. Schmidt, Jochen C. Reif and Yong Jiang

The original paper has been published and available online:

https://www.g3journal.org/content/9/12/4115

G3 Genes | Genomes | Genetics

INVESTIGATION

Selecting Closely-Linked SNPs Based on Local Epistatic Effects for Haplotype Construction Improves Power of Association Mapping

Fang Liu, Renate H. Schmidt, Jochen C. Reif,[1] and Yong Jiang[1]
Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, 06466 Stadt Seeland, Germany
ORCID IDs: 0000-0002-3774-552X (F.L.); 0000-0002-8037-3581 (R.H.S.); 0000-0002-2824-677X (Y.J.)

ABSTRACT Genome-wide association studies (GWAS) have gained central importance for the identification of candidate loci underlying complex traits. Single nucleotide polymorphism (SNP) markers are mostly used as genetic variants for the analysis of genotype-phenotype associations in populations, but closely linked SNPs that are grouped into haplotypes are also exploited. The benefit of haplotype-based GWAS approaches *vs.* SNP-based approaches is still under debate because SNPs in high linkage disequilibrium provide redundant information. To overcome some constraints of the commonly-used haplotype-based GWAS in which only consecutive SNPs are considered for haplotype construction, we propose a new method called functional haplotype-based GWAS (FH GWAS). FH GWAS is featured by combining SNPs into haplotypes based on the additive and epistatic effects among SNPs. Such haplotypes were termed functional haplotypes (FH). As shown by simulation studies, the FH GWAS approach clearly outperformed the SNP-based approach unless the minor allele frequency of the SNPs making up the haplotypes is low and the linkage disequilibrium between them is high. Applying FH GWAS for the trait flowering time in a large *Arabidopsis thaliana* population with whole-genome sequencing data revealed its potential empirically. FH GWAS identified all candidate regions which were detected in SNP-based and two other haplotype-based GWAS approaches. In addition, a novel region on chromosome 4 was solely detected by FH GWAS. Thus both the results of our simulation and empirical studies demonstrate that FH GWAS is a promising method and superior to the SNP-based approach even if almost complete genotype information is available.

KEYWORDS
haplotype
epistasis
genome-wide association study

Genome-wide association studies (GWAS) have been widely applied to identify candidate regions on chromosomes influencing complex traits in plant (Brachi *et al.* 2011), animal (Goddard and Hayes 2009) and human populations (McCarthy *et al.* 2008). The most commonly used genetic variants to test genotype-phenotype associations in GWAS are single nucleotide polymorphism (SNP) markers. Alternatively, SNPs can be combined into haplotypes which has been popular in association studies since the structure of human haplotype blocks was revealed (Gabriel *et al.* 2002; Cardon and Abecasis 2003). Empirical studies showed that haplotype-based GWAS was able to detect loci which failed to be identified in single SNP-based GWAS (Trégouët *et al.* 2009; Pryce *et al.* 2010). Nonetheless, contrasting results comparing the power of haplotype- and SNP-based GWAS were reported in previous studies (Lorenz *et al.* 2010) and whether it is beneficial to use haplotypes as variants in GWAS has to be evaluated on a case-by-case basis (Long and Langley 1999).

Potential advantages for testing associations between phenotypes and haplotypes, instead of SNP markers include: haplotypes may exploit epistatic interactions among markers within the haplotype blocks (Schaid 2004); contain more information on whether two alleles are identical by decent (Meuwissen and Goddard 2000);

Copyright © 2019 Liu *et al.*
doi: https://doi.org/10.1534/g3.119.400451
Manuscript received June 17, 2019; accepted for publication October 8, 2019; published Early Online October 11, 2019.
This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Supplemental material available at figshare: https://doi.org/10.25387/g3.8967986.
[1] Corresponding authors: Department of Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstr. 3, 06466 Stadt Seeland, Germany. E-mail: reif@ipk-gatersleben.de; jiang@ipk-gatersleben.de

utilize the information from evolutionary history (Durrant *et al.* 2004) and provide more power than single SNPs when multiple alleles contribute to the trait (Morris and Kaplan 2002). There are, however, also drawbacks when using haplotypes as variants in association tests. Adding irrelevant markers to a possible causal genetic variant will dilute the contrasts among haplotype allele classes (Clark 2004). A haplotype consisting of k SNPs may have up to 2^k different haplotype alleles, which will increase the degree of freedom and hence reduce the power of test (Zhao *et al.* 2003).

Among factors affecting the power of haplotype-based GWAS approaches, a fundamental one is how the haplotypes are constructed. The widely used methods group SNPs by sliding-windows of fixed or variable length (Lin *et al.* 2004; Huang *et al.* 2007), by the linkage disequilibrium (LD) between adjacent SNPs (Barrett *et al.* 2005) or by the diversity of haplotypes across samples (Zhang *et al.* 2002; Anderson and Novembre 2003). Common to these methods is that only consecutive SNPs, which are often in high LD, are combined into haplotypes. Consequently, in many cases the haplotypes are not much more informative than a single SNP because the SNPs in high LD provide redundant information (Laramie *et al.* 2007). This may provide one explanation for the contradicting results reported in the literature comparing the power of haplotype- and SNP-based GWAS approaches. Other methods were developed to search for haplotypes consisting of most informative and possibly non-consecutive SNPs within a certain region (Laramie *et al.* 2007; Yu and Schaid 2007; Abo *et al.* 2008; Yang *et al.* 2008; Dai *et al.* 2009; Knuppel *et al.* 2012). Despite their potential, the high computational burden associated with these methods restricted their use mainly to association studies for candidate gene regions.

In this study we addressed these limitations by developing a new method of constructing haplotypes, taking epistatic effects among SNPs into account. Epistasis has been identified as an important contributor to the genetic variation of complex quantitative traits (Carlborg and Haley 2004; Mackay 2014). It has been reported for two- or three-locus examples that a model involving haplotype effects can be reparametrized into one including the main and epistatic effects among markers constituting the haplotypes (Conti and Gauderman 2004; Schaid 2004). More recently this relationship between haplotype and marker effects was formally proved in the framework of genome-wide prediction for homozygous populations (Jiang *et al.* 2018). Capitalizing on these theoretical findings, we exploited epistatic effects among markers for constructing haplotypes and implemented this novel strategy in haplotype-based GWAS for a large *Arabidopsis thaliana* population generated by the 1001 Genomes Consortium (1001 Genomes Consortium 2016). The results were compared to those obtained with two commonly used haplotype-based GWAS methods as well as the single SNP-based approach, and underlined the ability to detect hidden marker-trait associations using the newly devised strategy. Moreover, simulation studies revealed factors which determine whether the developed GWAS approach outperforms the single SNP-based method.

MATERIALS AND METHODS

Throughout the manuscript, a combination of (possibly non-consecutive) SNPs was termed haplotype. By haplotype effect we meant to consider the effects of all possible alleles together. When referring to a specific allele, the term haplotype allele was used.

The baseline model for genome-wide association mapping

A standard linear mixed model controlling the structure of genetic relatedness or the polygenic background effects (Yu *et al.* 2006) was used for genome-wide association mapping. In this study the model was used for testing single SNP effects, epistatic effects among several SNPs and haplotype effects. It can be uniformly described as following:

$$y = 1_n\mu + X\beta + g + e \qquad (1)$$

where y is a n-dimensional vector of observed phenotypic values (n is the number of genotypes), 1_n is a vector of one's, μ is a common intercept term, β represents the effects of the variables (SNPs, interactions among SNPs or haplotype alleles) being tested, X stands for the corresponding design matrix, g is the n-dimensional vector of genotypic effects and e is the residual term. In the model we assume that μ and β are fixed effects, g and e are random effects and $g \sim N(0, \sigma_g^2 K)$, $e \sim N(0, \sigma_e^2 I)$, where K is a marker-derived kinship matrix, I is the identity matrix, σ_g^2 and σ_e^2 are the corresponding variance components. Distance matrix was calculated with Rogers' distance (Reif *et al.* 2005) and K was equaling one minus distance matrix. To reduce the computational load, an acceleration algorithm was implemented in which the linear mixed model was transferred to a simple linear model by applying eigen-decomposition to the kinship matrix (Lippert *et al.* 2011). The significance of β was assessed by t-test.

The general procedure of functional haplotype-based GWAS (FH GWAS)

In genomic prediction, it was demonstrated that modeling haplotype effects is equivalent to modeling main and epistatic effects among markers within the haplotype block, except that the two models assume different covariance structures for the unknown parameters (Jiang *et al.* 2018). The theory also applies to GWAS and in this case the two models are strictly equivalent because the parameters to be tested are assumed to be fixed effects (Equation 1) and hence without any covariance structure. Based on this theory, we developed a new haplotype-based GWAS approach, FH GWAS, with haplotypes based on the main and epistatic effects among SNPs. FH GWAS consists of the following four steps summarized in Figure 1.

Step 1: Preselecting SNPs to be combined into haplotypes: GWAS for single SNPs is performed using the linear mixed model (Equation 1) and a mild threshold without correction for multiple testing is applied to identify candidate SNPs (*e.g.*, $P < 0.01$). SNPs whose P values do not pass the threshold are excluded in subsequent analyses.

Step 2: Constructing functional haplotypes: In this step, candidate SNPs showing significant local epistatic effects are grouped into haplotypes. First we need to determine two parameters: the window size for searching haplotypes (denoted by w) and the number of SNPs in each haplotype (denoted by l). Theoretically the choice of these two parameters can be arbitrary. But in practice one needs to consider the linkage disequilibrium in the population, the computational load and the power of the association test (More details were discussed in the Discussion section). Once the parameters are chosen, GWAS model (Equation 1) is then performed for any l-tuple of SNPs within the window size w, including the additive effects of each SNP and the digenic epistatic effects for each pair of SNPs. That is, the entries in the vector β include a_i $(1 \le i \le l)$ and aa_{ij} $(1 \le i < j \le l)$, where a_i denotes the additive effect of the i-th SNP, aa_{ij} denotes the epistatic effects between the i-th and the j-th SNP. In the case that l is small, higher-order epistatic effects can also be included in the model. Next we determine the number of significant additive and epistatic effects (again under a mild threshold) required for grouping the l-tuple of

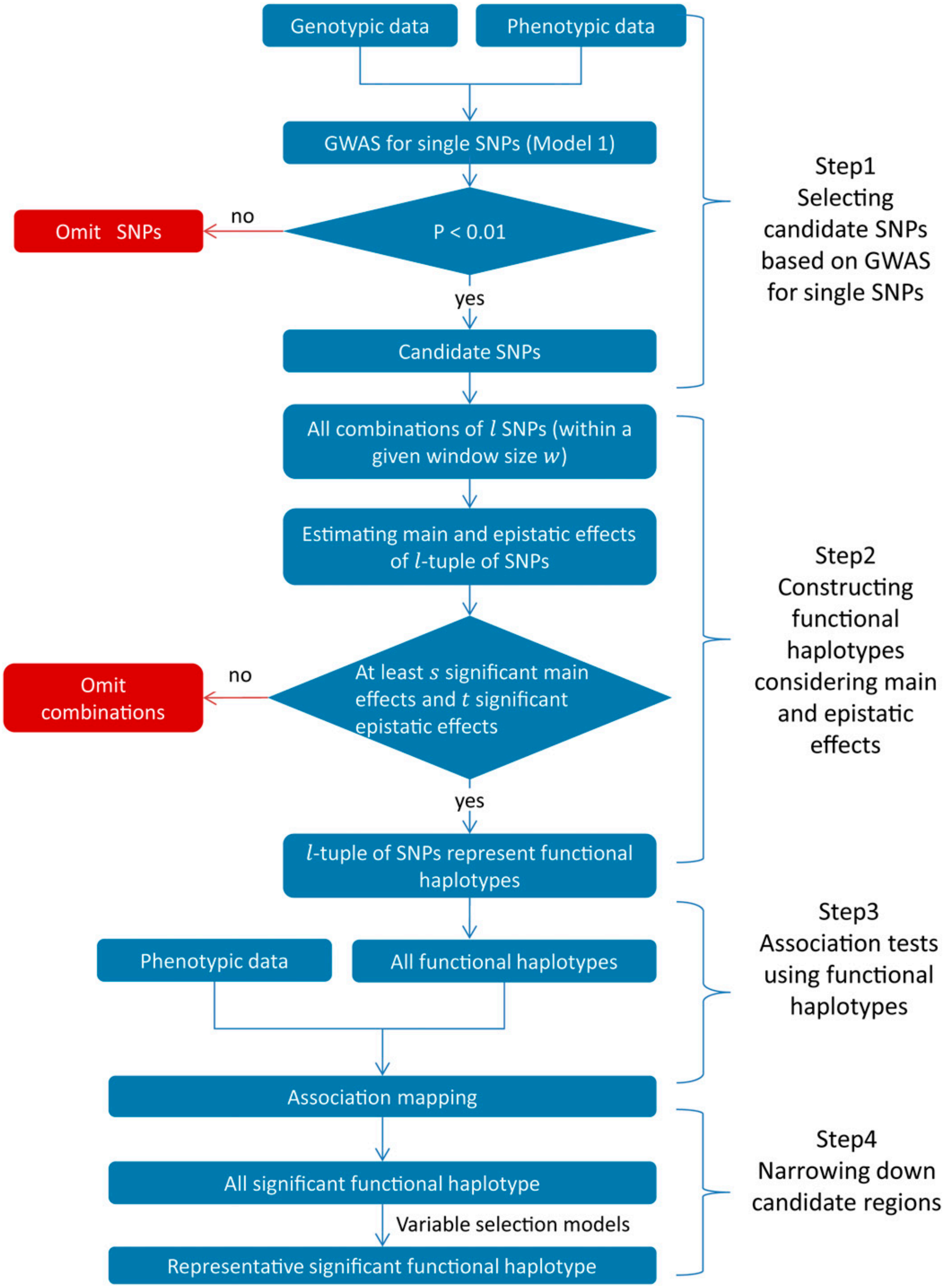

Genotypic data
Phenotypic data
GWAS for single SNPs (Model 1)
Omit SNPs
no
P < 0.01
yes
Candidate SNPs
Step1
Selecting
candidate SNPs
based on GWAS
for single SNPs
All combinations of l SNPs (within a given window size w)
Estimating main and epistatic effects of l-tuple of SNPs
Omit combinations
no
At least s significant main effects and t significant epistatic effects
yes
l-tuple of SNPs represent functional haplotypes
Step2
Constructing
functional
haplotypes
considering main
and epistatic
effects
Phenotypic data
All functional haplotypes
Step3
Association tests
using functional
haplotypes
Association mapping
All significant functional haplotype
Variable selection models
Representative significant functional haplotype
Step4
Narrowing down
candidate regions

Figure 1 The workflow for FH GWAS.

SNPs into a haplotype, *i.e.*, when there are at least s ($s \leq l$) significant additive effects and t ($t \leq l(l-1)/2$) significant epistatic effects, the l-tuple of SNPs is combined into a haplotype. Since main as well as epistatic effects are taken into account for haplotype construction we coined the term functional haplotypes. Importantly, for each trait to be analyzed in a given population, a different set of candidate SNPs and functional haplotypes will be obtained.

Step 3: GWAS using functional haplotypes: All resulting functional haplotypes are applied in conjunction with phenotypic data for GWAS using the linear mixed model (Equation 1). Significant functional haplotypes are then identified using a stringent genome-wide threshold corrected for multiple testing, *e.g.*, $P < 0.05$ after Bonferroni (Dunn 1961) or Benjamini-Hochberg correction (Benjamini and Hochberg 1995).

Step 4: Narrowing Down candidate regions: In each region in which significant functional haplotypes are detected in Step 3, we fitted all significant functional haplotypes in a variable-selection model (*e.g.*, the stepwise linear regression model (Draper and Smith 2014) or the least absolute shrinkage and selection operator (LASSO, Tibshirani 1996) to select representative significant functional haplotypes. In any region where significant functional haplotypes are found, the span of all representative haplotypes is considered as a final candidate region.

Two other methods constructing haplotypes

To compare FH GWAS with existing haplotype-based GWAS approaches, we considered the following two commonly-used methods of constructing haplotypes (Figure S1).

The overlapping sliding-window approach: The sliding window approach constructs haplotypes with a fixed window length, *i.e.*, the number of adjacent SNPs (Huang *et al.* 2007). If the window is moved with a certain step size which is smaller than the window length adjacent windows are overlapping. In this study we chose the window length to be three, which is consistent with the length of our functional haplotypes, and the step size to be one.

The linkage disequilibrium approach: The linkage disequilibrium (LD) approach groups SNPs into a haplotype if the LD between every two adjacent SNPs is equal or greater than a certain threshold, which allows physically close and non-randomly associated SNPs to be grouped together in the same haplotype (Barrett *et al.* 2005). In our study, the r^2 statistic was used to measure LD (Hill and Robertson 1968) and the threshold was set to 0.9. Note, that in this method the constructed haplotypes may have different lengths. In all cases in which SNPs were not grouped into any haplotype, the single SNPs were considered as haplotypes of length one.

GWAS based on haplotypes constructed via the above two methods are referred as SWH GWAS and LDH GWAS respectively. The significance of haplotypes was also tested using the linear mixed model (Equation 1).

Data sets

The study was based on published data of *Arabidopsis thaliana* from the 1001 Genomes Consortium (1001 Genomes Consortium 2016). The genotypic data contained 1,134 accessions with 11,458,975 single-nucleotide polymorphisms (SNPs). The phenotypic data that were considered was flowering time for plants grown at two different temperatures (10° and 16°), which included phenotypic values for 1,163 and 1,123 accessions respectively. Combining the genotypic and phenotypic data, 1,003 (970) accessions were used in the 10° (16°) data set. In the following the two data sets were referred as data set FT10 (10°) and FT16 (16°) respectively. Only bi-allelic SNPs were considered for the analyses. After removing the SNPs with missing rate above or equaling 0.1, the remaining missing values were imputed with IMPUTE2 (Howie *et al.* 2009; Howie *et al.* 2012). Linkage phases were determined by SHAPEIT (Delaneau *et al.* 2011). SNPs with minor allele frequency (MAF) below 0.05 were also removed. For subsequent analyses the resulting 756,005 and 754,656 SNPs were used for data set FT10 and FT16, respectively.

Comparing FH GWAS with other methods using empirical data

We compared the performance of FH GWAS with that of SWH GWAS, LDH GWAS and the single SNP-based approach with the *Arabidopsis thaliana* data sets described in the previous section. The genome-wide thresholds for the different approaches were generally determined as $P < 0.05$ after Bonferroni correction for multiple testing (Dunn 1961). Thus for the SNP, SWH and LDH GWAS approaches, the thresholds were $P < 0.05/m$, where m is the number of SNPs or haplotypes constructed in total. The proportion of phenotypic variance explained by each of the significant SNPs or haplotypes was calculated as the adjusted R^2 in a linear regression model with intercept and the testing variable. For regions in which significant associations were detected in GWAS, annotated genes were retrieved from Araport11 (Cheng *et al.* 2017).

Implementation of FH GWAS: In the procedure of preselecting SNPs, we filtered the markers with the threshold $P < 0.01$. Then we set the window size for searching haplotypes to be 50 kb and the number of SNPs in each haplotype to be three. Thus the linear mixed model (1) was performed for any triplet of candidate SNPs within 50 kb, testing the additive effects of each SNP, the epistatic effects for each pair of SNPs and the three-way epistatic effects. If at least two of the additive effects were significant with $P < 0.05$ and at least two of the pairwise epistatic effects were significant with $P < 0.1$, the triplet of SNPs was grouped into a functional haplotype. In the test of all functional haplotypes, we again applied the Bonferroni correction for multiple testing. But the threshold for FH GWAS needed further adjustment to account for the pre-testing procedure for single SNP effects and epistatic effects. So a more stringent threshold was determined as $P < 0.05/(m+c)$, where m is the number of functional haplotypes and c is the number of tests performed in the pre-testing procedure. To select representative significant functional haplotypes, we used the stepwise linear regression model (Draper and Smith 2014) and applied a bidirectional elimination procedure minimizing the Schwarz Bayesian Criterion (Schwarz 1978).

GWAS considering markers in perfect LD

SNPs in perfect LD ($r^2 = 1$) are virtually identical in GWAS models in the sense that they have the same estimated effects and P values. Thus for each group of SNPs in perfect LD, we recorded their positions and performed only one test in GWAS. This approach was termed SNP_{LD} GWAS. Let n_{LD} be the number of SNPs adjusted for perfect LD, meaning that SNPs in perfect LD were counted only once. Then the threshold for SNP_{LD} GWAS was $P < 0.05/n_{LD}$.

For any two haplotypes consisting of three SNPs, they may share k SNPs (k = 0, 1, 2). If the remaining 3-k pairs of SNPs are in perfect LD respectively, the two haplotypes can be treated as identical in GWAS for the same reason as above. Thus our FH GWAS approach can also be adjusted by considering SNPs in perfect LD, which was termed FH_{LD} GWAS. Let m_{LD} be the adjusted number of functional haplotypes and c_{LD} be the adjusted number of tests performed in the pre-testing procedure. Then the new threshold for FH_{LD} GWAS was determined as $P < 0.05/(m_{LD}+c_{LD})$.

Decay of linkage disequilibrium

The genome-wide decay of LD in the population of data set FT10 was estimated by a non-linear regression model using Hill and Weir's function (Hill and Weir 1988). The same method was used to estimate the decay of LD for the five candidate regions detected in GWAS.

Simulation study

Phenotypic data were simulated based on genotypic data of the 1,003 *Arabidopsis thaliana* accessions described previously (1001 Genomes Consortium 2016). Considering computational load, the simulations were restricted to all bi-allelic SNPs mapping to chromosome 2 regardless of MAF, in total 279,038 SNPs. In the simulation procedure three SNPs were always selected within a 50 kb window and main and epistatic effects were assigned to them. To clarify the influence of LD and MAF on the performance of haplotype-based GWAS, three ranges of LD (0-0.2, 0.3-0.6 and 0.7-1) between each pair of selected SNPs, three ranges of MAF (0-0.1, 0.2-0.3 and 0.4-0.5) for the selected SNPs and combinations thereof were considered. For each of the resulting nine simulation scenarios, the main effect of each SNP, the epistatic effects between each pair of SNPs and the three-term interaction effect were set to account for 6%, 3% and 1% of the explained proportion of genetic variance, respectively and the heritability was set to be 0.8. All remaining SNPs were required to contribute equally to the remaining proportion of explained genetic variance to simulate genetic background. Based on the resulting simulated phenotypic data, association mapping was performed (model 1) using the three SNPs of a particular haplotype individually and the haplotype. For each scenario, simulations were repeated 1,000 times.

Data availability

All data analyzed in this study have been published previously (1001 Genome Consortium 2016). Phenotypic data were downloaded from AraPheno (https://arapheno.1001genomes.org/phenotypes/?sort=study&page=1). Genomic data were downloaded from the 1001 Genomes data center (http://1001genomes.org/data/GMI-MPI/releases/v3.1/, the data '1001genomes_snp-short-indel_only_ACGTN.v*cf*.gz' was used in this study). FH GWAS was implemented using R (R Core Team 2017). The source code and sample data sets which are subsets of the original data set for running the code can be found at https://github.com/Fangv1/Functional_haplotype_GWAS/tree/master/. Supplemental files are available at figshare. Supplemental material available at figshare: https://doi.org/10.25387/g3.8967986.

RESULTS

FH GWAS outperformed SNP-based and two other haplotype-based GWAS approaches

In this study, data for flowering time in *Arabidopsis thaliana* accessions that had been cultivated at 10° and 16° were analyzed (1001 Genomes Consortium 2016), these compilations are referred to as data sets FT10 and FT16, respectively. Data set FT10 encompassed 1,003 accessions and after quality control 756,005 biallelic SNPs remained for subsequent analyses. To assess the performance of the proposed FH GWAS, we compared its results to those of single SNP-based GWAS and two other haplotype-based approaches in which haplotypes were either constructed using sliding-windows (SWH GWAS) or by considering LD of consecutive SNPs (LDH GWAS). The number of SNPs grouped into haplotypes and number of haplotypes analyzed in GWAS varied between the three approaches (Table S1). Applying Bonferroni correction for multiple testing (Dunn 1961) ($P < 0.05$) resulted in significance thresholds of $-\log_{10}P = 7.03$, $-\log_{10}P = 7.18$ and $-\log_{10}P = 6.50$ for LDH, SWH and FH GWAS, respectively. But for FH GWAS it was necessary to apply a further correction to account for the pre-testing procedure for single SNP effects and epistatic effects which preceded the construction of functional haplotypes (see **Materials and Methods** for details). Implementing this correction resulted in a more stringent threshold of $-\log_{10}P = 8.29$. Applying the different GWAS approaches, significant associations were found in five chromosome regions (Figure 2A-2D). Importantly, all regions that were identified by SNP-based, LDH and/or SWH GWAS were also found by FH GWAS. Four regions, I, III, IV and V, were identified by all methods, but region II on chromosome 4 solely showed significant association with flowering time using FH GWAS. For each of the significant functional haplotypes detected in region II, an additional association test was performed with only the main effects of the three SNPs in the haplotype. We found that in all cases the –log(P) values decreased by three to five orders of magnitude. This clearly showed the important contribution of epistatic effects to the overall effect of a functional haplotype.

For each haplotype the P value of the SNP for which the lowest P value had been observed in single SNP-based GWAS was compared to the one of the corresponding haplotype. The proportion of haplotypes showing significant associations that contained at least one SNP, which had passed the significance threshold in SNP-based GWAS, varied between the three different haplotype-based GWAS approaches (Figure 2E-2G). The highest proportion was found with 90.91% for LDH GWAS and the lowest one with 29.33% for FH GWAS (Table 1).

Accounting for linkage disequilibrium in functional haplotype-based GWAS

Strikingly, FH GWAS identified several thousand significant haplotypes whereas SNP-based GWAS and the other two haplotype-based GWAS approaches revealed few significant associations (Table 1, Figure 2). Inspection of the significant functional haplotypes in a given chromosome region revealed many subsets sharing one or two SNPs. For example, nine of the 15 significant haplotypes in region II had two SNPs in common. Moreover, the SNPs distinguishing these nine significant haplotypes were in high LD to each other (Figure S2). This exemplifies that many different significant functional haplotypes may result in cases in which significant haplotypes are made up of SNPs which are in high LD with other SNPs in the region. Taking into account which SNPs are in perfect LD to each other it is possible to restrict the FH GWAS analysis to those haplotypes which provide non-redundant information regarding additive and epistatic effects, called $\mathrm{FH_{LD}}$ GWAS hereafter. In data set FT10, the number of SNP combinations to be tested could be reduced in this manner from 8,932,265 to 2,460,993. Instead of 157,526 functional haplotypes in FH GWAS only 44,759 resulted in $\mathrm{FH_{LD}}$ GWAS. However, owing to a less stringent threshold of $-\log_{10}P = 7.79$ for $\mathrm{FH_{LD}}$ GWAS compared to $-\log_{10}P = 8.29$ for FH GWAS the number of significant associations increased (Table S2). Regardless whether FH GWAS or $\mathrm{FH_{LD}}$ GWAS were used multiple significant haplotypes were found in regions I to V. In addition, a single haplotype passed the significance threshold in $\mathrm{FH_{LD}}$ GWAS on chromosome 3 (Table S2, Figure 2, Figure S3).

Representative significant haplotypes narrowed down the candidate regions

Depending on the region, the mean size of the significant functional haplotypes, defined as the distance in base pairs between the outermost SNPs of a particular haplotype, varied from 10.6 to 41.3 kb in FH GWAS. Moreover, the size of chromosome segments in which overlapping significant functional haplotypes were found differed, ranging from 54.3 to 167.2 kb (Table S3). The sizes of the significant functional haplotypes in conjunction with their high number hampered the search for candidate genes. Variable selection methods were therefore used to reduce the number of significant functional haplotypes (see Materials and Methods for details). For data set FT10, two to six and two to eight representative haplotypes were selected per region in FH GWAS and $\mathrm{FH_{LD}}$ GWAS, respectively (Table S2). Taking into account the overlaps between all representative significant functional haplotypes of a given region and/or the area between them, small regions with few genes were detected (Figure 3, Figure 4, Figure S4, Table S4). In all cases a candidate gene was identified among these genes for which a role in flowering time control had been documented previously. *FT* (Kardailsky *et al.* 1999; Corbesier *et al.* 2007) represents a candidate gene for region I on chromosome 1 (Figure 3A). *DOG1* (Huo *et al.* 2016) and *FLC* (Michaels and Amasino 1999; Li *et al.* 2014) are part of regions III

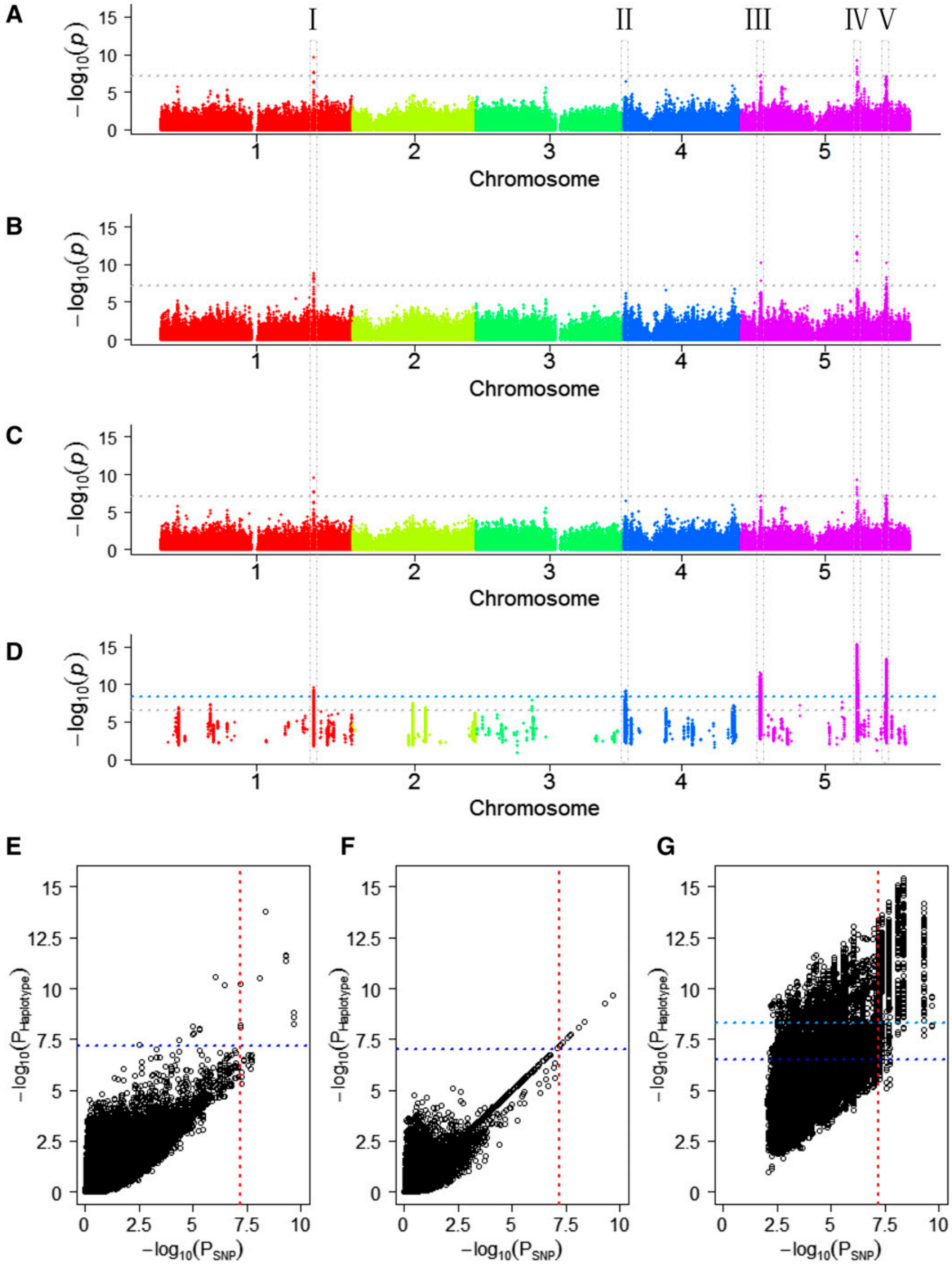


Figure 2 Association mapping results using four different GWAS approaches. (A-D) Manhattan plots illustrate the results for a single SNP-based (A) and three haplotype-based GWAS approaches for data set FT10 (B-D). Positions of SNPs or haplotypes on the five chromosomes are shown on the x axis relative to their $-\log_{10}(P)$ values on the y axis. Haplotypes were constructed based on overlapping sliding-windows (B), linkage disequilibrium (C) or by using the functional haplotype approach (D). Thresholds after Bonferroni correction for multiple testing (Dunn 1961) ($P < 0.05$) are displayed as horizontal dotted gray lines. Taking into account the pre-testing procedure for single SNP main and epistatic effects implemented in the functional haplotype approach a more stringent threshold resulted that is indicated as a pale blue dotted line in panels (D) and (G). (E-G) Plots showing the $-\log_{10}(P)$ values of haplotypes on the y axis relative to the $-\log_{10}(P)$ values established by SNP-based GWAS for the most significant SNP of a corresponding haplotype on the x axis. The P value relationships for SWH, LDH and FH based GWAS are illustrated in panels (E), (F) and (G), respectively. Thresholds after Bonferroni correction for multiple testing (Dunn 1961) ($P < 0.05$) are indicated as horizontal dotted dark blue lines for haplotypes and vertical dotted red lines for single SNPs. The five regions in which significant associations were found were denoted with I to V and are marked by stippled lines.

and IV on chromosome 5, respectively (Figure 4). In these three areas, the significant SNPs that were significantly associated with the trait flowering time were also located in the candidate gene itself or in its immediate vicinity. It is important to note that the proportions of phenotypic variance explained by the representative significant haplotypes were in four out of five analyzed regions higher than those determined for any of the SNPs in these chromosome segments (Figure 4, Table S5). In region V on chromosome 5, the only SNP significantly associated with the trait flowering time mapped approximately 60 kb apart from the region with the candidate gene *VIN3* (Sung and Amasino 2004) which was indicated by the representative significant functional haplotypes in FH GWAS (Figure 3B). In region II that had only been detected using FH GWAS and FH_{LD} GWAS, *CCT*/*CRP*/*MED12* (Imura *et al.* 2012) was identified as candidate gene (Figure S2, Figure S4).

FH GWAS for the trait flowering time at two different growth temperatures

Association studies in which the trait flowering time had been comparatively analyzed for accessions cultivated at 10° and 16° had revealed fewer significant SNP associations in the latter data set (1001 Genomes Consortium 2016). It was therefore of interest to extend the performance comparisons of SNP-based and FH GWAS to data set FT16, in which phenotypic data for 970 accessions and 754,655 biallelic SNPs that had passed quality control had been compiled. Two regions showing significant associations were identified by SNP-based GWAS as well

Table 1 Summary of significant associations in different genome-wide association studies obtained for the trait flowering time for *Arabidopsis thaliana* accessions using data set FT10

	SNP GWAS	SWH GWAS		LDH GWAS		FH GWAS	
Number of significant assocations	SNP	H_a	H_b	H_a	H_b	H_a	H_b
I (Chr1)	4	7	4	3	0	71	61
II (Chr4)	0	0	0	0	0	15	15
III (Chr5)	1	2	1	2	1	701	565
IV (Chr5)	5	6	1	5	0	19030	13900
V (Chr5)	1	5	2	1	0	4952	2963
Total	11	20	8	11	1	24769	17504

H_a refers to all significant haplotypes.
H_b represents significant haplotypes which did not contain any significant SNP.

as FH GWAS using data set FT16 (Figure S5), these corresponded to regions III and IV that had also been found for plants cultivated at 10°. In contrast, region II on chromosome 4 was solely identified by FH GWAS, regardless which of the two data sets was analyzed (Figure 2, Figure S5). Significant associations in regions I and V were not found in data set FT16. However, for FT16, FH GWAS detected in three additional regions located on chromosomes 1, 2 and 3 between one and three significant haplotypes associated with the trait flowering time (Figure S5).

The three regions, which were identified by FH GWAS in both data sets, were analyzed in more detail. A comparison of the results revealed that the chromosome segments in which the representative significant functional haplotypes were found showed large overlaps in both data sets (Table S4) implying that the same three candidate genes underlie the trait flowering time in these regions (Figure S6). Interestingly, such congruence was not observed if for example the candidate SNPs that were considered for the construction of functional haplotypes were compared in the two data sets. Although around 650 candidate SNPs were identified in each of the two data sets, only 390 were in common. Similar results were found by assessing the SNPs that were grouped into haplotypes (Table S6). Although 67,865 and 57,479 functional haplotypes had been considered in GWAS using the two data sets, only 3,606 were in common between both data sets. A similar trend was seen if only those haplotypes were considered that had passed the GWAS significance thresholds. None of the representative significant functional haplotypes were identical in the two data sets (Table S6).

The influence of linkage disequilibrium and minor allele frequencies on the power of functional haplotype-based GWAS

Simulation studies were performed to gain insight under which circumstances FH GWAS outperforms SNP-based GWAS. Specifically, it was analyzed how the minor allele frequency (MAF) of the SNPs making up a particular haplotype and the LD between them influenced the results of FH GWAS, therefore three MAF and LD ranges each as well as all of their combinations were considered (see **Materials and Methods** for details). The P value distributions obtained for the haplotypes using the nine different simulation scenarios are shown in Figure 5 side-by-side with the results for the most significant SNPs of the different haplotypes. Mean P values were inversely correlated with the MAF range in FH GWAS and SNP-based GWAS, regardless which LD range was analyzed. In scenarios in which the MAF range was kept constant, inverse correlations were seen between the mean P values and the LD range. Exceptionally, analysis of the highest MAF range revealed very similar mean P values in case of FH GWAS for the three different LD ranges. In four out of the nine scenarios tested, the mean P values obtained for FH GWAS clearly outperformed those of SNP-based GWAS, in each of these four scenarios more than 96% of the haplotypes revealed lower P values compared to the values that had been established by SNP-based GWAS for the most significant SNPs of these haplotypes (Table S7). This was not the case in the five scenarios in which the lowest MAF range and/or the highest LD range were analyzed. The same trends were observed regarding the proportion of phenotypic variance explained by the haplotypes and the most significant SNPs of the different haplotypes (Figure S7).

DISCUSSION

We devised a haplotype-based GWAS approach, FH GWAS, for studying complex quantitative traits which capitalizes on a novel way in which main and epistatic effects among markers are considered to group SNPs into haplotypes. In FH GWAS we first select SNPs with a mild threshold for main effects and then search for combinations of consecutive and/or non-consecutive SNPs in a genomic region of defined size requiring certain significance for epistatic effects. In this way, only those SNPs having true contribution to the haplotype effects via additive and/or epistatic effects are combined into functional haplotypes. Thus, FH GWAS is able to overcome the constraints of combining redundant SNPs in high LD into haplotypes and meanwhile it avoids exhausted search for optimal combinations of SNPs which is too time-consuming. It is therefore expected to be more powerful than SNP-based and other haplotype-based GWAS approaches, which was confirmed by the empirical analyses for the trait flowering time in *Arabidopsis thaliana* using the data from the 1001 Genomes Consortium (1001 Genomes Consortium 2016). Our FH GWAS approach detected not only all regions, which were detected in the SNP-based and the other two haplotype-based approaches, but also a new candidate region on chromosome 4 for plants cultivated at 10° and 16° (Figure 2, Figure S5). The FH GWAS approach can be generally applied to any quantitative trait in any homozygous species for which populations with appropriate SNP coverage and of suitable size are available. If multiple traits are studied, the functional haplotypes have to be constructed for each trait separately as the tests of marker main and epistatic effects are trait-dependent. Thus, FH GWAS enhances the power of GWAS in a way that is tailored for each trait, however, it has a higher computational load than other haplotype-based GWAS approaches in which solely consecutive SNPs are considered for haplotype construction.

On the implementation of functional haplotype-based GWAS

The first step of FH GWAS is a mild preselection of SNPs according to their main effects in order to reduce the computational load for the remaining steps. Thus, it is necessary for high density SNP data sets generated for example by whole genome sequencing projects as used in this study (1001 Genomes Consortium 2016). Theoretically, the significance of a haplotype effect can be solely a result of significant epistatic effects, or cumulative (non-)significant main and epistatic effects among the SNPs. The preselection of SNPs is therefore dispensable and can be omitted if the computational load is acceptable.

In the second step of the procedure, the construction of functional haplotypes, there are two important parameters to be determined, namely the size of the window in which the functional haplotypes are constructed and the number of SNPs to be grouped into haplotypes. The window size is essentially determined by the extent of LD in the population, however, gene density should also be considered. A too small window size leads to high LD among markers within the window,

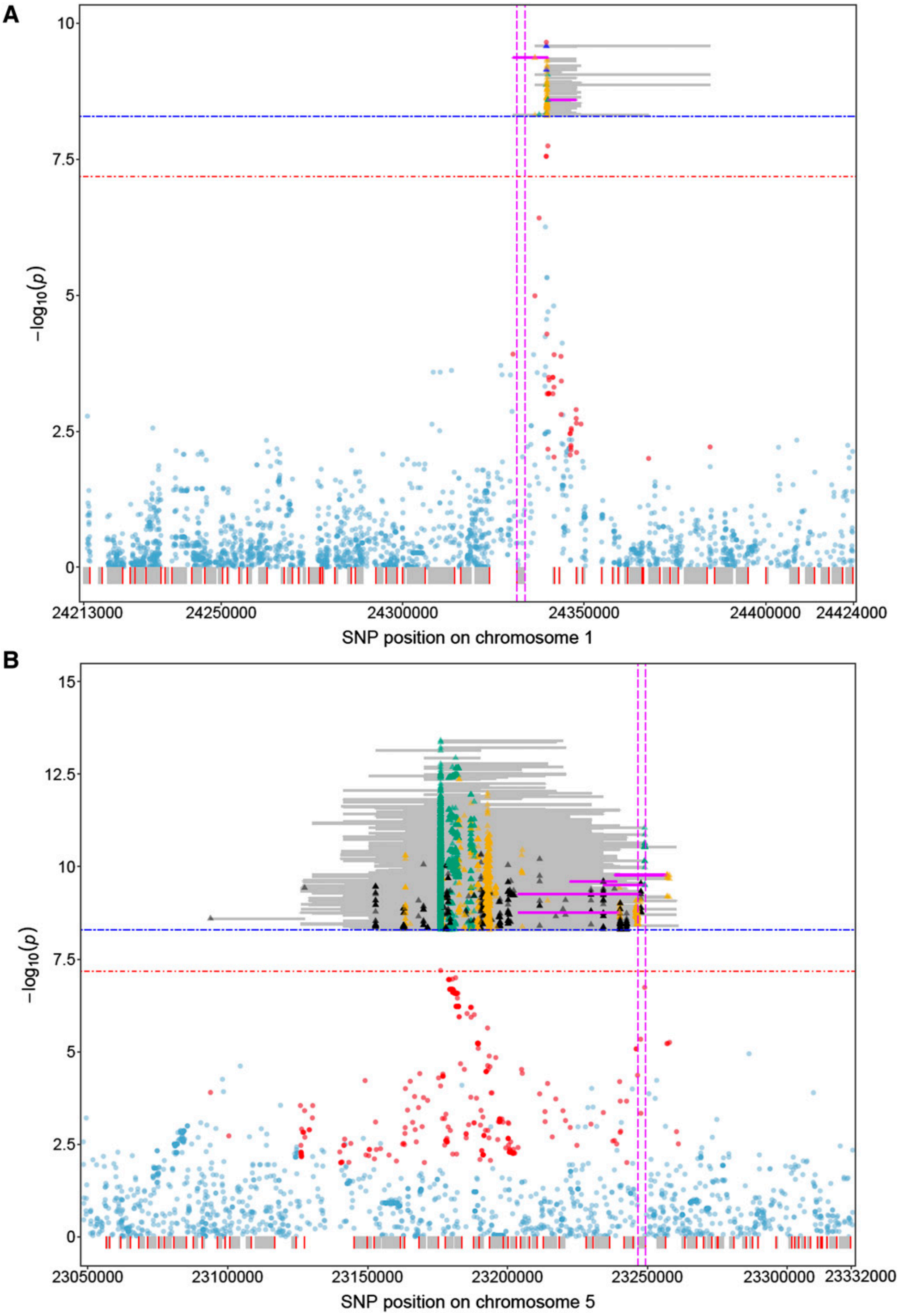


Figure 3 Details of significant associations for the trait flowering time revealed by SNP-based and functional haplotype-based GWAS in two chromosome regions. Panels (A) and (B) refer to the analysis of data set FT10 for regions I and V, respectively. SNP positions on the different chromosomes are shown on the x axis relative to the corresponding $-\log_{10}(P)$ values on the y axis. The depicted regions reflect the chromosome segments for which overlapping functional haplotypes had been obtained, but only those functional haplotypes are shown which passed the stringent adjusted significance threshold of $-\log_{10}(P) = 8.29$ as gray or pink lines. Pink lines highlight representative significant functional haplotypes. The positions of the first and third SNP of a particular haplotype on the chromosome mark the beginning and end of the line, respectively. A colored triangle indicates the SNP of a haplotype for which the lowest P value was observed by SNP-based GWAS. P values ranging from 1×10^{-4} to 1×10^{-2}, 1×10^{-6} to 1×10^{-4}, 1×10^{-8} to 1×10^{-6} are represented as black, orange and green triangles, respectively. Blue triangles represent P values smaller than 1×10^{-8}. The translucent pale blue and red dots correspond to the P values of SNPs obtained in single SNP-based GWAS, red dots represent those SNPs that were part of significant functional haplotypes. Below the x axis the coding regions of genes in the region are shown as gray boxes, 5′-regions are indicated as red lines. Two vertical pink dashed lines are used to mark the position of the coding region of the candidate gene. The red and blue horizontal stippled lines correspond to the significance thresholds for single SNP-based and FH GWAS, respectively.

reducing the advantage of haplotypes according to the results of the simulation study (Figure 5), whereas a too large window size may yield functional haplotypes that span large regions on the chromosome involving many candidate genes. For the *Arabidopsis thaliana* population considered in this study, the window size was set to be 50 kb, where the LD (measured as r^2) decayed to 0.03 (Figure S8A). On average 22 genes mapped to intervals of this size in the *Arabidopsis thaliana* Col-0 genome (Arabidopsis Genome Initiative 2000). Interestingly, in the region with the steepest LD decay (Figure S8B), region I, median and mean haplotype sizes were substantially smaller than in the other four regions (Table S3).

The number of SNPs in each haplotype is directly relevant to the power of association test, which decreases as the number of haplotype alleles increases. Usually only a small number of SNPs can be afforded unless the population size is very large, because the number of haplotype alleles grows exponentially with an increasing number of SNPs constituting the haplotype. It is also limited by the computational load because allowing more SNPs in a haplotype results in more possible

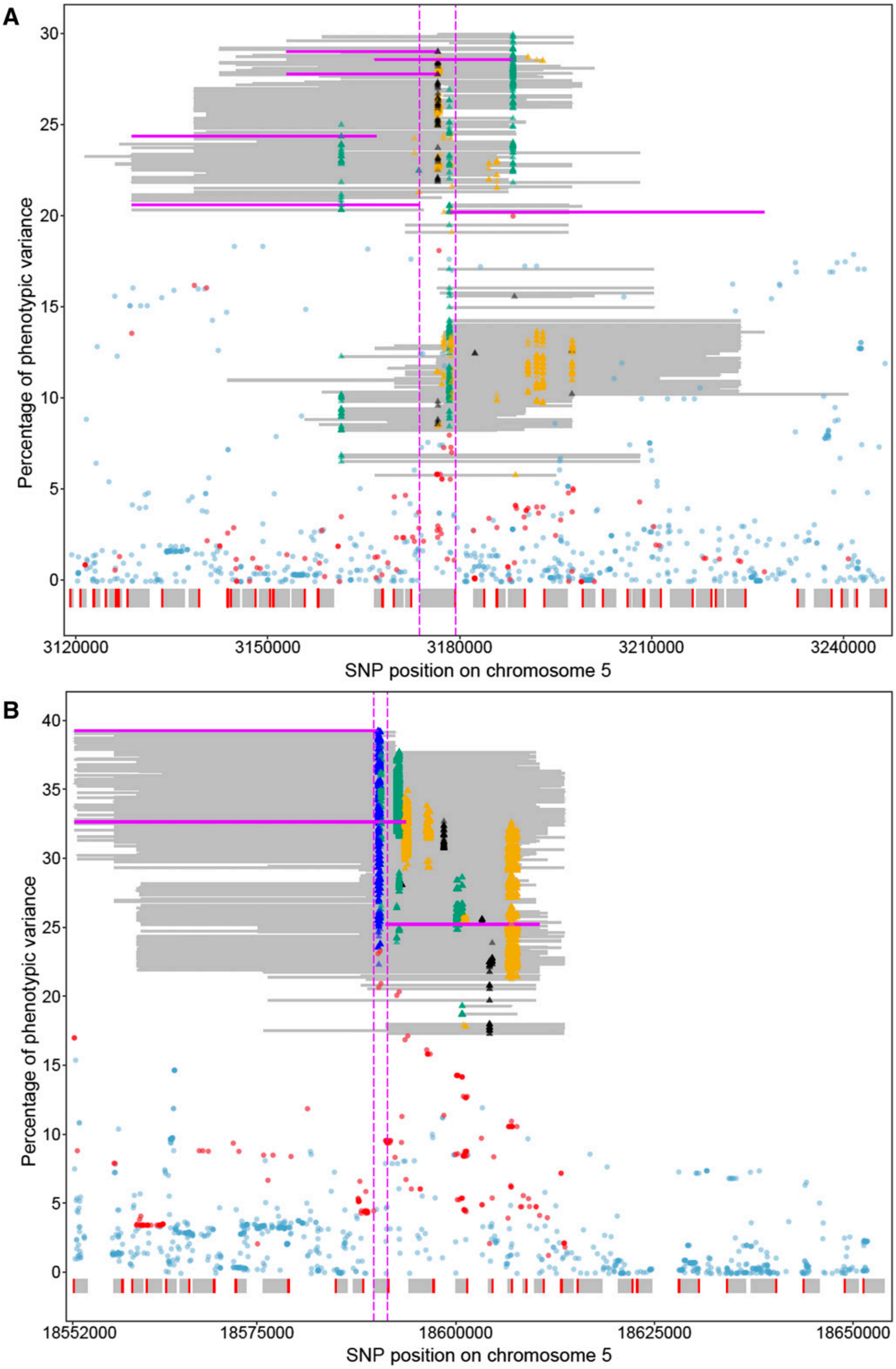


Figure 4 Proportions of explained phenotypic variance for the trait flowering time obtained by SNP-based and functional haplotype based GWAS in two different chromosome regions. Details for regions III and IV are illustrated for data set FT10 in panels (A) and (B), respectively. SNP positions on chromosome 5 and percentages of adjusted R^2 values are shown on the x and y axes, respectively. Chromosome segments are illustrated for which overlapping functional haplotypes had been obtained, but only significant functional haplotypes are displayed as gray or pink lines. Representative significant functional haplotypes are indicated by pink lines. The beginning and end of the individual lines represent the chromosome positions of the first and third SNP of a particular haplotype, respectively. The SNP for which the lowest P value of a given significant functional haplotype was obtained is indicated as a colored triangle. Black, orange and green triangles represent P values ranging from 1×10^{-4} to 1×10^{-2}, 1×10^{-6} to 1×10^{-4}, 1×10^{-8} to 1×10^{-6}, respectively. Blue triangles mark P values smaller than 1×10^{-8}. Percentages of R^2 determined for SNPs are displayed as translucent pale blue or red dots, those SNPs that were part of significant functional haplotypes are depicted in red. The coding regions of genes are shown as gray boxes and red lines represent 5′-regions. The position of the coding region of the candidate gene is marked by two vertical pink dashed lines.

combinations of SNPs to be tested. Thus in this study the number of SNPs in each haplotype block was set to be three.

Functional haplotypes boosted power of GWAS by exploiting statistical epistasis

The construction of functional haplotypes rests upon interaction effects among markers, which was termed statistical epistasis in quantitative genetics (Moore and Williams 2005). In general, the estimation of statistical epistasis is not directly relevant to biological mechanisms of gene interactions (Carlborg and Haley 2004), although some simulation studies showed that various functional dependency patterns of genes could result in significant statistical epistasis (Gjuvsland *et al.* 2007). As we observed many significant functional haplotypes consisting of SNPs with non-significant main effects even in the region where

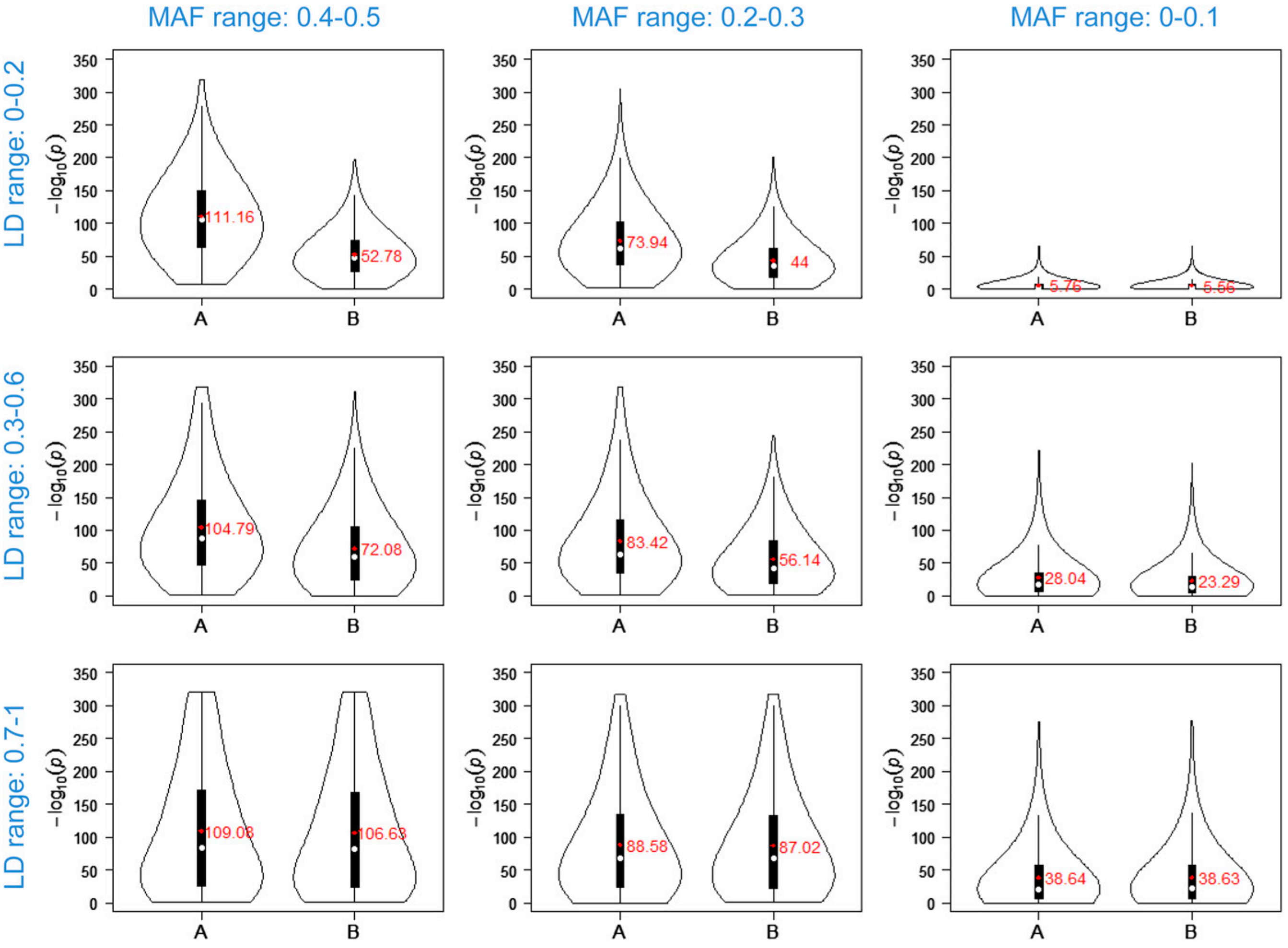


Figure 5 Comparison of P value distributions for FH GWAS and SNP-based GWAS obtained for nine different simulation scenarios. The violin plots show the distributions of P values after 1,000 simulation runs. Plots are arranged in order of decreasing MAF and increasing LD range. The y axis corresponds to the -$\log_{10}$(P values). 'A' represents the P values of haplotypes and 'B' the P values of the most significant SNPs of the particular haplotypes. The black vertical line corresponds to the 95% confidence interval and the black vertical box represents the interquartile range. The white and red dots mark the median and mean values, respectively. The latter values are indicated in the plots.

SNPs with strong main effect were detected (Figure 3B), a variable-selection algorithm was applied to select representative haplotypes. This step was of crucial importance to narrow down the regions, which needed to be inspected for the presence of candidate genes. In each candidate region, detailed analyses of the representative haplotypes revealed several distinct two- or three-locus genotype-phenotype patterns. Moreover, although three of the candidate genes were identified in two different data sets, none of the representative significant functional haplotypes were identical in these two data sets (Table S5). These findings made it unlikely that the statistical epistasis exploited by the significant haplotypes reflected a biological mechanism of gene interactions but also revealed that the cumulative statistical epistatic effects among SNPs in haplotypes indeed enhanced the power of FH-GWAS. Hence, the approach is useful for detecting new candidate regions, which cannot be detected using SNP-based or other haplotype-based GWAS approaches. Previously, haplotype-based methods were used to boost power in GWAS mainly for incomplete genotype data (McCarthy *et al.* 2008), whereas our study showed that FH GWAS is a promising method even if almost complete genotype information is available such as whole-genome sequencing data.

Further development of functional haplotype-based GWAS

In this study, FH GWAS was applied to an *Arabidopsis thaliana* population consisting of pure homozygous lines. Hence, the haplotype phase was known and only the additive-by-additive epistasis was considered in the construction of functional haplotypes. A generalization of the FH GWAS method for heterozygous populations is possible as algorithms inferring haplotype phases (Browning and Browning 2011) can be applied if the haplotype phase is unknown. It may, however, be necessary to consider other types of epistasis, additive-by-dominance and dominance-by-dominance, when constructing functional haplotypes. Note, that in these cases the relationship between haplotype effects and marker epistatic effects was only illustrated in two- or three-locus examples but not formally proved in general case (Conti and Gauderman 2004; Schaid 2004; Jiang *et al.* 2018). Thus, further theoretical and empirical studies are needed to develop an optimal strategy of FH GWAS for heterozygous populations.

ACKNOWLEDGMENTS

Fang Liu was supported with funds from China Scholarship Council (CSC). Yong Jiang was supported by the Federal Ministry of Education and Research of Germany (Grant FKZ031B0184A).

LITERATURE CITED

Abo, R., S. Knight, J. Wong, A. Cox, and N. J. Camp, 2008 hapConstructor: automatic construction and testing of haplotypes in a Monte Carlo framework. Bioinformatics 24: 2105–2107. https://doi.org/10.1093/bioinformatics/btn359

Anderson, E. C., and J. Novembre, 2003 Finding haplotype block boundaries by using the minimum-description-length principle. Am. J. Hum. Genet. 73: 336–354. https://doi.org/10.1086/377106

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815. https://doi.org/10.1038/35048692

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly, 2005 Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265. https://doi.org/10.1093/bioinformatics/bth457

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57: 289–300.

Brachi, B., G. P. Morris, and J. O. Borevitz, 2011 Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol. 12: 232. https://doi.org/10.1186/gb-2011-12-10-232

Browning, S. R., and B. L. Browning, 2011 Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. 12: 703–714. https://doi.org/10.1038/nrg3054

Cardon, L. R., and G. R. Abecasis, 2003 Using haplotype blocks to map human complex trait loci. Trends Genet. 19: 135–140. https://doi.org/10.1016/S0168-9525(03)00022-2

Carlborg, O., and C. S. Haley, 2004 Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. 5: 618–625. https://doi.org/10.1038/nrg1407

Cheng, C. Y., V. Krishnakumar, A. P. Chan, F. Thibaud-Nissen, S. Schobel *et al.*, 2017 Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 89: 789–804. https://doi.org/10.1111/tpj.13415

Clark, A. G., 2004 The role of haplotypes in candidate gene studies. Genet. Epidemiol. 27: 321–333. https://doi.org/10.1002/gepi.20025

Conti, D. V., and W. J. Gauderman, 2004 SNPs, haplotypes, and model selection in a candidate gene region: The SIMPle analysis for multilocus data. Genet. Epidemiol. 27: 429–441. https://doi.org/10.1002/gepi.20039

Corbesier, L., C. Vincent, S. H. Jang, F. Fornara, Q. Z. Fan *et al.*, 2007 *FT* protein movement contributes to long-distance signaling in floral induction of *Arabidopsis*. Science 316: 1030–1033. https://doi.org/10.1126/science.1141752

Dai, J. Y., M. Leblanc, N. L. Smith, B. Psaty, and C. Kooperberg, 2009 SHARE: an adaptive algorithm to select the most informative set of SNPs for candidate genetic association. Biostatistics 10: 680–693. https://doi.org/10.1093/biostatistics/kxp023

Delaneau, O., J. Marchini, and J. F. Zagury, 2011 A linear complexity phasing method for thousands of genomes. Nat. Methods 9: 179–181. https://doi.org/10.1038/nmeth.1785

Draper, N. R., and H. Smith, 2014 *Applied regression analysis*, John Wiley & Sons, Hoboken, NJ.

Dunn, O. J., 1961 Multiple Comparisons among Means. J. Am. Stat. Assoc. 56: 52–64. https://doi.org/10.1080/01621459.1961.10482090

Durrant, C., K. T. Zondervan, L. R. Cardon, S. Hunt, P. Deloukas *et al.*, 2004 Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am. J. Hum. Genet. 75: 35–43. https://doi.org/10.1086/422174

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science 296: 2225–2229. https://doi.org/10.1126/science.1069424

Gjuvsland, A. B., B. J. Hayes, S. W. Omholt, and O. Carlborg, 2007 Statistical epistasis is a generic feature of gene regulatory networks. Genetics 175: 411–420. https://doi.org/10.1534/genetics.106.058859

Goddard, M. E., and B. J. Hayes, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10: 381–391. https://doi.org/10.1038/nrg2575

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38: 226–231. https://doi.org/10.1007/BF01245622

Hill, W. G., and B. S. Weir, 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33: 54–78. https://doi.org/10.1016/0040-5809(88)90004-4

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, 2012 Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44: 955–959. https://doi.org/10.1038/ng.2354

Howie, B. N., P. Donnelly, and J. Marchini, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5: e1000529. https://doi.org/10.1371/journal.pgen.1000529

Huang, B. E., C. I. Amos, and D. Y. Lin, 2007 Detecting haplotype effects in genomewide association studies. Genet. Epidemiol. 31: 803–812. https://doi.org/10.1002/gepi.20242

Huo, H. Q., S. H. Wei, and K. J. Bradford, 2016 *DELAY OF GERMINATION1 (DOG1)* regulates both seed dormancy and flowering time through microRNA pathways. Proc. Natl. Acad. Sci. USA 113: E2199–E2206. https://doi.org/10.1073/pnas.1600558113

Imura, Y., Y. Kobayashi, S. Yamamoto, M. Furutani, M. Tasaka *et al.*, 2012 *CRYPTIC PRECOCIOUS/MED12* is a novel flowering regulator with multiple target steps in *Arabidopsis*. Plant Cell Physiol. 53: 287–303. https://doi.org/10.1093/pcp/pcs002

Jiang, Y., R. H. Schmidt, and J. C. Reif, 2018 Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. G3 (Bethesda) 8: 1687–1699. https://doi.org/10.1534/g3.117.300548

Kardailsky, I., V. K. Shukla, J. H. Ahn, N. Dagenais, S. K. Christensen *et al.*, 1999 Activation tagging of the floral inducer *FT*. Science 286: 1962–1965. https://doi.org/10.1126/science.286.5446.1962

Knuppel, S., J. Esparza-Gordillo, I. Marenholz, H. G. Holzhutter, A. Bauerfeind *et al.*, 2012 Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. BMC Med. Genet. 13: 8. https://doi.org/10.1186/1471-2350-13-8

Laramie, J. M., J. B. Wilk, A. L. Destefano, and R. H. Myers, 2007 HaploBuild: an algorithm to construct non-contiguous associated haplotypes in family based genetic studies. Bioinformatics 23: 2190–2192. https://doi.org/10.1093/bioinformatics/btm316

Li, P., D. Filiault, M. S. Box, E. Kerdaffrec, C. Van Oosterhout *et al.*, 2014 Multiple *FLC* haplotypes defined by independent cis-regulatory variation underpin life history diversity in *Arabidopsis thaliana*. Genes Dev. 28: 1635–1640. https://doi.org/10.1101/gad.245993.114

Lin, S., A. Chakravarti, and D. J. Cutler, 2004 Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat. Genet. 36: 1181–1188. https://doi.org/10.1038/ng1457

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson *et al.*, 2011 FaST linear mixed models for genome-wide association studies. Nat. Methods 8: 833–835. https://doi.org/10.1038/nmeth.1681

Long, A. D., and C. H. Langley, 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 9: 720–731.

Lorenz, A. J., M. T. Hamblin, and J. L. Jannink, 2010 Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. PLoS One 5: e14079. https://doi.org/10.1371/journal.pone.001407

Mackay, T. F. C., 2014 Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat. Rev. Genet. 15: 22–33. https://doi.org/10.1038/nrg3627

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.*, 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9: 356–369. https://doi.org/10.1038/nrg2344

Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155: 421–430.

Michaels, S. D., and R. M. Amasino, 1999 *FLOWERING LOCUS C* encodes a novel *MADS* domain protein that acts as a repressor of flowering. Plant Cell 11: 949–956. https://doi.org/10.1105/tpc.11.5.949

Moore, J. H., and S. M. Williams, 2005 Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. BioEssays 27: 637–646. https://doi.org/10.1002/bies.20236

Morris, R. W., and N. L. Kaplan, 2002 On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet. Epidemiol. 23: 221–233. https://doi.org/10.1002/gepi.10200

Pryce, J. E., S. Bolormaa, A. J. Chamberlain, P. J. Bowman, K. Savin *et al.*, 2010 A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J. Dairy Sci. 93: 3331–3345. https://doi.org/10.3168/jds.2009-2893

R Core Team, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for statistical Computing, Vienna.

Reif, J. C., A. E. Melchinger, and M. Frisch, 2005 Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci. 45: 1–7. https://doi.org/10.2135/cropsci2005.0001

Schaid, D. J., 2004 Evaluating associations of haplotypes with traits. Genet. Epidemiol. 27: 348–364. https://doi.org/10.1002/gepi.20037

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Stat. 6: 461–464. https://doi.org/10.1214/aos/1176344136

Sung, S. B., and R. M. Amasino, 2004 Vernalization in *Arabidopsis thaliana* is mediated by the *PHD* finger protein *VIN3*. Nature 427: 159–164. https://doi.org/10.1038/nature02195

1001 Genomes Consortium, 2016 1,135 Genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell 166: 481–491. https://doi.org/10.1016/j.cell.2016.05.063

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B 58: 267–288.

Trégouët, D. A., I. R. Konig, J. Erdmann, A. Munteanu, P. S. Braund *et al.*, 2009 Genome-wide haplotype association study identifies the *SLC22A3-LPAL2-LPA* gene cluster as a risk locus for coronary artery disease. Nat. Genet. 41: 283–285. https://doi.org/10.1038/ng.314

Yang, Y., S. S. Li, J. W. Chien, J. Andriesen, and L. P. Zhao, 2008 A systematic search for SNPs/haplotypes associated with disease phenotypes using a haplotype-based stepwise procedure. BMC Genet. 9: 90. https://doi.org/10.1186/1471-2156-9-90

Yu, J. M., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208. https://doi.org/10.1038/ng1702

Yu, Z. X., and D. J. Schaid, 2007 Sequential haplotype scan methods for association analysis. Genet. Epidemiol. 31: 553–564. https://doi.org/10.1002/gepi.20228

Zhang, K., M. H. Deng, T. Chen, M. S. Waterman, and F. Z. Sun, 2002 A dynamic programming algorithm for haplotype block partitioning. Proc. Natl. Acad. Sci. USA 99: 7335–7339. https://doi.org/10.1073/pnas.102186799

Zhao, H. Y., R. Pfeiffer, and M. H. Gail, 2003 Haplotype analysis in population genetics and association studies. Pharmacogenomics 4: 171–178. https://doi.org/10.1517/phgs.4.2.171.22636

Communicating editor: P. Morrell

Exome association analysis sheds light onto leaf rust (*Puccinia triticina*) resistance genes currently used in wheat breeding (*Triticum aestivum* L.)

Published in:

Plant biotechnology journal (2020)

DOI: 10.1111/pbi.13303

Authors: Fang Liu, Yusheng Zhao, Sebastian Beier, Yong Jiang, Patrick Thorwarth, C. Friedrich H. Longin, Martin Ganal, Axel Himmelbach, Jochen C. Reif and Albert W. Schulthess

The original paper has been published and available online:

https://onlinelibrary.wiley.com/doi/full/10.1111/pbi.13303

Plant Biotechnology Journal

Open Access

Plant Biotechnology Journal (2020) **18**, pp. 1396–1408

doi: 10.1111/pbi.13303

Exome association analysis sheds light onto leaf rust (*Puccinia triticina*) resistance genes currently used in wheat breeding (*Triticum aestivum* L.)

Fang Liu[1], Yusheng Zhao[1], Sebastian Beier[1], Yong Jiang[1], Patrick Thorwarth[2], C. Friedrich H. Longin[2], Martin Ganal[3], Axel Himmelbach[1], Jochen C. Reif[1,*] and Albert W. Schulthess[1]

[1]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Stadt Seeland, Germany*
[2]*State Plant Breeding Institute, University of Hohenheim, Stuttgart, Germany*
[3]*TraitGenetics GmbH, Gatersleben, Germany*

Received 9 September 2019;
revised 6 November 2019;
accepted 17 November 2019.

**Correspondence* (Tel +49 (0)394825840; fax 039482/5 - 137; email reif@ipk-gatersleben.de)

Keywords: leaf rust, hybrid wheat, exome capture, association mapping, candidate genes, independent validation.

Summary

Resistance breeding is crucial for a sustainable control of leaf rust (*Puccinia triticina*) in wheat (*Triticum aestivum* L.) while directly targeting functional variants is the Holy Grail for efficient marker-assisted selection and map-based cloning. We assessed the limits and prospects of exome association analysis for severity of leaf rust in a large hybrid wheat population of 1574 single-crosses plus their 133 parents. After imputation and quality control, exome sequencing revealed 202 875 single-nucleotide polymorphisms (SNPs) covering 19.7% of the high-confidence annotated gene space. We performed intensive data mining and found significant associations for 2171 SNPs corresponding to 50 different loci. Some of these associations mapped in the proximity of the already known resistance genes *Lr21*, *Lr34-B*, *Lr1* and *Lr10*, while other associated genomic regions, such as those on chromosomes 1A and 3D, harboured several annotated genes putatively involved in resistance. Validation with an independent population helped to narrow down the list of putative resistance genes that should be targeted by fine-mapping. We expect that the proposed strategy of intensive data mining coupled with validation will significantly influence research in plant genetics and breeding.

Introduction

Wheat (*Triticum aestivum* L.) is the world's second most cultivated cereal after maize and provides one-fifth of the calories intake of human population (FAO, 2019). Leaf rust, caused by *Puccinia triticina* f. sp. *tritici*, is one of the most widespread wheat diseases, which can cause up to 40% loss of wheat yield mainly by reducing kernel weight and decreasing the number of kernels per spike (Khan *et al.*, 2013). Resistance breeding, an economic and environment friendly approach, is critical for the sustainable control of wheat leaf rust (Oliver, 2014). Many researchers have studied the genetic architecture of leaf rust in order to efficiently increase the resistance level of cultivars. For instance, more than 70 leaf rust resistance genes (*Lr*) have been identified in wheat (Kassa *et al.*, 2017) and a few of them have been cloned, including the seedling stage resistance genes *Lr1*, *Lr10* and *Lr21*, as well as the adult plant resistance *Lr34* and *Lr67* loci (Cloutier *et al.*, 2007; Feuillet *et al.*, 2003; Huang *et al.*, 2009; Krattinger *et al.*, 2011; Moore *et al.*, 2015). Most of the discovered *Lr* genes confer all-stage resistance and are race-specific, with only a few exceptions like *Lr34*, *Lr46* and *Lr67* which confer non–race-specific resistance during adult plant stages (da Silva *et al.*, 2018). Race-specific resistance proves to be ineffective after a few years of introduction because of the high mutation rate or virulence dynamics of pathogen populations (Lowe *et al.*, 2011; McCallum *et al.*, 2016). Thus, researchers and breeders are trying to understand the diversity of resistance genes currently used in elite breeding populations and are continuously searching for novel *Lr* genes.

Genome-wide association mapping is often used to dissect the genetic architecture of important agronomic traits (Gong *et al.*, 2017) such as leaf rust resistance of wheat (Gao *et al.*, 2016; Juliana *et al.*, 2018; Kertho *et al.*, 2015; Maccaferri *et al.*, 2010). Association mapping can provide a high mapping resolution, particularly in genetically diverse populations, and in some cases facilitates the detection of functional quantitative trait nucleotides (Yano *et al.*, 2016). The high resolution, however, requires the characterization of the mapping population with a high density of markers. Given the advances in next-generation sequencing techniques and reduced sequencing costs, whole genome sequencing (WGS) has been suggested as a method to characterize the genetic variants of mapping populations (Schneeberger, 2014). Nevertheless, the price of WGS in wheat remains high due to the large genome and its allohexaploid nature, making it necessary to have enough coverage to distinguish homologs and homeologs (Appels *et al.*, 2018). Therefore, resistance gene enrichment sequencing (RenSeq) focusing on genes exclusively encoding intracellular nucleotide-binding/leucine-rich repeat immune receptor proteins has been suggested. In fact, combining RenSeq with association genetics allowed to clone four stem rust resistance genes in wheat (Arora *et al.*, 2019). Exome capture sequencing is an alternative solution to dramatically reduce sequencing costs by focusing on gene

© 2019 The Authors. *Plant Biotechnology Journal* published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

coding regions (Mo *et al.*, 2018). The potential of using exome sequencing has been demonstrated in a pioneering study in wheat where genes underlying wheat improvement and environmental adaptation could be identified (He *et al.*, 2019).

Our study is based on a large elite wheat population (Longin *et al.*, 2013; Zhao *et al.*, 2013) including ~1800 single-cross hybrids and their 135 parental lines adapted to the growing conditions of Central Europe: The population was phenotyped in multi-environmental field trials for leaf rust resistance and fingerprinted using exome capture sequencing. Extensive data mining facilitated to broaden our insights into the diversity of leaf rust resistance genes currently used in wheat breeding in Central Europe. Resistance genes *Lr1, Lr10, Lr21 and Lr34-B* are already exploited by breeders but also novel candidate regions on chromosome 1A or 3D were detected. The latter were validated in an independent population of 128 single-cross hybrids, which facilitated to narrow down the list of putative resistance genes. We expect that the outcomes will benefit marker-assisted selection of leaf rust resistance and represent promising targets to clone novel resistance genes.

Results

Exome sequencing revealed a broad nucleotide and haplotype diversity

Since two female parents failed to produce meaningful read coverage during exome capture sequencing and enough seed was available for 1604 of the potential 120 × 15 = 1800 single-cross hybrid combinations, our study is based on 1574 hybrids generated by crossing 118 female and 15 male elite winter wheat lines. The 133 wheat lines were selected to cover a broad range of diversity currently exploited in elite breeding in Central Europe (Zhao *et al.*, 2015). We performed exome capture sequencing of the 133 parental lines based on the NimbleGen array (Winfield *et al.*, 2012) and using an Illumina HiSeq 2500 platform. This resulted in 10.6 billion 100 bp reads (10.4 billion paired-end and 200 million single-end) that were mapped against the reference genome of ‘Chinese Spring’ (Appels *et al.*, 2018), unravelling 7 253 398 single-nucleotide polymorphism (SNP) sites. Only 0.37% (about 40 million reads) of the reads could not be mapped to the reference sequence. The mean coverage of called sites amounted to 1.5 (Figure S1). After imputation using FILLIN (Swarts *et al.*, 2014), we selected SNPs with minor allele frequency (MAF) larger or equal than 0.05 and missing rate smaller than 0.05, resulting in 202 875 SNPs used for subsequent analyses. Although rare resistance/susceptible loci may be underdetected by using a MAF threshold of 0.05, this filtering criteria should avoid the increased false-positive rate expected for rare variants in large-scale exome association studies of human diseases (Akle *et al.*, 2015). Prediction of the functional effect revealed that 22 166 SNPs induced non-synonymous variants and 144 687 out of the 202 875 SNPs (71.32%) were located in genic regions flanked by an upstream and downstream window of 1kb (Table S1). We identified SNPs for 21 249 genes (about 19.7% of all 107 891 high-confidence genes, Figure S2), with 13 399 genes (63.1%) exhibiting at least two SNPs.

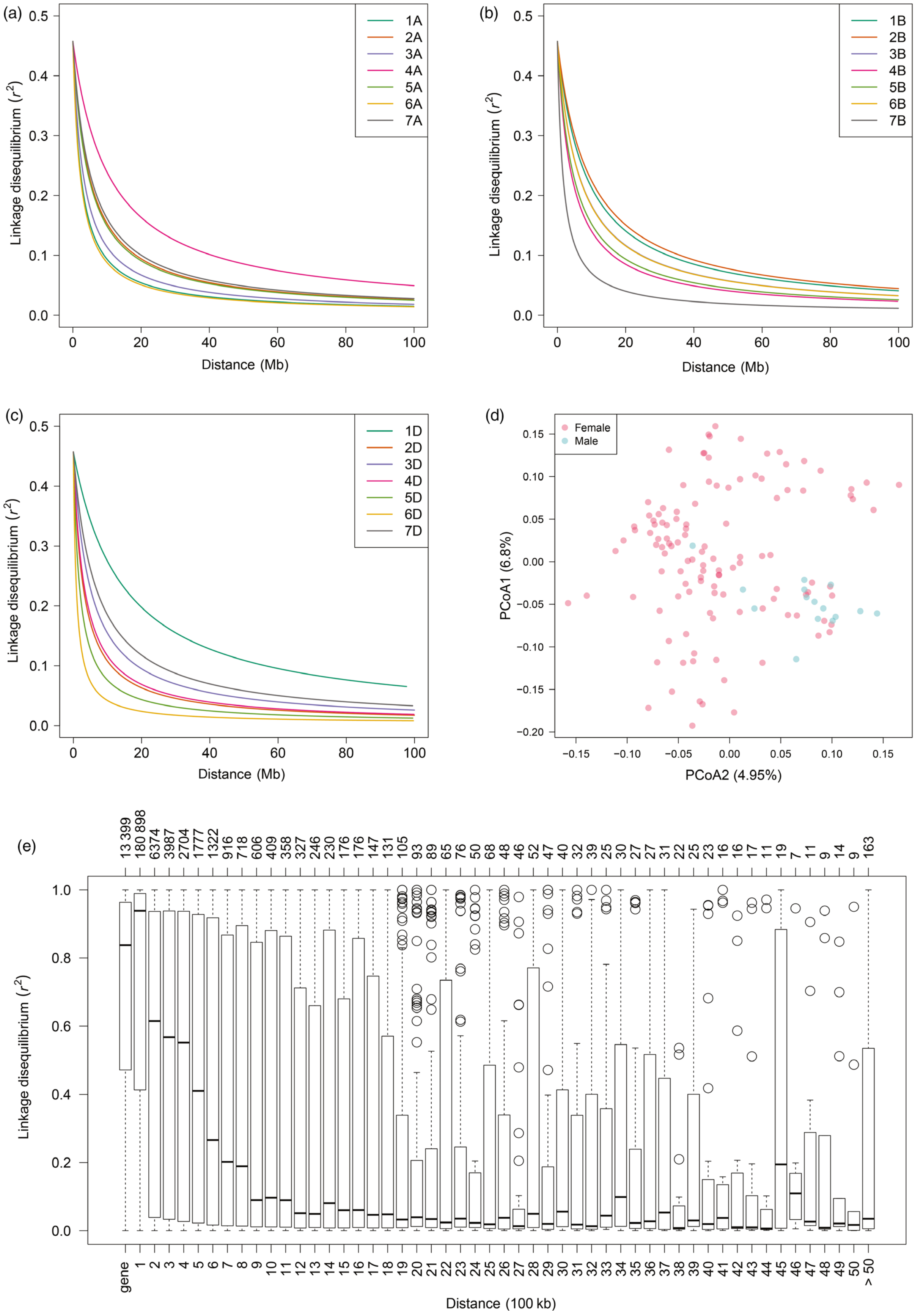
Principal coordinate analysis (PCoA) on the pairwise Rogers’ distances suggested an absence of obvious subpopulations among the 133 parents (Figure 1d), which is in line with previous findings based on a 90k SNP array (Wurschum *et al.*, 2013). These results were further supported by the tight positive correlation (Pearson’s $r = 0.911$, $P < 0.001$ using a Mantel test, Figure S3) between the genetic distances estimated using the exome sequencing and those estimated based on the 90k SNP array.

Molecular and nucleotide diversity was greater for the B than for the A genome and these two genomes, in turn, have a much higher diversity than the D genome (Figures S4 and S5). Analysis of linkage disequilibrium (LD) showed that the LD decay varied among subgenomes and chromosomes. On average, the estimated LD decayed to an $r^2 = 0.2$ at a distance of 7 Mb (Figure 1 and Table S2). LD decayed faster in the A genome than in B and D genomes, while LD curves were virtually the same for B and D genomes (Figure S6). Even though most studies have reported that LD decays slower for the D genome (e.g. Chen *et al.*, 2012; Liu *et al.*, 2017; Lopes *et al.*, 2015; Sukumaran *et al.*, 2015), there have been also some studies showing LD decay values for the D genome that are similar to or even lower than those of the A and/or B genomes (Sehgal *et al.*, 2017; Zhang *et al.*, 2013). One possible explanation for this discrepancy are the differences in number of marker pairs as a function of physical distance observed in our study (Figure S7). In this sense, a great proportion (~27%) of marker pairs in the D genome is located within very short physical distances, whereas this percentage is clearly smaller for the A and B genomes. This higher density of marker pairs at short physical distances for the D genome can cause a leverage effect that artificially forces the LD curve to be fitted towards the origin of the plot. Therefore, LD comparisons among genomes should take this issue into consideration. Chromosomes 1D and 6D presented the slowest and fastest decay, respectively, and fell below the $r^2 = 0.2$ threshold at 1 and 20 Mb, correspondingly. This rapid decay revealed the broad haplotype diversity in the underlying mapping population. Considering that a gene is the basic physical and functional unit of heredity, we estimated the LD within genes and found that the LD varied between SNPs within the same gene. In this respect, about 25% of the estimated LD values were less than $r^2 = 0.5$ (Figure 1e) while some marker pairs even approached linkage equilibrium.

Bimodal distribution of the hybrid performance for leaf rust severity

The 1574 hybrids and their 133 parents were phenotyped for leaf rust severity in 5 environments, that is year × location combinations (Table 1). Correlations between the performances of different environments ranged between 0.37 and 0.89, which suggests the existence of genotype × environment interaction effects influencing leaf rust severity. According to these

Figure 1 Linkage disequilibrium decay and diversity analysis in a wheat population composed by 1574 hybrids plus their 118 female and 15 male parent lines. Linkage disequilibrium (LD, as r^2) decay plots as a function of physical distance (Mb) within each chromosome for subgenomes A (a), B (b) and D (c). (d) Diversity among the 133 parent lines of the studied hybrid wheat population portrayed in a biplot of the first two principal coordinates from a principal coordinate analysis on the pairwise Rogers’ distance matrix calculated using exome capture single-nucleotide polymorphisms (SNPs) profiles. (e) Boxplot charts showing the distributions of average LD within a gene and of LD between adjacent SNPs. Bins in the lower *x*-axis correspond to the region defined by a gene or to regions defined by adjacent SNPs separated by certain physical distance (100 Kb). The upper *x*-axis shows the number of SNP pairs belonging to each corresponding bin in the lower *x*-axis.


(a)
Linkage disequilibrium (r^2)
0.5
0.4
0.3
0.2
0.1
0.0
0
20
40
60
80
100
Distance (Mb)
1A
2A
3A
4A
5A
6A
7A
(b)
Linkage disequilibrium (r^2)
1B
2B
3B
4B
5B
6B
7B
Distance (Mb)
(c)
Linkage disequilibrium (r^2)
1D
2D
3D
4D
5D
6D
7D
Distance (Mb)
(d)
Female
Male
PCoA1 (6.8%)
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
PCoA2 (4.95%)
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
(e)
13 399
180 898
6374
3987
2704
1777
1322
916
718
606
409
358
327
246
230
176
176
147
131
105
93
89
65
76
50
68
48
46
52
47
40
32
39
25
30
27
27
31
22
25
23
16
16
17
11
19
7
11
9
14
9
163
Linkage disequilibrium (r^2)
1.0
0.8
0.6
0.4
0.2
0.0
gene
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
> 50
Distance (100 kb)

estimates, the most abrupt changes in phenotype rankings are expected for HHOF2012 when compared with ROS2012 and ROS2013. We adjusted for the effects of environments and obtained best linear unbiased estimations (BLUEs) for the 1707 genotypes. The BLUEs were widely distributed ranging from 0.5 to 7.4 for leaf rust severity assessed by using a 1 (fully resistant) to 9 (fully susceptible) scoring scale (Figure 2a). The heritability amounted to 0.81 for hybrids and 0.82 for parent lines, while variation due to general combining abilities (GCAs) was 16.12 times the variance attributed to specific combining abilities (SCAs) (Table S4). Overall, the male parents were more susceptible than the female parents, while the severity of hybrids followed a bimodal distribution (Figure 2a). We decided to look in more detail into this bimodal distribution and split the hybrids into two subpopulations according to the resistance of the female parents: The hybrids based on crosses involving the top 25% resistant female parents form the $\text{Top}_{25\%}$ subpopulation and the hybrids from the remaining 75% female parents formed the $\text{Inferior}_{75\%}$ subpopulation. In this regard, the BLUEs of these two subpopulations followed two different normal distributions with a mean of 1.42 and 3.61, respectively (Figure 2b). This trend was also visible – but less pronounced – when splitting the population into a $\text{Top}_{50\%}$ and $\text{Inferior}_{50\%}$ subpopulation. Interestingly, we observed much more pronounced average midparent heterosis for $\text{Top}_{25\%}$ with −39.5% compared to 4.69% for $\text{Inferior}_{75\%}$ (Figure 2c). Hierarchical cluster analysis was used to search for genetic similarities among the parents of the $\text{Top}_{25\%}$ and $\text{Inferior}_{75\%}$ population. The phylogenetic tree revealed a tendency for female parents from $\text{Top}_{25\%}$ to cluster together, with the exceptions of lines F004, F011, F021, F029, F059, F086, F098, F100, F101 and F112 (Figure 2d).

Detection of subpopulation specific marker-trait associations pointing to known resistance genes

Using a mixed linear model to correct for population stratification, we firstly performed association mapping for leaf rust in the total population (Figure 3a, d). In this scan, 1565 SNPs exceeded the significant threshold of $-\log_{10}(P\text{-value}) = 5.44$, defined by applying the multiple-test correction suggested by Gao (Gao *et al.*, 2008; Figure S8). The 1565 SNPs trace back to 45 independent loci. The most significant SNPs were located on chromosome 4A and mapped as close as 7.2 Mb from the previously described resistance gene *Lr34-B*; a homolog of *Lr34* (Krattinger *et al.*, 2011). *Lr34* maps on chromosome 7D (Dakouri *et al.*, 2010; Dyck, 1987) and functions in adult plants by encoding an ATP-binding cassette (ABC) transporter. Interestingly, *Lr34-B* does not map on chromosome 7B as it would be expected due to chromosome homology, which is explained by a translocated segment from 7BS to the 4A chromosome in Chinese Spring (Krattinger *et al.*, 2011). In addition to *Lr34-B*, several significantly associated loci on chromosome 1D mapped approximately 0.4Mb away from another previously described causal gene: *Lr21* (Huang *et al.*, 2009; Figure 3d, Table 3).

We explored whether the phenotypic structure of our mapping population – as indicated by the bimodal distribution of the phenotypic values – has an effect on our association mapping results (Figure 3a). To do so, we divided the total population into resistant ($\text{Top}_{25\%}$, $\text{Top}_{50\%}$) and susceptible subpopulations ($\text{Inferior}_{75\%}$, $\text{Inferior}_{50\%}$) as outlined in detail above (Figure 3). The amount of markers differed in the subpopulations owing to the quality control of minor allele frequency (MAF, Table 2) with following ranking: Total > $\text{Inferior}_{75\%}$ > $\text{Top}_{50\%} \approx \text{Inferior}_{50\%}$- > $\text{Top}_{25\%}$. Among the significant SNPs that were detected in the total population, 429 (27.4%) were detected again as significantly associated in at least one of the four subpopulations (Figure S8). In particular, SNP *S15_2077073*, which is located proximal to *Lr21*, was significant again in the subpopulation $\text{Top}_{25\%}$ while significant associations for loci mapping as close as 7.1 Mb away from *Lr34-B* were also successfully identified in subpopulation $\text{Inferior}_{75\%}$ and $\text{Top}_{50\%}$ (Table 3 and Figure 3h, j). Interestingly, for both of the resistant subpopulations $\text{Top}_{25\%}$ and $\text{Top}_{50\%}$, we detected new marker-trait associations mapping as close as 1.4 Mb away from the known leaf rust resistance gene *Lr1* (Cloutier *et al.*, 2007), which is located on chromosome 5D (Figure 3j, i). Moreover, a strong marker-trait association on chromosome 1A was exclusively observed in the subpopulation $\text{Inferior}_{50\%}$. The SNPs of this region were located 5.9 Mb away from the CC-NBS-LRR type resistant gene *Lr10* (Feuillet *et al.*, 2003; Table 3).

Validation of marker-trait associations in an independent population

We used an independent wheat population, further denoted as validation set that comprised 128 hybrids from crosses between 24 female and 16 male parents to validate the detected marker-trait associations. Genomic data for the validation set were obtained again by exome capture sequencing, resulting in 129 818 SNPs with MAF larger or equal than 0.05 and missing rate smaller than 0.05. Out of the 1565 SNPs presenting significant marker-trait associations in the mapping population comprising the 1707 wheat genotypes, 466 were polymorphic in the independent validation set. These polymorphic SNPs reflect 15 independent loci. From the 466 SNPs, 23 were significant in the validation set at a threshold of $P < 0.01$ after applying the method for correction for multiple testing suggested by Gao *et al.* (2008). Marker effects for the 23 SNPs were estimated in the population of the 1,707 hybrids and used to predict the leaf rust severity of the hybrids in the validation set. The predicted and observed phenotypic values were significantly (P-value = 2.676×10^{-7}) correlated with a Pearson correlation of 0.384.

Independent validation facilitates to narrow down the list of putative resistance genes

We detected in the population of 1707 genotypes, a pronounced peak spanning a 25 Mb region on chromosome 3D (590–615 Mb). The SNPs were in high LD, which makes the identification of the underlying candidate gene difficult (Figure S9a). Interestingly, the diversity and pattern of LD among SNPs in this region were different in the validation set (Figure S9b), which

Table 1 Correlations among environment-specific and across-environment best linear unbiased estimations of leaf rust severity scores of a hybrid population (1574 hybrids plus their 118 female and 15 male parent lines) tested in five environments

Correlation	HAD2012	HHOF2012	ROS2012	ROS2013	BLUEs
BOH2012	0.66	0.45	0.71	0.71	0.89
HAD2012		0.42	0.52	0.52	0.75
HHOF2012			0.38	0.37	0.59
ROS2012				0.70	0.87
ROS2013					0.85

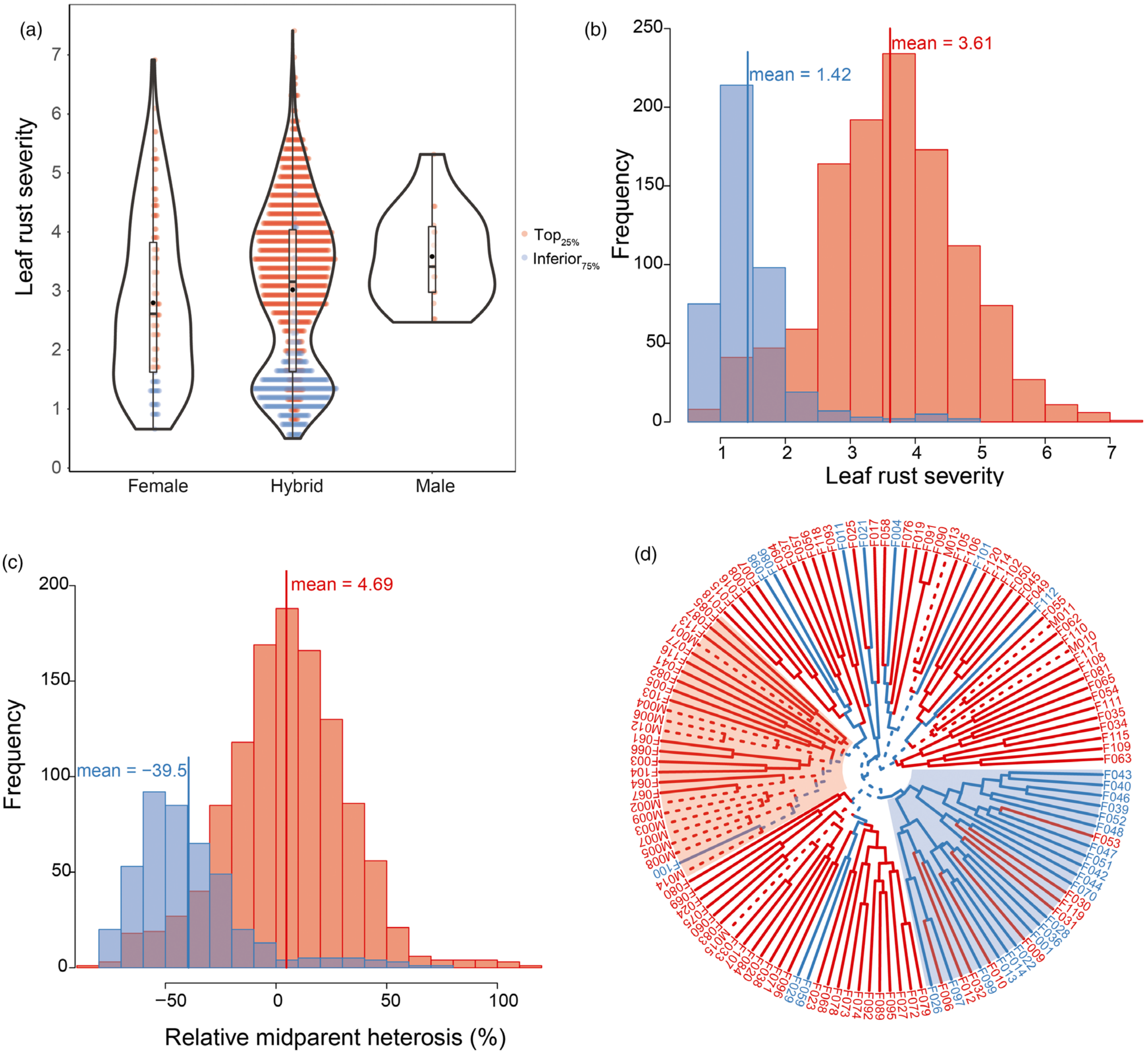


Figure 2 Bimodal distribution of leaf rust severity in a wheat population of 1574 hybrids plus their 118 female and 15 male parent lines. Genotypes from Top$_{25\%}$ and Inferior$_{75\%}$ subpopulations are indicated in blue and red, respectively. (a) Leaf rust severity scores (1 = fully resistant and 9 = fully susceptible) shown according to the different groups (females, hybrids and males). (b) Histogram of leaf rust severity scores of hybrids. (b) Histogram of midparent heterosis. (d) Dendrogram constructed by performing hierarchical clustering based on the pairwise Rogers' distance matrix among 133 parent genotypes calculated using exome capture single-nucleotide polymorphisms profiles. Solid and dashed lines represent the female and male parents, respectively.

allowed us to narrow down the list of candidate SNPs to 6 SNPs that were significant in the validation set. Within the 1Mb candidate region (597.1–598.1 Mb), four protein-coding genes (TraesCS3D01G513000, TraesCS3D01G513200, TraesCS3D01G513400 and TraesCS3D01G513500) were annotated as potential disease resistance-related genes and each of them encoded the NB-ARC domain, an important part of many plant resistance proteins (Van Ooijen *et al.*, 2008). These four genes are a promising target for further functional validation strategies such as virus-induced gene silencing or overexpression.

Another example was the region around 532.5–533.2 Mb on chromosome 1A. We identified 14 SNPs significantly associated with leaf rust severity in the population of 1707 genotypes. In

Table 2 Composition, size and number of informative exome capture sequencing single-nucleotide polymorphisms (SNPs) of each population/subpopulation for genome-wide association analysis

	Female	Male	Hybrids	SNPs
Total population	118	15	1574	202 875
Top$_{25\%}$ subpopulation	30	15	425	162 327
Inferior$_{75\%}$ subpopulation	88	15	1149	191 377
Top$_{50\%}$ subpopulation	59	15	788	180 942
Inferior$_{50\%}$ subpopulation	59	15	786	184 498
Validation population	24	16	128	112 587

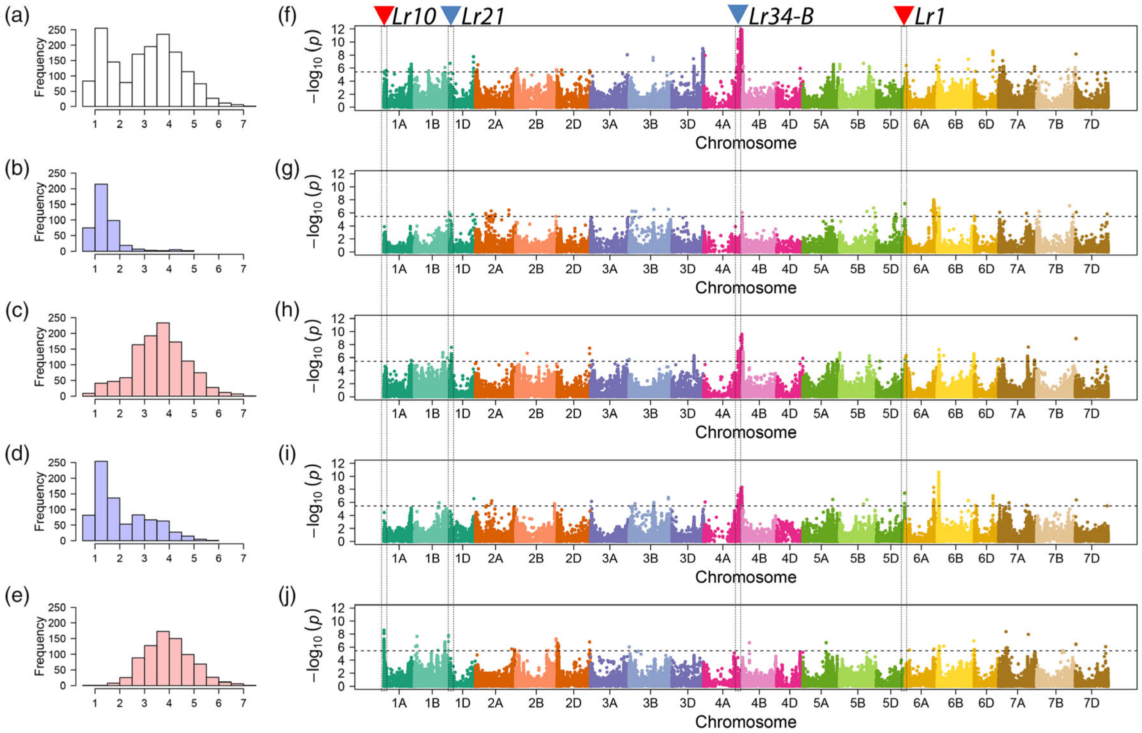


Figure 3 Genome-wide exome association scans for additive effects underlying leaf rust severity in a hybrid wheat population and its different subpopulations. Left panel: Histogram of leaf rust severity in: (a) a hybrid wheat population of 1574 hybrids plus their 118 female and 15 male parent lines and its subpopulations (b) Top$_{25\%}$, (c) Inferior$_{75\%}$, (d) Top$_{50\%}$ and (e) Inferior$_{50\%}$. Right panel: Manhattan plots of genome-wide exome association scans for additive effects underlying leaf rust severity in: (f) total population, (g) subpopulation Top$_{25\%}$, (h) Inferior$_{75\%}$, (i) Top$_{50\%}$ and (j) Inferior$_{50\%}$. –log$_{10}$(*P*-value)s of the significance test are plotted against physical positions on chromosome. Black horizontal dashed lines indicate the genome-wide multiple test corrected significance threshold for association analysis. The candidate region of *Lr10, Lr21, Lr1* and *Lr34-B* homologous gene of *Lr34* is marked with vertical dashed lines and triangles. Blue triangles mean that these loci were detected in total population and subpopulations, while red triangles mean those are only significant in subpopulations.

Table 3 Significantly associated single-nucleotide polymorphisms (SNP)s from exome capture sequencing that map closest to already known leaf rust resistant genes and located within a 10 Mb window away from the known candidate gene

Population	SNP	Chromosome	Position (bp)	*P-value*†	Gene‡	Distance (Mb)
Total	S15_2077073	1D	2 077 073	1.88E-07	*Lr21*	0.4
Total	S4_669444522	4A	669 444 522	3.86E-11	*Lr34-B*	7.2
Top$_{25\%}$	S15_2077073	1D	2 077 073	1.95E-06	*Lr21*	0.4
Top$_{25\%}$	S19_554396794	5D	554 396 794	3.76E-08	*Lr1*	7.5
Inferior$_{75\%}$	S15_99767	1D	99 767	2.80E-06	*Lr21*	2.4
Inferior$_{75\%}$	S4_669441424	4A	669 441 424	1.09E-07	*Lr34-B*	7.1
Inferior$_{75\%}$	S19_560500848	5D	560 500 848	1.94E-06	*Lr1*	1.4
Inferior$_{75\%}$	S7_55412966	7A	55 412 966	3.21E-06	*Lr34-B*	5.4
Top$_{50\%}$	S4_669444182	4A	669 444 182	7.17E-08	*Lr34-B*	7.2
Top$_{50\%}$	S19_554396794	5D	554 396 794	3.46E-08	*Lr1*	7.5
Inferior$_{50\%}$	S1_3668532	1A	3 668 532	2.45E-09	*Lr10*	5.9

[*] *P-value* of the significance test for additive effects.

[†] Resistance genes: *Lr21* (Huang *et al.*, 2009), *Lr34-B* (Krattinger *et al.*, 2011)*, Lr1* (Cloutier *et al.*, 2007) and *Lr10* (Feuillet *et al.*, 2003).

total, 24 genes were located in this candidate region and the LD among SNPs was very high (r^2> 0.65). Interestingly, the extent of LD decreased in the validation set and only 5 out of the 14 SNPs surpassed the significance threshold in the validation set (Figure 4). These SNPs were located at 532.7Mb and were linked with each other (r^2 > 0.84). We detected three genes in this region, which are putatively related with disease resistance. Two of them (TraesCS1A01G345400 and TraesCS1A01G345500) were annotated with protein kinase activity and the remaining one (TraesCS1A01G345600) encodes the LRR receptor-like serine/threonine-protein kinase domain. In more detail, annotated genes TraesCS1A01G345400 and TraesCS1A01G345500 presented

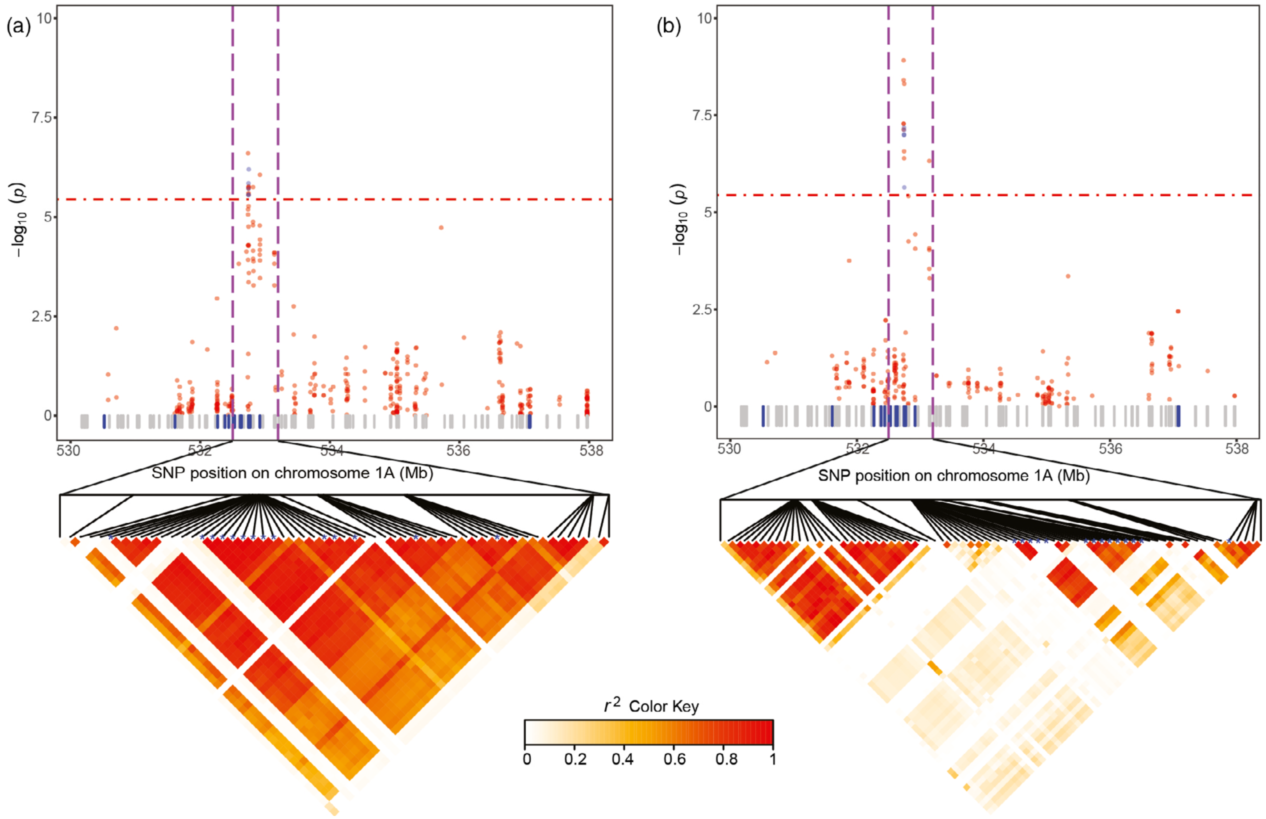


Figure 4 Candidate region associated with leaf rust resistance on chromosome 1A in a hybrid wheat population and narrowed down using an independent validation population. Manhattan plots showing the significant exome associations for additive effects underlying leaf rust within a candidate region on chromosome 1A found in: (a) a hybrid wheat population of 1574 hybrids plus their 118 female and 15 male parent lines and (b) a validation population of 128 hybrids plus their 24 female and 16 male parent lines. $-\log_{10}$(*P*-value)s of the significance test of additive effects are plotted against physical positions on chromosome 1A. Red horizontal dot-dashed lines indicate the multiple test-corrected significance thresholds for association analysis. Single-nucleotide polymorphisms (SNPs) significantly associated in the two data sets are shown as blue points and other SNPs are shown as red points. The genes with annotated resistance function and others are shown as vertical boxes in blue and grey, respectively. The upper-triangular halves of the linkage disequilibrium (LD, as r^2) matrices between SNPs within the candidate region are shown as heat maps below Manhattan plots. Blue stars in LD plots indicate the physical positions of SNPs with significant associations.

SNPs with significant associations in both populations, while TraesCS1A01G345600 has no significant SNPs in any of the studied populations (Figure S18). Similarly, as for the narrowed-down region on chromosome 3D, this region on chromosome 1A carries promising targets for further detailed validation studies.

Discussion

Exploiting environmentally stable QTL for durable resistance breeding

Our study showed that the phenotypic variation of leaf rust resistance is influenced by major and minor-effect loci (Figure S10) with the most important major effect loci located on chromosome 4A proximal to *Lr34-B*. Thus and as often suggested by several authors (Nelson *et al.*, 2018), a strategy combining major- and minor-effect genes could provide durable resistance. Moreover, resistance loci, especially those with major effects, should ideally provide race nonspecific resistance, because this type of resistance has proven to be longer lasting as compared to the race-specific one (Nelson *et al.*, 2018). For instance, the *Lr34* gene has provided resistance against several pathogens, including leaf rust, stem rust, stripe rust and powdery mildew, for over 100 years (Moore *et al.*, 2015). Obviously, whether or not these loci would provide durable resistance will be determined by the evolution of leaf rust populations in the field, which is, to a greater extent, determined by good and bad practices of integrated pest management used by wheat growers (Mundt, 2014). Nonetheless, even though the term durability can only be defined in a retrospective fashion (Nelson *et al.*, 2018), analysing the environmental stability of marker-trait associations may give some insights into it. At this stage, loci whose associations are environmentally unstable are obviously too risky to be used in marker-assisted selection. Thus, we fitted a multiple linear regression model for each marker on the environmental BLUEs of genotypes that included main environment and marker effects plus their interactions. Interestingly, these analyses revealed that only for 3.4 % of the associated loci, the marker by environment interaction components explained an equal or higher amount of variation on leaf rust severity as compared to the main locus effects (Figure S12). On the other hand, the most environmentally stable locus was located in the 4A QTL region, with main effects explaining 13.6-fold the amount of variation explained by the interaction components. Moreover, this QTL region harboured 6 additional highly environmentally stable loci, with main effects explaining at least 10 times the amount of leaf rust severity variation attributed to their interactions with the environment. Considering that we could not find significant associations for SNPs directly targeting *Lr34-B* in our population, our results point to potentially new sources of resistance mapping on chromosome 4A. In fact, some of these strongly associated loci were located within annotated genes with putative disease resistance activity (Table S5). Interestingly, by sorting SNPs according to the physical map positions of the reference sequence of Chinese Spring, associated loci portrayed two regions on chromosome 4A influencing leaf rust resistance, whose peaks were separated by 43 Mb (Figure S11). This observation was surprising considering the very high LD ($r^2 = 0.9$) between both peaks. One plausible explanation for this discrepancy is a lack of structural collinearity between the reference genome and those of our studied population due to insertions and deletions, translocations, among other genome rearrangements (Dvorak *et al.*, 2018; Helguera *et al.*, 2015; Thind *et al.*, 2018). Nonetheless, reverse and forward genetic techniques would be necessary to elucidate if these highly promising loci mapping on chromosome 4A confer new sources of resistance against leaf rust or, to the contrary, belong actually to *Lr34-B.*

Data mining broadened insights into the gene portfolio currently used in wheat breeding in Central Europe

Factors influencing the statistical power for QTL detection in association mapping are very well known (Myles *et al.*, 2009). Although the size of the association mapping population is certainly the key factor influencing statistical power, the ability to detect true QTL signals can also be improved by a decreased correlation between genetic and phenotypic similarity as well as by increased frequencies of rare alleles at functional loci. In this respect, our strategy of subdividing the total hybrid population into four different subpopulations based on the bimodal leaf rust severity distribution of parents decreased the correlation between genetic and phenotypic distances ($r_{RD,PD}$) in subpopulations Inferior$_{75\%}$, Top$_{50\%}$ and Inferior$_{50\%}$ as compared to that of the total population (Table S3). Therefore, a general increment in QTL detection power was expected for these three subpopulations, obviously, at expenses of the general power achieved by an increased size in the total population. Moreover, some associations found in the proximity of already known resistance loci such as *Lr1* and *Lr10* were only detected in subpopulations (Figure 3, Table 3), which further highlights the advantage of our strategy. In addition, analysing subpopulations improved QTL detection power by increasing MAF in some cases. For instance and compared to the total population, $-\log_{10}$(*P-value*)s of loci on chromosome 6A were higher in subpopulations Top$_{25\%}$ and Top$_{50\%}$ (Figure 3g, i and Figure S15), while MAF of most of these loci was concomitantly higher in both subpopulations. A similar observation was done for loci on chromosome 6B in the Top$_{50\%}$ subpopulation (Figures 3i and S14). Nonetheless, there were some increases in QTL detection power in subpopulations whose causal factors could not be elucidated. For example, associations in the proximity of *Lr10* were only detected in the Inferior$_{50\%}$ subpopulation. However, neither the differences in MAF (Figures 3j and S17), nor the changes in $r_{RD,PD}$ nor the presence/absence of a confounding genetic background (i.e. when genetic distances are more/less correlated with those distances portrayed by associated loci; Table S3), can explain the improved detection ability for this QTL in this particular subpopulation (Figure 3f-j). Although some driving forces underlying the improved QTL detection ability remain hidden for some subpopulations, the combined analysis of a large population plus its subpopulations increased the number of associations by ~37% compared to the total population (Figure S8), thus providing a robust strategy for QTL detection in our study.

Exploiting dominance effects through hybrid wheat breeding

The quantitative inheritance of leaf rust resistance is predominantly of additive nature, although past studies have shown that dominance effects at some loci also contribute to the genetic variation (Ahamed *et al.*, 2004; Jacobs and Broers, 1989; Navabi *et al.*, 2003). The additive nature is also supported by the associations for leaf rust severity detected by our approach, with 2151 (45 independent loci, 90%) being of additive type (Figure 3f-j). Despite this, 20 (5 independent loci, 10%) prominent dominance association effects were detected. Among them, a locus on chromosome 5A appeared as highly significant ($-\log_{10}$(*P*-value) = 8) in the Top$_{25\%}$ subpopulation (Figure S13). In this respect, hybrid breeding provides a straightforward manner for their exploitation. This is in particular attractive considering that midparent heterosis reached desired negative values of up to −82.89% (Figures 2c, S14).

Prospects of exome sequencing association studies in resistance breeding: a critical view

Association mapping has gained much popularity in the plant breeding community because it provides a very straightforward and cheap way to discover new marker-trait associations that could be exploited by means of marker-assisted selection. Nevertheless, true genetic linkage to the functional causal variant(s) underlying trait(s) is not always the cause of marker-trait association(s). In this respect, linkage disequilibrium decay between associated marker(s) and functional variant(s) in the material under selection will reduce the efficiency of marker-assisted selection (Lande and Thompson, 1990). Theoretically, one way to overcome this limitation is to focus on protein-coding regions (Hayes and Szucs, 2006), that is genes, by relying on targeted sequence methods such as exome capture. In this sense, whole-exome association mapping has proven to be beneficial for dissecting human diseases in the past (Carter *et al.*, 2014; Guo *et al.*, 2018; Kim *et al.*, 2012) and lately, exome association mapping has been also conducted using plant populations (He *et al.*, 2019; Henry *et al.*, 2014; Looseley *et al.*, 2017; Pont *et al.*, 2019; Russell *et al.*, 2016). For instance, exome capture revealed regions containing genes that are associated to traits involved in adaptation and also subjected to selection due to domestication and plant breeding in a population composed of 487 genotypes of wheat and related species (Pont *et al.*, 2019). Taking into account the size of our association mapping population, we expected to find associations that were narrowed down to the level of true functional associations. It is important to consider that due to the presence/absence nature of genetic variants conferring resistance (Arora *et al.*, 2019), insertions and deletions may play a central role when detecting candidate resistance genes based on a reference genome. In this sense, our candidate search was confined to those resistance genes annotated for Chinese Spring. In a first step, we considered already known *Lr* genes (Table 3) as a kind of proof-of-concept and in the best case; we were able to approach *Lr* genes as close as 0.4 Mb. As discussed for the *Lr34-B* on chromosome 4A; a lack of structural collinearity with the reference genome can explain an imprecise mapping of known leaf rust resistance loci and their associations in our mapping population. In addition, limited allelic variation and recombination rates are also plausible causes for this restricted ability. For instance, even though 10 SNPs were found and tested for significant associations within *Lr1*, their low MAF values (MAF = 0.05–0.08) limited our detection ability. Particularly, this gene was presumably almost fixed in all parent lines due to selection and breeding. We were able to overcome some of these challenges by using an independent population as a way to narrow down and statistically validate new associations. This strategy allowed us to increase mapping precision by, for example, narrowing down an extensive 25-Mb region on chromosome 3D found as associated in the population of 1707 genotypes to a 1-Mb region in the validation population (Figure S8). Despite this and partly because regions harbouring SNPs in strong LD and with similar allele frequency result in similar *P-values* in association mapping, we were not able to validate and narrow down associations to the level of one single gene. This last ability may also be limited in our study by the overall low gene coverage of exome capture (Figure S2, Table S1). On the other hand and due to the evolution of resistance genes, many of them lie together within clusters of highly linked resistance genes (Dilbirligi *et al.*, 2004; Liu *et al.*, 2015); an issue that clearly challenges the detection of single functional variants. At this stage, we expect that deep next-generation sequencing approaches or whole genome sequencing can shed some light into this issue. Alternatively, if the goal is the detection of new functional genes with resistance activity against biotrophic fungi, RenSeq approaches (Steuernagel *et al.*, 2018) targeting sequences such as nucleotide-binding site–leucine-rich repeats (NB-LRRs) with known participation in plant resistance should be also appropriate.

Experimental procedures

Plant material and phenotypic data analyses

The phenotypic data are based on 1749 wheat genotypes including 10 checks, 1604 F1 factorial hybrids, and their 120 female and 15 male parental lines (Longin *et al.*, 2013; Zhao *et al.*, 2013). Leaf rust disease severities were evaluated based on natural infection or deliberate inoculation in four locations (Böhnshausen, Hadmersleben, Harzhof and Rosenthal) in 2012 as described in detail elsewhere (Gowda *et al.*, 2014; Longin *et al.*, 2013). An additional field trial was conducted based on natural infection in Rosenthal during 2013 using the same experimental procedures. The leaf rust disease severities were scored at the date of flowering on the flag leaf using a scale from 1 (fully resistant) to 9 (fully susceptible) referring to the guidelines of the German Federal Plant Variety Office (Bundessortenamt, 2000).

Best linear unbiased estimations (BLUEs) of genotypes, variance components, and broad-sense heritabilities for parent lines and hybrids were estimated as outlined in detail elsewhere (Zhao *et al.*, 2015).

Genotypic data and diversity analyses

The 135 elite parental lines were genotyped with exome capture sequencing using an Illumina HiSeq 2500 platform. Sequencing data were mapped to the reference genome of Chinese Spring (Appels *et al.*, 2018). This landrace is susceptible against leaf rust at the seedling stage (Li *et al.*, 2010), but carries adult plant resistance (Dyck, 1991; Kerber and Aung, 1999). Details of the bioinformatics pipelines used for read mapping and variant calling are described in a previous study (Milner *et al.*, 2019). Briefly, BWA-MEM (Li, 2013) and SAMtools (Li *et al.*, 2009) were used to align reads to the reference sequence and convert them to binary format (BAM), respectively. GATK (DePristo *et al.*, 2011; version v3.8) was applied to realign reads near indels. Variant calling was performed with the SAMtools/BCFtools pipeline (version 1.6; Li, 2011).

A custom awk script was used for gentle filtering of variants, retaining VCF file entries with a minimum number of reads set to one for homozygous and two for heterozygous calls, respectively. Minimum SNP quality was set to 40. The resulting VCF file was imported into R statistical environment (version 3.4.3) for further filtering. Applying the SeqArray package (Zheng *et al.*, 2017) we set polymorphisms detected on chrUn to missing and filtered remaining SNPs for a minor allele count of at least one and a minimum number of present calls of 0.05. Missing genotype calls were imputed with FILLIN (Swarts *et al.*, 2014) from the TASSEL5 (Bradbury *et al.*, 2007) software suit.

After imputation with FILLIN (Swarts *et al.*, 2014), only bi-allelic variants with MAF ≥ 0.05 and missing rate < 0.05 were used for subsequent analyses. Following quality control, two female parental lines were excluded, resulting in SNP profiles for 118 female and 15 male parental lines. The genotypes of 1574 hybrids were derived from the genotypes of their parental lines. Prediction of the functional effect was performed with the tool SnpEff version 4.3 (Cingolani *et al.*, 2012) based on the IWGSC_v1.1 gene models of high confidence. Nucleotide diversity π was calculated with 1Mb non-overlapping sliding window using the software vcftools version 0.1.12b (Danecek *et al.*, 2011). The linkage disequilibrium (LD) of each chromosome was calculated using the r^2 statistic (Hill and Robertson, 1968). We applied a non-linear regression model described by Hill and Weir (Hill and Weir, 1988) to estimate the LD decay. We used the average r^2 of all SNPs within the same gene to represent LD of genes. LD within a specific genomic region was calculated based on parental lines and visualized with the R package LDheatmap (Shin *et al.*, 2006). Principal coordinate analysis and hierarchical cluster analysis were performed based on pairwise Rogers' distances among genotypes (Reif *et al.*, 2005). In a previous study (Wurschum *et al.*, 2013), parents were characterized by using a 90K Infinium SNP chip (Wang *et al.*, 2014) and pairwise Rogers' distances based on these marker profiles were also considered for comparative purposes. These analyses were performed within R environment (version 3.4.3).

Genome-wide association mapping

Genome-wide association mapping was implemented in R environment (version 3.4.3) using a linear mixed model considering additive and dominance effects (Zhao *et al.*, 2013). In brief, the model can be described as follows:

$$y = 1_n\mu + Aa + Dd + g + e, \qquad (1)$$

where y are the observed phenotypic values, 1_n corresponds to a n-length vector of ones, μ denotes a common intercept term, a and d represent the additive and dominance effect of the tested SNP, respectively, while A and D stand for the design matrices relating y to a an d, correspondingly, g is a vector of genotypic effects or polygenic background effects and e indicates the error term of the model. For each tested SNP, genotypes homozygous for the first allele, heterozygous and homozygous for the alternative allele were coded as -1, 0 and 1, respectively, in the case of the A matrix. In the D matrix, homozygous and heterozygous genotypes were coded as 0 and 1, respectively. For the model (1), we assumed that μ, a and d are fixed factors, while g and e were considered random, with $g \sim N\left(0, \sigma_g^2 K\right)$ and $e \sim N\left(0, \sigma_e^2 I\right)$, where K is a marker-derived kinship matrix, I is an identity matrix, σ_g^2 and σ_e^2 are the corresponding variance components. Each kinship coefficient between two parent genotypes within K was computed as twice the difference of one minus the corresponding Rogers' distance (Reif *et al.*, 2005) between them. In the case of hybrids, the additive polygenic background is decomposed as the sum of the general combining abilities of female (GCA_F) and male (GCA_M) parents, thus, the kinship matrices for hybrids model the covariance among GCA effects of the respective female and male parents (Bernardo, 1994). Linear mixed models for the phenotypic data analyses as well as for association mapping were solved using the ASReml-R package (Butler *et al.*, 2009).

Validation of marker-trait associations in an independent population

The validation population is a fraction of another large hybrid wheat population, which consists of 41 male lines, 189 female lines and their 1815 single-cross hybrids produced using an incomplete factorial mating design. The 230 parental lines were tested together with 1815 hybrids and 11 commercial check varieties in 7 environments in un-replicated field trials based on an alpha design with block size equals to 11 plots. Infection of genotypes with leaf rust occurred naturally and was scored at the date of flowering on the flag leaf as described in detail above.

Across environments, BLUEs of lines and hybrids from validation population were obtained as outlined in detail elsewhere (Zhao *et al.*, 2015). For 24 female and 16 male parental lines, exome capture data were obtained as already detailed in the 'Genotypic data and diversity analyses' section. The genotypes of hybrids were deduced according to the genotypes of their parents. The 40 lines served as parents for 128 hybrids, which were denoted in the following as the validation population.

To validate the significant SNPs found in the population of the 1707 genotypes (133 parental lines and 1574 hybrids), we first identified common markers between the two populations. Then, all the common markers were used to predict the phenotype of lines in the validation population using a linear model, in which the markers were sorted by physical position on the chromosome. Finally, we calculated the Pearson correlation coefficient between the predicted and observed phenotypic values.

Narrow down candidate regions combining information of the two populations and identification of candidate genes

For the novel candidate regions, we used all the markers in those regions that were found in the validation population and performed association tests based on a linear regression model. The detected potential disease resistance-related genes (R genes) were annotated with the pipeline RGAugury (Li *et al.*, 2016). R genes were classified as CC (coiled-coil), NBS (nucleotide-binding site), CN (CC-NBS), NL (NBS-LRR), CNL (CC-NBS-LRR), RLK (receptor-like kinase), RLP (receptor-like protein) and TM-CC (Transmembrane-CC). Moreover, we double-checked the functional annotation of these genes in the candidate region IWGSC (Appels *et al.*, 2018).

Acknowledgements

We would like to thank Martin Mascher and Beat Keller for their valuable comments that helped us to improve the manuscript. We also want to thank Peter Morrell for his advices for the population genomic analyses.

Funding

FL was supported with funds from China Scholarship Council (CSC; Grant no. 201606760065). The experimental work of the validation set as well as YJ and PT were supported by the German Federal Ministry of Food and Agriculture within the ZUCHTWERT Project (Grant FKZ0103010). The Federal Ministry of Education and Research of Germany is acknowledged for funding AWS (Grant no. FKZ031B0184B).

Conflict of interest

The authors declare that they have no conflict of interest and the experiments comply with the current laws of Germany.

Author contributions

YZ and JCR designed the study; MG generated genomic data and SB processed it; PT, YZ and FL curated phenotypic and genomic data; FL performed the analyses with the support of YZ and YJ; FL and AWS wrote the paper with input from all co-authors.

References

Ahamed, M.L., Singh, S.S., Sharma, J.B. and Ram, R.B. (2004) Evaluation of inheritance to leaf rust in wheat using area under disease progress curve. *Hereditas*, **141**, 323–327.

Akle, S., Chun, S., Jordan, D.M. and Cassa, C.A. (2015) Mitigating false-positive associations in rare disease gene discovery. *Hum. Mutat.* **36**, 998–1003.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C.J. *et al*. (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, 661.

Arora, S., Steuernagel, B., Gaurav, K., Chandramohan, S., Long, Y., Matny, O., Johnson, R. *et al.* (2019) Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nat. Biotechnol.* **37**, 139–143.

Bernardo, R. (1994) Prediction of maize single-cross performance using Rflps and information from related hybrids. *Crop Sci* **34**, 20–25.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Bundessortenamt (2000) *Richtlinien für die Durchführung von landwirtschaftlichen Wertprüfungen und Sortenversuchen*. Hannover, Germany: Landbuch Verlagsgesellschaft mbH.

Butler, D., Cullis, B.R., Gilmour, A. and Gogel, B. (2009) *ASReml-R reference manual*. Brisbane: The State of Queensland, Department of Primary Industries and Fisheries.

Carter, S.L., Brastianos, P.K., McFadden, D.G., Papagiannakopoulos, T., Taylor-Weiner, A., Cibulskis, K., Jacks, T. *et al*. (2014) Modes of metastasis evolution in human and murine cancer revealed by whole exome sequencing. *Cancer Res.* **74**, 3992.

Chen, X.J., Min, D.H., Yasir, T.A. and Hu, Y.G. (2012) Genetic diversity, population structure and linkage disequilibrium in elite Chinese winter wheat investigated with SSR markers. *Plos One*, **7**, e44510.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J. *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly*, **6**, 80–92.

Cloutier, S., McCallum, B.D., Loutre, C., Banks, T.W., Wicker, T., Feuillet, C., Keller, B. *et al*. (2007) Leaf rust resistance gene Lr1, isolated from bread wheat (*Triticum aestivum* L.) is a member of the large psr567 gene family. *Plant Mol. Biol.* **65**, 93–106.

Dakouri, A., McCallum, B.D., Walichnowski, A.Z. and Cloutier, S. (2010) Fine-mapping of the leaf rust Lr34 locus in *Triticum aestivum* (L.) and characterization of large germplasm collections support the ABC transporter as essential for gene function. *Theoret. Appl. Genet.* **121**, 373–384.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E. *et al*. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A. *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491.

Dilbirligi, M., Erayman, M., Sandhu, D., Sidhu, D. and Gill, K.S. (2004) Identification of wheat chromosomal regions containing expressed resistance genes. *Genetics*, **166**, 461–481.

Dvorak, J., Wang, L., Zhu, T., Jorgensen, C.M., Luo, M.C., Deal, K.R., Gu, Y.Q. *et al*. (2018) Reassessment of the evolution of wheat chromosomes 4A, 5A, and 7B. *Theoret. Appl. Genet.* **131**, 2451–2462.

Dyck, P.L. (1987) The association of a gene for leaf rust resistance with the chromosome-7d suppressor of stem rust resistance in common wheat. *Genome*, **29**, 467–469.

Dyck, P.L. (1991) Genetics of adult-plant leaf rust resistance in Chinese spring and sturdy wheats. *Crop Sci.* **31**, 309–311.

FAO (2019) *Food and agriculture organisation of the United Nations, FAOSTAT*, http://www.fao.org/faostat/en/#data

Feuillet, C., Travella, S., Stein, N., Albar, L., Nublat, A. and Keller, B. (2003) Map-based isolation of the leaf rust disease resistance gene Lr10 from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proc. Natl. Acad. Sci. USA*, **100**, 15253–15258.

Gao, X.Y., Stamier, J. and Martin, E.R. (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **32**, 361–369.

Gao, L.L., Turner, M.K., Chao, S.M., Kolmer, J. and Anderson, J.A. (2016) Genome wide association study of seedling and adult plant leaf rust resistance in elite spring wheat breeding lines. *PLoS ONE,* **11**, e0148671.

Gong, J.Y., Miao, J.S., Zhao, Y., Zhao, Q., Feng, Q., Zhan, Q.L., Cheng, B.Y. *et al*. (2017) Dissecting the genetic basis of grain shape and chalkiness traits in hybrid rice using multiple collaborative populations. *Mol. Plant*, **10**, 1353–1356.

Gowda, M., Zhao, Y., Wurschum, T., Longin, C.F., Miedaner, T., Ebmeyer, E., Schachschneider, R. *et al.* (2014) Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. *Heredity (Edinb)*, **112**, 552–561.

Guo, W., Zhu, X.H., Yan, L.Y. and Qiao, J. (2018) The present and future of whole-exome sequencing in studying and treating human reproductive disorders. *J. Genet. Genom.* **45**, 517–525.

Hayes, P. and Szucs, P. (2006) Disequilibrium and association in barley: Thinking outside the glass. *Proc. Natl. Acad. Sci. USA*, **103**, 18385–18386.

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K. *et al*. (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet*. **51**, 896–904.

Helguera, M., Rivarola, M., Clavijo, B., Martis, M.M., Vanzetti, L.S., Gonzalez, S., Garbus, I. *et al*. (2015) New insights into the wheat chromosome 4D structure and virtual gene order, revealed by survey pyrosequencing. *Plant Sci.* **233**, 200–212.

Henry, I.M., Nagalakshmi, U., Lieberman, M.C., Ngo, K.J., Krasileva, K.V., Vasquez-Gross, H., Akhunova, A. *et al.* (2014) Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. *Plant Cell*, **26**, 1382–1397.

Hill, W.G. and Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theoret. Appl. Genet.* **38**, 226–231.

Hill, W.G. and Weir, B.S. (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* **33**, 54–78.

Huang, L., Brooks, S., Li, W.L., Fellers, J., Nelson, J.C. and Gill, B. (2009) Evolution of new disease specificity at a simple resistance locus in a crop-weed complex: reconstitution of the Lr21 gene in wheat. *Genetics*, **182**, 595–602.

Jacobs, T. and Broers, L.H.M. (1989) The inheritance of host plant effect on latency period of wheat leaf rust in spring wheat. 1. Estimation of gene-action and number of effective factors in F1, F2 and backcross generations. *Euphytica*, **44**, 197–206.

Juliana, P., Singh, R.P., Singh, P.K., Poland, J.A., Bergstrom, G.C., Huerta-Espino, J., Bhavani, S. *et al*. (2018) Genome-wide association mapping for resistance to leaf rust, stripe rust and tan spot in wheat reveals potential candidate genes. *Theoret. Appl. Genet.* **131**, 1405–1422.

Kassa, M.T., You, F.M., Hiebert, C.W., Pozniak, C.J., Fobert, P.R., Sharpe, A.G., Menzies, J.G. *et al*. (2017) Highly predictive SNP markers for efficient selection of the wheat leaf rust resistance gene Lr16. *BMC Plant Biol.* **17**, 45.

Kerber, E.R. and Aung, T. (1999) Leaf rust resistance gene lr34 associated with nonsuppression of stem rust resistance in the wheat cultivar canthatch. *Phytopathology*, **89**, 518–521.

Kertho, A., Mamidi, S., Bonman, J.M., McClean, P.E. and Acevedo, M. (2015) Genome-wide association mapping for resistance to leaf and stripe rust in winter-habit hexaploid wheat landraces. *PLoS ONE*, **10**, e0129580.

Khan, M.H., Bukhari, A., Dar, Z.A. and Rizvi, S.M. (2013) Status and strategies in breeding for rust resistance in wheat. *Agric. Sci.* **4**, 292.

Kim, J.J., Park, Y.M., Baik, K.H., Choi, H.Y., Yang, G.S., Koh, I., Hwang, J.A. *et al*. (2012) Exome sequencing and subsequent association studies identify five amino acid-altering variants influencing human height. *Hum. Genet.* **131**, 471–478.

Krattinger, S.G., Lagudah, E.S., Wicker, T., Risk, J.M., Ashton, A.R., Selter, L.L., Matsumoto, T. *et al*.(2011) Lr34 multi-pathogen resistance ABC transporter: molecular analysis of homoeologous and orthologous genes in hexaploid wheat and other grass species. *Plant J.* **65**, 392–403.

Lande, R. and Thompson, R. (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, **124**, 743–756.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li, H. (2013) *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv preprint arXiv:1303.3997.

Li, Z.F., Xia, X.C., He, Z.H., Li, X., Zhang, L.J., Wang, H.Y., Meng, Q.F. *et al*. (2010) Seedling and slow rusting resistance to leaf rust in Chinese wheat cultivars. *Plant Dis.* **94**, 45–53.

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S. and You, F.M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genom.* **17**, 852.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. and Genome Project Data Processing, S. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu, Y.Q., Wu, H., Chen, H., Liu, Y.L., He, J., Kang, H.Y., Sun, Z.G. *et al*. (2015) A gene cluster encoding lectin receptor kinases confers broad-spectrum and durable insect resistance in rice. *Nat. Biotechnol.* **33**, 301.

Liu, J.D., He, Z.H., Rasheed, A., Wen, W.E., Yan, J., Zhang, P.Z., Wan, Y.X. *et al*. (2017) Genome-wide association mapping of black point reaction in common wheat (*Triticum aestivum* L.). *BMC Plant Biol.* **17**, 220.

Longin, C.F., Gowda, M., Muhleisen, J., Ebmeyer, E., Kazman, E., Schachschneider, R., Schacht, J. *et al*. (2013) Hybrid wheat: quantitative genetic parameters and consequences for the design of breeding programs. *Theoret. Appl. Genet.* **126**, 2791–2801.

Looseley, M.E., Bayer, M., Bull, H., Ramsay, L., Thomas, W., Booth, A., Canto, C.D. *et al*. (2017) Association mapping of diastatic power in UK Winter and spring barley by exome sequencing of phenotypically contrasting variety sets. *Front. Plant Sci*. **8**, 1566.

Lopes, M.S., Dreisigacker, S., Pena, R.J., Sukumaran, S. and Reynolds, M.P. (2015) Genetic characterization of the wheat association mapping initiative (WAMI) panel for dissection of complex traits in spring wheat. *Theor. Appl. Genet.* **128**, 453–464.

Lowe, I., Cantu, D. and Dubcovsky, J. (2011) Durable resistance to the wheat rusts: integrating systems biology and traditional phenotype-based research methods to guide the deployment of resistance genes. *Euphytica*, **179**, 69–79.

Maccaferri, M., Sanguineti, M.C., Mantovani, P., Demontis, A., Massi, A., Ammar, K., Kolmer, J.A. *et al*. (2010) Association mapping of leaf rust response in durum wheat. *Mol. Breed.* **26**, 189–228.

McCallum, B.D., Hiebert, C.W., Cloutier, S., Bakkeren, G., Rosa, S.B., Humphreys, D.G., Marais, G.F. *et al*. (2016) A review of wheat leaf rust research and the development of resistant cultivars in Canada. *Can. J. Plant Path.* **38**, 1–18.

Milner, S.G., Jost, M., Taketa, S., Mazon, E.R., Himmelbach, A., Oppermann, M., Weise, S. *et al*. (2019) Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* **51**, 319.

Mo, Y.J., Howell, T., Vasquez-Gross, H., de Haro, L.A., Dubcovsky, J. and Pearce, S. (2018) Mapping causal mutations by exome sequencing in a wheat TILLING population: a tall mutant case study. *Mol. Genet. Genom.* **293**, 463–477.

Moore, J.W., Herrera-Foessel, S., Lan, C.X., Schnippenkoetter, W., Ayliffe, M., Huerta-Espino, J., Lillemo, M. *et al*. (2015) A recently evolved hexose transporter variant confers resistance to multiple pathogens in wheat. *Nat. Genet.* **47**, 1494–1498.

Mundt, C.C. (2014) Durable resistance: A key to sustainable management of pathogens and pests. *Infect. Genet. Evol.* **27**, 446–455.

Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z.W., Costich, D.E. and Buckler, E.S. (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell*, **21**, 2194–2202.

Navabi, A., Singh, R.P., Tewari, J.P. and Briggs, K.G. (2003) Genetic analysis of adult-plant resistance to leaf rust in five spring wheat genotypes. *Plant Dis.* **87**, 1522–1529.

Nelson, R., Wiesner-Hanks, T., Wisser, R. and Balint-Kurti, P. (2018) Navigating complexity to breed disease-resistant crops. *Nat. Rev. Genet.* **19**, 21–33.

Oliver, R.P. (2014) A reassessment of the risk of rust fungi developing resistance to fungicides. *Pest Manag. Sci.* **70**, 1641–1645.

Van Ooijen, G., Mayr, G., Kasiem, M.M.A., Albrecht, M., Cornelissen, B.J.C. and Takken, F.L.W. (2008) Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *J. Exp. Bot.* **59**, 1383–1397.

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D. *et al*. (2019) Tracing the ancestry of modern bread wheats. *Nat. Genet.* **51**, 905–911.

Reif, J.C., Melchinger, A.E. and Frisch, M. (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Sci.* **45**, 1–7.

Russell, J., Mascher, M., Dawson, I.K., Kyriakidis, S., Calixto, C., Freund, F., Bayer, M. *et al*. (2016) Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nat. Genet.* **48**, 1024.

Schneeberger, K. (2014) Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat. Rev. Genet.* **15**, 662–676.

Sehgal, D., Autrique, E., Singh, R., Ellis, M., Singh, S. and Dreisigacker, S. (2017) Identification of genomic regions for grain yield and yield stability and their epistatic interactions. *Sci. Rep.* **7**, 41578.

Shin, J.-H., Blay, S., McNeney, B. and Graham, J. (2006) LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* **16**, 1–10.

da Silva, G.B.P., Zanella, C.M., Martinelli, J.A., Chaves, M.S., Hiebert, C.W., McCallum, B.D. and Boyd, L.A. (2018) Quantitative trait loci conferring leaf rust resistance in hexaploid wheat. *Phytopathology*, **108**, 1344–1354.

Steuernagel, B., Witek, K., Krattinger, S.G., Ramirez-Gonzalez, R.H., Schoonbeek, H.-J., Yu, G., Baggs, E. *et al*. (2018) Physical and transcriptional organisation of the bread wheat intracellular immune receptor repertoire. *bioRxiv*, 339424.

Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P. and Reynolds, M.P. (2015) Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theoret. Appl. Genet.* **128**, 353–363.

Swarts, K., Li, H.H., Navarro, J.A.R., An, D., Romay, M.C., Hearne, S., Acharya, C. *et al*. (2014) Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *Plant Genome* **7**, 3.

Thind, A.K., Wicker, T., Muller, T., Ackermann, P.M., Steuernagel, B., Wulff, B.B.H., Spannagl, M. *et al*. (2018) Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome dynamics between two wheat cultivars. *Genome Biol.* **19**, 104.

Wang, S.C., Wong, D.B., Forrest, K., Allen, A., Chao, S.M., Huang, B.E., Maccaferri, M. *et al*. (2014) Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **12**, 787–796.

Winfield, M.O., Wilkinson, P.A., Allen, A.M., Barker, G.L.A., Coghill, J.A., Burridge, A., Hall, A. *et al*. (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* **10**, 733–742.

Wurschum, T., Langer, S.M., Longin, C.F.H., Korzun, V., Akhunov, E., Ebmeyer, E., Schachschneider, R. *et al*. (2013) Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. *Theoret. Appl. Genet.* **126**, 1477–1486.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P.C., Hu, L., Yamasaki, M. *et al*. (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* **48**, 927.

Zhang, K.P., Wang, J.J., Zhang, L.Y., Rong, C.W., Zhao, F.W., Peng, T., Li, H.M. *et al*. (2013) Association analysis of genomic loci important for grain weight control in elite common wheat varieties cultivated with variable water and fertiliser supply. *PLoS ONE*, **8**, e57853.

Zhao, Y.S., Gowda, M., Wurschum, T., Longin, C.F.H., Korzun, V., Kollers, S., Schachschneider, R. *et al*. (2013) Dissecting the genetic architecture of frost tolerance in Central European winter wheat. *J. Exp. Bot.* **64**, 4453–4460.

Zhao, Y.S., Li, Z., Liu, G.Z., Jiang, Y., Maurer, H.P., Wurschum, T., Mock, H.P. *et al.* (2015) Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. *Proc. Natl. Acad. Sci. USA*, **112**, 15624–15629.

Zheng, X.W., Gogarten, S.M., Lawrence, M., Stilp, A., Conomos, M.P., Weir, B.S., Laurie, C. *et al*. (2017) SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, **33**, 2251–2257.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Histogram of the mean depth (number of unique reads) of exome capture sequencing.

Figure S2 Histogram of the number of single-nucleotide polymorphisms per gene captured by exome sequencing.

Figure S3 Comparison between Rogers' distances calculated using exome capture sequencing (RD_{EC}) and the 90K SNP array data set (RD_{90K_SNP}) for the 133 parents of the population composed by 1574 wheat hybrids.

Figure S4 Number of single-nucleotide polymorphisms per wheat chromosome captured by exome sequencing.

Figure S5 Genome-wide distribution of average exome nucleotide diversity.

Figure S6 Linkage disequilibrium (r^2) as a function of physical distance (Mb) for the three different wheat subgenomes.

Figure S7 Percentual distribution of marker pairs in each of the three wheat genomes as a function of the physical distance (Mb).

Figure S8 Venn diagram showing the amount of overlapping associations for leaf rust severity found by genome-wide exome association scans in a hybrid wheat population and its corresponding subpopulations.

Figure S9 Extensive genomic region associated with leaf rust resistance on chromosome 3D detected by a genome-wide exome association scan for additive effects in a vast hybrid wheat

population and narrowed down to a smaller candidate region using an independent validation population.

Figure S10 Histogram of the percentages of genetic variance (R^2/h) explained by single-nucleotide polymorphisms found as associated with leaf rust severity in a hybrid wheat population of 1574 hybrids plus their 118 female and 15 male parent lines.

Figure S11 Close-up view of the two association peaks within a candidate region on chromosome 4A found by a genome-wide exome association scan for additive and dominance effects underlying leaf rust severity in a hybrid wheat population of 1574 hybrids plus their 118 female and 15 male parents.

Figure S12 Histogram of the ratio of phenotypic variation explained by main additive and dominance effects ($R^2_{\text{main\_effects}}$) over that explained by interaction effects with environments ($R^2_{\text{interaction\_effects}}$) of single-nucleotide polymorphisms found as associated with leaf rust severity in a hybrid population (1574 hybrids plus their 118 female and 15 male parent lines) tested in five environments.

Figure S13 Genome-wide exome association scans for dominance effects underlying leaf rust severity in a vast hybrid wheat population and its different subpopulations.

Figure S14 Distribution of leaf rust severity scores in the validation population.

Figure S15 Some general statistics in a vast hybrid wheat population and its different subpopulations regarding associations mapping on chromosome 6A.

Figure S16 Some general statistics in a vast hybrid wheat population and its different subpopulations regarding associations mapping on chromosome 6B.

Figure S17 Some general statistics in a vast hybrid wheat population and its different subpopulations regarding associations mapping on chromosome 1A.

Figure S18 Candidate genes underlying significant associations on chromosome 1A.

Table S1 Functional location and type of substitution (synonymous and non-synonymous) within the coding sequence for single-nucleotide polymorphisms (SNPs) present in a hybrid wheat population of 1574 hybrids plus their 118 female and 15 male parent lines.

Table S2 Estimated average physical distance (Mb) according to different wheat chromosomes where the linkage disequilibrium (r^2) fell below $r^2 = 0.2$ in a hybrid wheat population of 1574 hybrids plus their 118 female and 15 male parent lines

Table S3 Expected impact of population stratification during genome-wide exome association scans for leaf rust severity on associations of single-nucleotide polymorphisms (SNPs) mapping close to *Lr10* in a hybrid population of 1574 hybrids plus their 118 female and 15 male parents and its subpopulations $\text{Top}_{25\%}$, $\text{Inferior}_{75\%}$, $\text{Top}_{50\%}$ and $\text{Inferior}_{50\%}$.

Table S4 Estimates of variances components and heritabilities (h^2) for leaf rust severity in a hybrid wheat population composed by 1,574 hybrids produced by factorial crosses between 118 female (F) and 15 male (M) parents.

Table S5 Putative genes underlying significant regions on chromosome 4A.

Haplotype-based genome-wide association increases the predictability of leaf rust (*Puccinia triticina*) resistance in wheat.

Published in:

Journal of Experimental Botany (2020)

DOI: 10.1093/jxb/eraa387

Authors: Fang Liu, Yong Jiang, Yusheng Zhao, Albert W. Schulthess, and Jochen C. Reif

Abstract

Resistance breeding is crucial for sustainable control of wheat leaf rust and single nucleotide polymorphism (SNP)-based genome-wide association studies (GWAS) are widely used to dissect leaf rust resistance. Unfortunately, GWAS based on SNPs often explained only a small proportion of the genetic variation. We compared SNP-based GWAS with a method based on functional haplotypes (FH) considering epistasis in a comprehensive hybrid wheat mapping population composed of 133 parents plus their 1574 hybrids and characterized with 626 245 high-quality SNPs. In total, 2408 and 1 139 828 significant associations were detected in the mapping population by using SNP-based and FH-based GWAS, respectively. These associations mapped to 25 and 69 candidate regions, correspondingly. SNP-based GWAS highlighted two already-known resistance genes, *Lr22a* and *Lr34-B*, while FH-based GWAS detected associations not only on these genes but also on two additional genes, *Lr10* and *Lr1*. As revealed by a second hybrid wheat population for independent validation, the use of detected associations from SNP-based and FH-based GWAS reached predictabilities of 11.72% and 22.86%, respectively. Therefore, FH-based GWAS is not only more powerful for detecting associations, but also improves the accuracy of marker-assisted selection compared with the SNP-based approach.

The original paper has been published and available online:

https://doi.org/10.1093/jxb/eraa387

General Discussion

Genetic variant for GWAS: SNP or haplotype?

Theoretically, the haplotype-based GWAS can provide additional benefits compared to SNP-based GWAS. Some empirical studies indeed demonstrated that haplotype-based GWAS was able to detect loci, which failed to be identified in single SNP-based GWAS (Pryce et al., 2010; Tregouet et al., 2009). However, contrasting results were also reported in previous studies regarding the performance of haplotype-based and SNP-based GWAS (Lorenz et al., 2010; Sato et al., 2016). Whether it is beneficial to use haplotypes as variants in GWAS depends on a case by case basis (Long and Langley, 1999). Thus, it is important to know in which situation haplotype-based GWAS would (or not) be recommended for association analysis. Several hypotheses have been proposed to clarify specific situations in which the haplotype-based approach might be beneficial: 1) haplotype-based GWAS can identify rare variants associated with quantitative traits or complex diseases (Wang and Lin, 2015); 2) haplotype-based association is expected to provide higher power of detection than single marker when the maker density is limited. Especially when medium-density SNP panels were used, haplotype could be in higher LD with the causative variants (Ding et al., 2015); and 3) haplotype-based approach can capture variants associated with structural variants (i.e. deletion, duplication, insertion, inversion, translocation etc.) that could not be accounted for in SNP-based GWAS (Sato et al., 2016).

For empirical data, the genetic architecture are complex and causal loci are usually unknow so that it's difficult to be used to illustrate the underlying reasons. Alternatively, simulation is often used as a methodology for testing hypothesis as it circumvents the problem that the causal loci are usually unknown in reality. However, as the simulated data are usually much simpler and in an ideal situation, the results always act as what they simulated. In this PhD work, we tackled the questions using not only simulations but also the empirical data from Arabidopsis and wheat (Fig. 4). The simulation studies were performed to illustrate under which circumstances FH-based GWAS outperforms SNP-based GWAS considering the influence of minor allele frequency (MAF) and LD (three MAF and LD ranges). The results of simulations revealed that when the MAF is low (MAF belong 0-0.1), both FH- and SNP-based GWAS have low power to detect the causal loci, but the FH-based approach provided higher power than the SNP-based approach when the LD between causal loci were medium (0.3-0.6). This may suggest that haplotype-based approaches could provide potential advantage in detecting rare variants but only in certain scenarios, which is partly consistent with the first hypothesis mentioned above. Besides, FH-based GWAS presented higher power when the LD between those SNPs used to construct haplotypes is low, while haplotypes consisting of SNPs with strong LD could not improve the power compared with the SNP-based GWAS. A possible explanation is that haplotypes cannot provide additional information when the SNPs are with high LD (Pinnaduwage and Briollais, 2005).

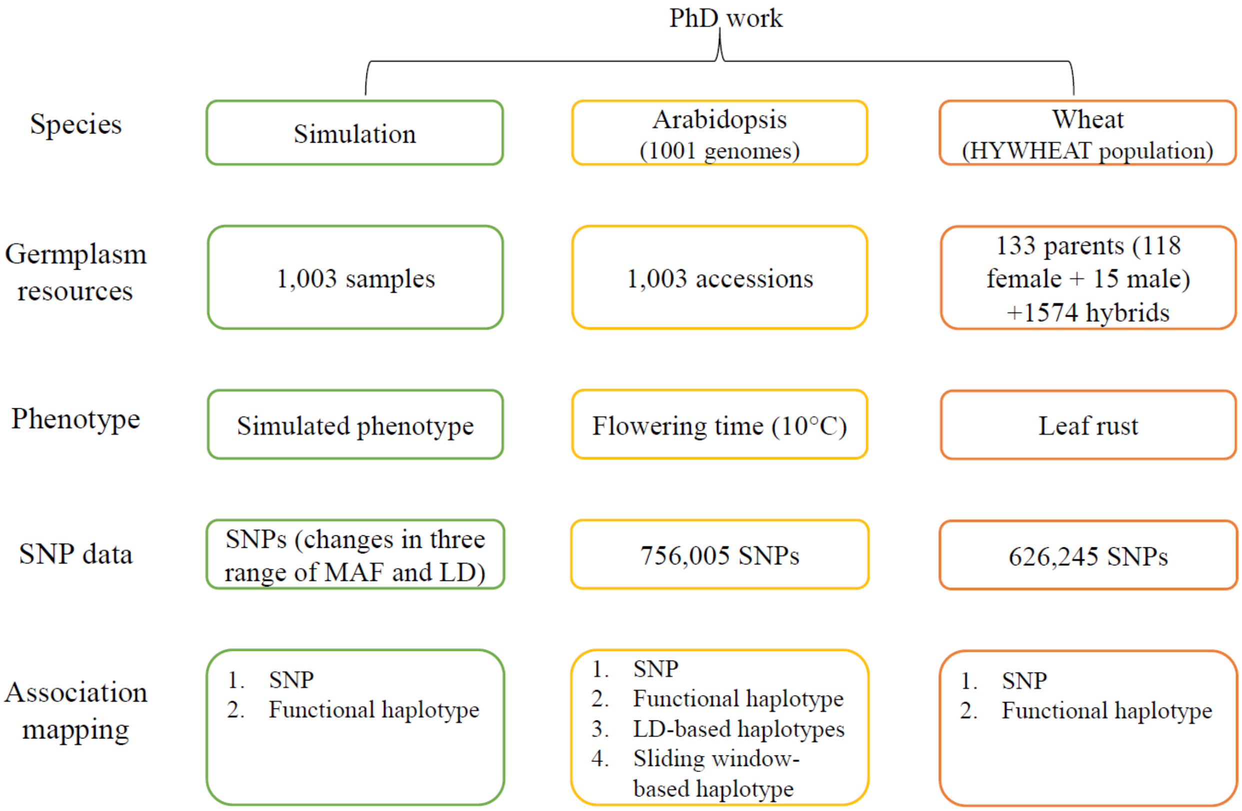


Fig. 4. Outline of the PhD work.

The second hypothesis is about maker density. As the decay of LD may be rapid in the diverse population, high-density markers are usually desired for SNP-based GWAS. In the HYWHEAT population, the decay of LD happened even within the genes (Fig. 5). Thus, in our study, numerous SNPs (626,245 SNPs obtained from exome capture, 48 per Mb) were used for GWAS. Although it was hypothesized that haplotype-based approaches are most beneficial in medium-density marker panels, we still observed the additional superiority of haplotype-based GWAS with respect to the explained phenotypic variation of leaf rust. Even though very high-density markers (756,005 SNPs from WGS, 6,353 per Mb) were used in Arabidopsis, the advantage of haplotype-based GWAS was observed again in the empirical data of flowering time. This suggested that despite using high-density marker panels, haplotype-based GWAS can still be performed as a complementary approach of SNP-based GWAS to find new marker-traits associations. In contrast, if the marker density is too low, there might be no difference between SNP-based and haplotype-based GWAS, because in this situation both approaches have low power (Pinnaduwage and Briollais, 2005).

Testing the third hypothesis is important but challenging because structural variation can frequently have functional impact on gene structure or dosage, however, the accurate calling and genotyping of structural variants in an individual genome is typically more challenging than those of SNPs (Fuentes et al., 2019). Some studies have revealed that haplotype-based GWAS can efficiently capture variants linked with the structural variants, such as copy number variants (Zhang et al., 2012). In our study, the FH-based GWAS found additional new candidate regions associated with leaf rust in wheat. Some of the candidate haplotypes are significant due to the cumulative additive effect or the local epistatic interaction of the SNPs or both, while others are in more complex scenario that the cumulative additive or epistatic effects do not exhibit great difference. This might possibly indicate that those haplotypes are linked with structural variants. However, further studies, e.g. providing the exact sequences around the significant regions, are needed to tackle the problem.

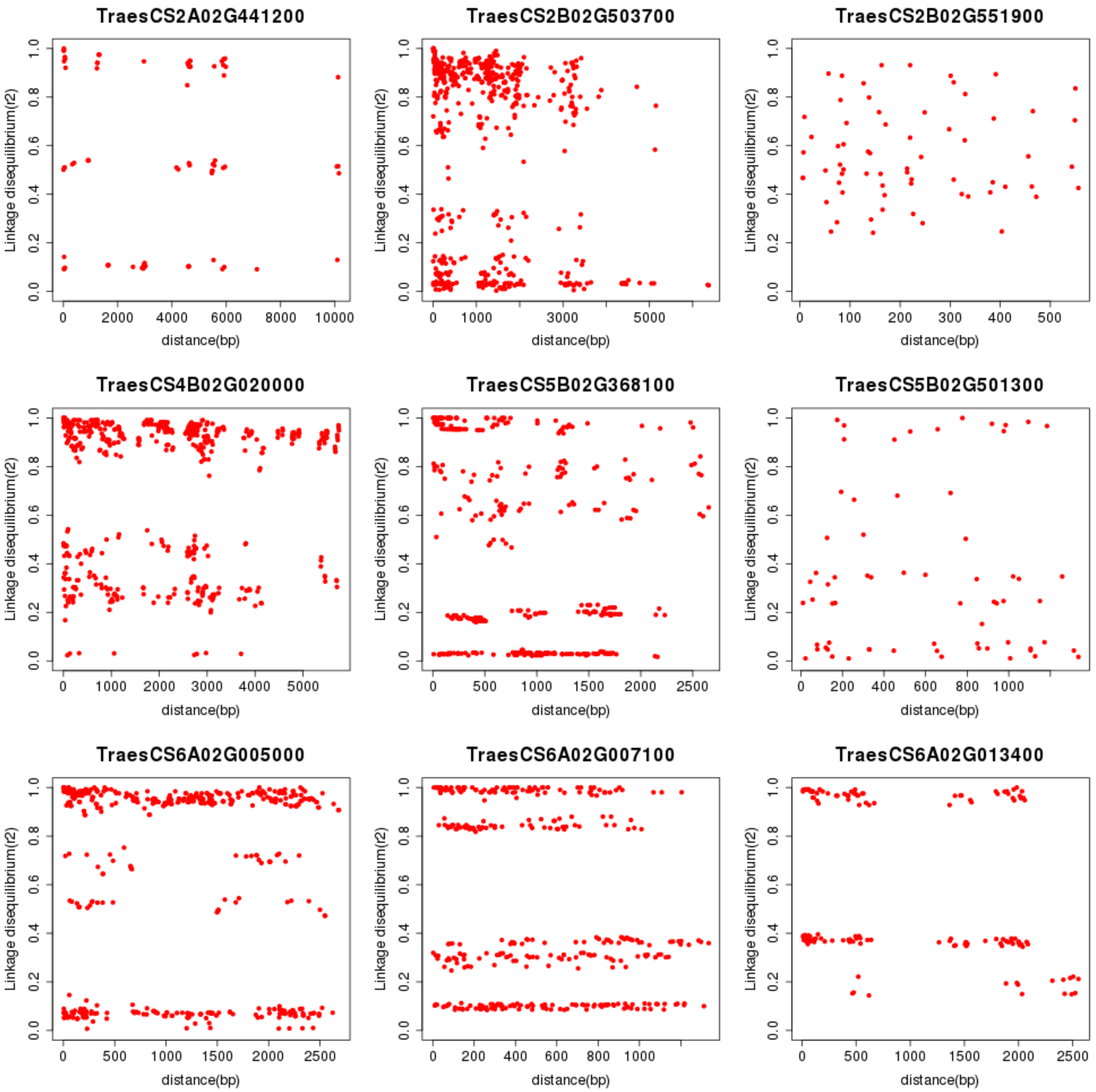


Fig. 5. Linkage disequilibrium among markers for a selected set of nine sequenced genes in the HYWHEAT population.

Functional haplotype and other haplotype based GWAS approaches

In this study, we have compared the developed FH-based GWAS approach with other two commonly used haplotype-based GWAS approaches (sliding-window-based haplotypes and LD-based haplotypes). We found that FH-based GWAS outperformed the other two approaches in terms of the power of detection, i.e. FH-based GWAS detected more significant regions. In particular, we found a novel candidate region on chromosome 4 associated with the trait flowering time in Arabidopsis. The possible reasons why FH-based GWAS performed better can be summarized as follows: 1) FH-based GWAS constructs haplotypes allowing for non-contiguous SNPs, which excludes the noise from trait-unrelated markers locating between the associated SNPs; 2) Additive as well as local epistatic effects are taken into account for haplotype construction, thus the significance of a haplotype effect can be solely a result of significant epistatic effects, or cumulative (non-)significant additive and epistatic effects among the SNPs. However, the application of FH-based GWAS might be hampered by the computational load which is highly dependent on the marker density and the chosen window size. For the data set with very high marker density, a mild preselection of SNPs according to their additive effects can greatly reduce the computational load for the remaining steps. Determining an appropriate window size is critical but challenging. A too small window size can lead to high LD among markers within the window, resulting in low efficiency of the FH-based GWAS as shown in our simulation study, while an oversized window may yield functional haplotypes spanning large genomic regions with many candidate genes.

In addition to the LD-based haplotype approach considered in this study, many other LD-based approaches were used in construction of haplotypes. For example, Haploview is a tool for haplotype pattern analysis and visualization, including three LD-based methods to define haplotypes (Barrett et al., 2005). The first method defines haplotype blocks using a 95% confidence bound for the LD measure D' (Gabriel et al., 2002). The second method was proposed by Wang and coauthors (Wang et al., 2002) and defines LD blocks as a set of contiguous and ordered SNP markers without evidence for recombination events that are identified by performing four-gamete test between each pairwise SNPs. The third method constructs LD blocks using the algorithm 'Solid Spine', in which the haplotype blocks comprise all the SNPs that are in strong LD with the first and last SNPs in the block (Barrett et al., 2005). However, all these three methods fail to consider possible correlation among LD blocks. Namely, they do not allow intermediate regions of low LD between strongly associated SNP pairs and tend to split them in to small blocks without considering the high correlation between LD blocks. In this regard, Kim and coauthors (Kim et al., 2018) proposed a new method called Big-LD that is implemented in R. This method overcomes the problem of interrupting low LD SNPs in the middle of high LD region by an agglomerative approach that firstly identify small communities of SNPs using the CLQ-D algorithm (a modified version of LD bin construction algorithm CLQ) (Yoo et al., 2015) and then merge these communities with high correlation. However, the haplotype blocks implemented in these LD-based approaches are still a subset of contiguous SNPs with high LD, which may lead to redundant information included in the haplotype. Besides, as shown in the simulation studies, the strong LD-based haplotypes did not perform better than single SNPs in GWAS. Therefore, they are not expected to outperform our FH-based GWAS approach.

Another haplotype-based GWAS considered in this study is based on fixed-size sliding window. The choice of window size is critical because too small window size results in loss of information whereas too large window size may introduce excessive noises. Furthermore, when the fixed window size is used for whole genome analysis, it may be problematic because the LD pattern changes in different regions. Alternatively, variable-length Markov chains (VLMCs) and variable-size sliding window avoid the problem of choosing an appropriate window size. The VLMCs method proposed by Browning (Browning, 2006) can automatically balance the degree of LD between markers and number of tests with maximal information extraction, which improves the power of detecting associations. For variable-size sliding window (Li et al., 2007), the maximum size of a sliding window is determined by the local haplotype diversity and the sample size. Variable-size sliding window should outperform the fixed-size sliding window especially for large-scale haplotype analyses (such as a whole-genome scan) where the LD patterns are expected to vary widely. Additionally, HapConstructor (Abo et al., 2008) constructs multi-loci SNP sets as haplotypes through a forward-backward stepwise process. Due to the high computational burden of exhaustive process, it has only been applied in the analysis of candidate regions instead of haplotype-based GWAS. HaploBlocker (Pook et al., 2019) defines and infers haplotype blocks based on linkage instead of the commonly used population-wide measures of LD. Overall, these haplotype-based approaches have different advantages and limitations (summarized in Table 2).

Table 2. Comparison of different haplotype-based approaches.

Method	Software/code	Advantages	Limitations	Reference
confidence interval method	Haploview	simple (using 95% confidence bound for LD measure D')	Not consider possible correlation among LD block	Gabriel et al., 2002
four-gamete test	Haploview	Consider evidence for recombination using four-gamete test	Not consider possible correlation among LD block	Wang et al., 2002
Solid-Spine of LD	Haploview	the first and last markers are in strong LD with all intermediate markers	Not consider possible correlation among LD block	Barrett et al., 2005
Big-LD	R package	considering possible correlation among LD block	Only construct contiguous SNP subsets for haplotypes	Kim et al., 2018
VLMCs (variable-length Markov chains)	R package	Automatically balance the degree of freedom and number of tests	Problem may happen for the data with complex LD patterns	Browning, 2006
variable-size sliding window	Code in R	window size depends on the basis of local haplotype diversity as well as sample size	detection power for the region with relatively low LD is low (<40%)	Li et al., 2007
hapConstructor	hapConstructor	allowing for non-contiguous SNPs for haplotype construction	Only for candidate regions owing to computational burden	Abo et al., 2008
HaploBlocker	R package	allowing for non-contiguous SNPs for haplotype construction	For phase-unknown genotype data, it needs additional phasing step	Pook et al., 2019
FH (functional haplotypes)	Code in R	allowing for non-contiguous SNPs for haplotype construction	Preselection is need for high marker-density data	Liu et al., 2019

Data mining considering population structure helps to increase the power of GWAS

The statistical power for QTL detection in association mapping are influenced by several factors including population size, population structure, allele frequency and effect size of the causal variant (Shin and Lee, 2015). Population structure can decrease power and are prone to cause false-positive results. Thus, many methods were proposed to correct for population structure, such as the method using the Q matrix (treating population structure as fix effect that can be obtained by genomic control (GC), structured association analysis (SA) or principal component analysis (PCA)), the MLM using the K (kinship) matrix or the Q+K model (Xu et al., 2017). Currently, although MLM has been successfully used to account for population structure, it may also mask true QTLs that are strongly correlated with population structure. It has been widely accepted that correcting for population structure in GWAS will cause many false negatives for such confounded traits (Huang and Han, 2014). For instance, only one gene (*ZmCCT*) was revealed in the GWAS of flowering time using a diverse maize association population (consisting of 500 inbred lines) because flowering time is a typical adaptive trait and is always confounded with population structure (Yang et al., 2013). Thus, for the traits confounded with population structure, it is not possible to identify marker-traits association using GWAS with routine MLM. One of possible solutions is to attempt different statistical models to explore the confounded traits. For example, Fixed and random model Circulating Probability Unification (FarmCPU) is based on a multiple loci linear mixed model with two parts: fixed effect model and a random effect model, controlling false positives and simultaneously reducing both false negatives and computing time (Liu et al., 2016), while Quantitative Trait Cluster Association Test (QTCAT) overcomes the need for population structure correction and controls false positives by accounting for the correlation between the markers (Klasen et al., 2016).

Another reasonable solution is subdividing the diverse population into subpopulations and analyzing each subpopulation independently. Taking rice for example, as we known that the genetic architecture in rice is quite different between the two subspecies: *indica* and *japonica*. Thus, Chen and coauthors (Chen et al., 2014) analyzed metabolites within different subpopulation and found that many loci under genetic control were distinct in the different subspecies. In our study of leaf rust in the total HYWHEAT population, the distribution of phenotypic values of leaf rust severity is not normal (which is desired in our statistical model conducting GWAS) but bimodal, and the correlation between genetic and phenotypic distances ($r_{RD,PD}$) is moderate (r=0.23, P=0.001) evaluated using all the parents. In this respect, the strategy is to divide the total hybrid population into four different subpopulations based on the bimodal distribution of parents. As expected, $r_{RD,PD}$ in subpopulations $Inferior_{75\%}$, $Top_{50\%}$ and $Inferior_{50\%}$ were decreased compared to that of the total population. Subsequently, the GWAS in subpopulations found additional new candidate regions that were close to some of the known resistant genes. Notably, the loci that were only detected in subpopulation using SNP-based GWAS were found again in FH-based GWAS using the total population. A possible reason is that the population structure is estimated using SNPs instead of haplotypes, suggesting that haplotype-based GWAS may provide a solution to detect the masked QTLs confounded with population structure.

To summarize, there is no single best GWAS method which is sufficient to dissect the genetic architecture of complex traits in all different situations. The FH-based GWAS is a promising and powerful complementary approach to the standard SNP-based approach.

Outlook

Despite great efforts to remove undesired noise (such as the different models for GWAS and the using of fairly stringent thresholds), false-positive association can still occur in GWAS due to the enormous number of statistical test and other unaccounted factors, i.e., small population size, rare allele frequency, inaccurate genotype calling at some variants. This calls for an independent validation process, which requires substantial amount of time (or labor) and is therefore often avoided to be incorporated into GWAS design. Such validation process can be functional validation (including genetic complementation, candidate gene over-expression and knock-out etc.) or independent validation using a different population. For independent validation, independence and choice of the validation population is critical. Firstly, the population size should be taken under consideration to provide sufficient statistic power. Otherwise, the significant markers or associations may fail to validate just because of the limited power of small population. Secondly, the relatedness between validation and detection population should also be cautiously considered to avoid the same population structure and genetic background. For instance, if a significant marker is detected in a parental population of wheat, and then the marker is validated in their offspring (i.e., F_1 population or F_2 population). In this case, there is high possibility that the two population have the same population structure and genetic background, which may lead to apparent validation of what is actually a false positive result. A convincing independent validation should be based on an appropriate validation with independent pedigree structure. The traditional aim of validation is to decrease false positive association and increase the level of trust. Our study revealed that validation could also benefit from narrowing down the candidate regions because the LD phase was different between detection and validation population, which would accelerate the speed of searching candidate genes. Thus, we suggest to incorporate validation process into future GWAS design.

Summary

In facing the challenge of meeting the demand of a growing population for sufficient food in times of climate change, plant breeders are striving to develop improved crops with higher productivity and better adaptability. A deeper understanding of the genetic basis for important traits is the key to marker-assisted selection (MAS), which allows efficient crop improvement. Currently, genome-wide association mapping (GWAS) based on single nucleotide polymorphisms (SNPs) is the most commonly used method to study the genetic architecture of complex traits. However, a single SNP explains only a small proportion of the genetic variation. In addition, epistatic interactions, which play a critical role in the regulation of many complex traits in plants, are difficult to detect by single SNP-based GWAS. Haplotype-based GWAS is a good alternative, since haplotypes can be used to take into account local epistasis between the SNPs. Therefore, a functional haplotype-based GWAS (FH-based GWAS) was developed in the context of this PhD thesis, which selects SNPs based on additive and epistatic effects in order to construct haplotypes. To test the performance of the FH-based GWAS, simulation studies were performed. These studies showed that FH-based GWAS outperformed SNP-based GWAS at a higher minor allele frequency (MAF) and lower linkage disequilibrium (LD) between the SNPs. Secondly, empirical data on the flowering time of the model plant *Arabidopsis* was used to compare FH-based GWAS with other approaches, including SNP-based and two further haplotype-based GWAS. Using FH-based GWAS, all candidate regions were identified that were also detected in SNP-based and two other haplotype-based approaches. Besides, a novel region on chromosome 4 was detected exclusively by FH-based GWAS. Thirdly, the FH-based GWAS was tested with data from the important crop wheat to study resistance to leaf rust, one of the most widespread diseases of wheat. In addition to the region found by SNP-based GWAS, new candidate regions were discovered by the FH-based GWAS. Furthermore, an independent validation showed that the predictabilities with FH-based GWAS was nearly doubled compared to SNP-based GWAS. In conclusion, FH-based GWAS offers a higher detection power compared to the SNP-based approach and can also improve the predictability and accuracy of MAS. Consequently, FH-based GWAS is a powerful approach to increase the selection gain in breeding.

Zusammenfassung

Angesichts der Herausforderung, den Bedarf einer wachsenden Bevölkerung nach genügend Nahrungsmittel in Zeiten des Klimawandels zu decken, streben Pflanzenzüchter danach, verbesserte Nutzpflanzen mit höherer Produktivität und besserer Anpassungsfähigkeit zu entwickeln. Ein vertieftes Verständnis der genetischen Basis wichtiger Merkmale ist der Schlüssel zur markergestützten Selektion (MAS), die es erlaubt Nutzpflanzen effizient zu verbessern. Gegenwärtig ist die genomweite Assoziationskartierung (GWAS), die auf einzelnen Nukleotidpolymorphismen (SNPs) basiert, die am häufigsten verwendete Methode, um die genetische Basis komplexer Merkmale zu untersuchen. Ein einziger SNP erklärt jedoch nur einen kleinen Teil der genetischen Variation. Hinzu kommt, dass epistatische Interaktionen, die bei der Regulation vieler komplexer Merkmale in Pflanzen eine entscheidende Rolle spielen, durch eine einzelne SNP-basierte GWAS nur schwer erfasst werden können. Alternativ bietet sich eine haplotypbasierte GWAS an, da durch Haplotypen die lokale Epistasie zwischen SNPs berücksichtigt werden kann. Daher wurde im Rahmen dieser Doktorarbeit eine funktionelle haplotypbasierte GWAS (FH-basierte GWAS) entwickelt, welche SNPs auf der Grundlage additiver und epistatischer Effekte auswählt, um Haplotypen zu konstruieren. Um die Leistungsfähigkeit der FH-basierten GWAS zu testen, wurden Simulationsstudien durchgeführt. Diese zeigen, dass GWAS auf FH-Basis bei einer höheren Frequenz kleinerer Allele (MAF) und eines geringeren Kopplungsungleichgewichts (LD) zwischen den SNPs deutlich besser abschneidet als GWAS auf SNP-Basis. Zweitens wurde mittels empirischer Daten zum Blühzeitpunkt der Modellpflanze *Arabidopsis* die FH-basierte GWAS mit anderen Ansätzen verglichen, zu denen die SNP-basierte und zwei weitere haplotypische GWAS gehören. Mittels FH-basierter GWAS wurden alle Kandidatenregionen identifiziert, die auch in SNP-basierten und den zwei anderen haplotyp-basierten Ansätzen erkannt wurden. Außerdem wurde eine neue Region auf Chromosom 4 ausschließlich durch FH-basierte GWAS nachgewiesen. Drittens wurde die FH-basierte GWAS mit Daten der wichtigen Kulturpflanze Weizen getestet, um die Resistenz gegen Blattrost, eine der am weitesten verbreiteten Krankheiten von Weizen, zu untersuchen. Zusätzlich zu der Region, die durch die SNP-basierte GWAS gefunden wurde, konnte durch die FH-basierte GWAS neue Kandidatenregionen entdeckt werden. Darüber hinaus ergab eine unabhängige Validierung, dass die Vorhersagbarkeiten mit FH-basierte GWAS im Vergleich zur SNP-basierten

GWAS nahezu verdoppelt werden konnten. Zusammenfassend lässt sich sagen, dass FH-basierte GWAS im Vergleich zum SNP-basierten Ansatz eine höhere Detektionskraft bietet und auch die Vorhersagbarkeit und Genauigkeit der MAS verbessern kann. Folglich ist FH-basierte GWAS ein leistungsfähiger Ansatz, um den Selektionsgewinn in der Züchtung zu steigern.

Reference

1001 Genomes Consortium (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166, 481-491.

Abo, R., Knight, S., Wong, J., Cox, A. and Camp, N.J. (2008) hapConstructor: automatic construction and testing of haplotypes in a Monte Carlo framework. Bioinformatics 24, 2105-2107.

Akdemir, D. and Jannink, J.L. (2015) Locally Epistatic Genomic Relationship Matrices for Genomic Association and Prediction. Genetics 199, 857-871.

Alqudah, A.M., Haile, J.K., Alomari, D.Z., Pozniak, C.J., Kobiljski, B. and Borner, A. (2020) Genome-wide and SNP network analyses reveal genetic control of spikelet sterility and yield-related traits in wheat. Sci Rep 10, 2098.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C.J., Stein, N., Choulet, F., Distelfeld, A., et al. (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, 661-+.

Atwell, S., Huang, Y.S., Vilhjalmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465, 627-631.

Barbier, E.B. and Hochard, J.P. (2018) The impacts of climate change on the poor in disadvantaged regions. Review of Environmental Economics and Policy 12, 26-47.

Bardel, C., Danjean, V., Hugot, J.P., Darlu, P. and Genin, E. (2005) On the use of haplotype phylogeny to detect disease susceptibility loci. Bmc Genet 6.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263-265.

Boeven, P.H.G., Longin, C.F.H., Leiser, W.L., Kollers, S., Ebmeyer, E. and Wurschum, T. (2016) Genetic architecture of male floral traits required for hybrid wheat breeding. Theor Appl Genet 129, 2343-2357.

Breseghello, F. and Coelho, A.S. (2013) Traditional and modern plant breeding methods with examples in rice (Oryza sativa L.). J Agric Food Chem 61, 8277-8286.

Browning, S.R. (2006) Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 78, 903-913.

Cardon, L.R. and Abecasis, G.R. (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19, 135-140.

Chan, E.K.F., Rowe, H.C., Corwin, J.A., Joseph, B. and Kliebenstein, D.J. (2011) Combining Genome-Wide Association Mapping and Transcriptional Networks to Identify Novel Genes Controlling Glucosinolates in Arabidopsis thaliana. Plos Biol 9.

Chao, D.Y., Silva, A., Baxter, I., Huang, Y.S., Nordborg, M., Danku, J., Lahner, B., Yakubova, E. and Salt, D.E. (2012) Genome-Wide Association Studies Identify Heavy Metal ATPase3 as the Primary Determinant of Natural Variation in Leaf Cadmium in Arabidopsis thaliana. Plos Genet 8.

Chen, W., Gao, Y.Q., Xie, W.B., Gong, L., Lu, K., Wang, W.S., Li, Y., Liu, X.Q., Zhang, H.Y., Dong, H.X., et al. (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nature Genetics 46, 714-721.

Cloutier, S., McCallum, B.D., Loutre, C., Banks, T.W., Wicker, T., Feuillet, C., Keller, B. and Jordan, M.C. (2007) Leaf rust resistance gene Lr1, isolated from bread wheat (Triticum aestivum L.) is a member of the large psr567 gene family. Plant Mol Biol 65, 93-106.

da Silva, G.B.P., Zanella, C.M., Martinelli, J.A., Chaves, M.S., Hiebert, C.W., McCallum, B.D. and Boyd, L.A. (2018) Quantitative Trait Loci Conferring Leaf Rust Resistance in Hexaploid Wheat. Phytopathology 108, 1344-1354.

Daly, M.J., Rioux, J.D., Schaffner, S.E., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. Nature Genetics 29, 229-232.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12, 499-510.

Ding, J.Q., Ali, F., Chen, G.S., Li, H.H., Mahuku, G., Yang, N., Narro, L., Magorokosho, C., Makumbi, D. and Yan, J.B. (2015) Genome-wide association mapping reveals novel sources of resistance to northern corn leaf blight in maize. Bmc Plant Biology 15.

Doust, A.N., Lukens, L., Olsen, K.M., Mauro-Herrera, M., Meyer, A. and Rogers, K. (2014) Beyond the single gene: How epistasis and gene-by-environment effects influence crop domestication. P Natl Acad Sci USA 111, 6178-6183.

Feuillet, C., Travella, S., Stein, N., Albar, L., Nublat, A. and Keller, B. (2003) Map-based isolation of the leaf rust disease resistance gene Lr10 from the hexaploid wheat (Triticum aestivum L.) genome. P Natl Acad Sci USA 100, 15253-15258.

Filiault, D.L. and Maloof, J.N. (2012) A genome-wide association study identifies variants underlying the Arabidopsis thaliana shade avoidance response. Plos Genet 8, e1002589.

Francisco, M., Joseph, B., Caligagan, H., Li, B.H., Corwin, J.A., Lin, C., Kerwin, R.E., Burow, M. and Kliebenstein, D.J. (2016) Genome Wide Association Mapping in Arabidopsis thaliana Identifies Novel Genes Involved in Linking Allyl Glucosinolate to Altered Biomass and Defense. Front Plant Sci 7.

Fuentes, R.R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J.F., Mohiyuddin, M., Wing, R.A., McNally, K.L., Tatarinova, T., Grigoriev, A., et al. (2019) Structural variants in 3000 rice genomes. Genome Res 29, 870-880.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002) The structure of haplotype blocks in the human genome. Science 296, 2225-2229.

Goddard, M. (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245-257.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M. and Toulmin, C. (2010) Food Security: The Challenge of Feeding 9 Billion People. Science 327, 812-818.

Gowda, M., Zhao, Y., Wuerschum, T., Longin, C.F.H., Miedaner, T., Ebmeyer, E., Schachschneider, R., Kazman, E., Schacht, J., Martinant, J.P., et al. (2014) Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. Heredity 112, 552-561.

Gu, S.M., Han, P., Ye, Z.P., Perkins, L.E., Li, J., Wang, H.Q., Zalucki, M.P. and Lu, Z.Z. (2018) Climate change favours a destructive agricultural pest in temperate regions: late spring cold matters. J Pest Sci 91, 1191-1198.

Guo, Y.F., Li, J., Bonham, A.J., Wang, Y.P. and Deng, H.W. (2009) Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. Eur J Hum Genet 17, 785-792.

Haberer, G., Young, S., Bharti, A.K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing, R.A., Rounsley, S., Birren, B., et al. (2005) Structure and architecture of the maize genome. Plant Physiol 139, 1612-1624.

Hartung, F. and Schiemann, J. (2014) Precise plant breeding using new genome editing techniques: opportunities, safety and regulation in the EU. Plant J 78, 742-752.

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K., Fritz, A., Hucl, P., Wiebe, K., et al. (2019a) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nature Genetics 51, 896-904.

He, S., Reif, J.C., Korzun, V., Bothe, R., Ebmeyer, E. and Jiang, Y. (2017) Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to Central Europe. Theor Appl Genet 130, 635-647.

He, T., Hill, C.B., Angessa, T.T., Zhang, X.Q., Chen, K., Moody, D., Telfer, P., Westcott, S. and Li, C. (2019b) Gene-set association and epistatic analyses reveal complex gene interaction networks affecting flowering time in a worldwide barley collection. Journal of experimental botany 70, 5603-5616.

Huang, B.E., Amos, C.I. and Lin, D.Y. (2007) Detecting haplotype effects in genomewide association studies. Genet Epidemiol 31, 803-812.

Huang, B.E., Verbyla, K.L., Verbyla, A.P., Raghavan, C., Singh, V.K., Gaur, P., Leung, H., Varshney, R.K. and Cavanagh, C.R. (2015) MAGIC populations in crops: current status and future prospects. Theor Appl Genet 128, 999-1017.

Huang, L., Brooks, S., Li, W.L., Fellers, J., Nelson, J.C. and Gill, B. (2009) Evolution of New Disease Specificity at a Simple Resistance Locus in a Crop-Weed Complex: Reconstitution of the Lr21 Gene in Wheat. Genetics 182, 595-602.

Huang, X.H. and Han, B. (2014) Natural Variations and Genome-Wide Association Studies in Crop Plants. Annu Rev Plant Biol 65, 531-551.

IWGSC, International Wheat Genome Sequencing Consortium, Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., investigators, I.w.-g.a.p., Pozniak, C.J., et al. (2018)

Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361.

Jamann, T.M., Balint-Kurti, P.J. and Holland, J.B. (2015) QTL mapping using high-throughput sequencing. In: *Plant Functional Genomics* pp. 257-285. Springer.

Jiang, Y., Schmidt, R.H., Zhao, Y.S. and Reif, J.C. (2017) A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. Nature Genetics 49, 1741-1746.

Jiang, Y., Schmidt, R.H. and Reif, J.C. (2018) Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers. G3 (Bethesda) 8, 1687-1699.

Juliana, P., Singh, R.P., Singh, P.K., Poland, J.A., Bergstrom, G.C., Huerta-Espino, J., Bhavani, S., Crossa, J. and Sorrells, M.E. (2018) Genome-wide association mapping for resistance to leaf rust, stripe rust and tan spot in wheat reveals potential candidate genes. Theor Appl Genet 131, 1405-1422.

Kassa, M.T., You, F.M., Hiebert, C.W., Pozniak, C.J., Fobert, P.R., Sharpe, A.G., Menzies, J.G., Humphreys, D.G., Harrison, N.R., Fellers, J.P., et al. (2017) Highly predictive SNP markers for efficient selection of the wheat leaf rust resistance gene Lr16. BMC Plant Biology 17, 45.

Khan, M.H., Bukhari, A., Dar, Z.A. and Rizvi, S.M. (2013) Status and strategies in breeding for rust resistance in wheat. Agricultural Sciences 4, 292.

Kim, H., Grueneberg, A., Vazquez, A.I., Hsu, S. and de los Campos, G. (2017) Will Big Data Close the Missing Heritability Gap? Genetics 207, 1135-1145.

Kim, S.A., Cho, C.S., Kim, S.R., Bull, S.B. and Yoo, Y.J. (2018) A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. Bioinformatics 34, 388-397.

Klasen, J.R., Barbez, E., Meier, L., Meinshausen, N., Buhlmann, P., Koornneef, M., Busch, W. and Schneeberger, K. (2016) A multi-marker association method for genome-wide association studies without the need for population structure correction. Nat Commun 7.

Knuppel, S., Esparza-Gordillo, J., Marenholz, I., Holzhutter, H.G., Bauerfeind, A., Ruether, A., Weidinger, S., Lee, Y.A. and Rohde, K. (2012) Multi-locus stepwise regression: a haplotype-based algorithm for finding genetic associations applied to atopic dermatitis. Bmc Med Genet 13.

Korte, A., Vilhjalmsson, B.J., Segura, V., Platt, A., Long, Q. and Nordborg, M. (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44, 1066-1071.

Krattinger, S.G., Lagudah, E.S., Wicker, T., Risk, J.M., Ashton, A.R., Selter, L.L., Matsumoto, T. and Keller, B. (2011) Lr34 multi-pathogen resistance ABC transporter: molecular analysis of homoeologous and orthologous genes in hexaploid wheat and other grass species. Plant J 65, 392-403.

Laramie, J.M., Wilk, J.B., DeStefano, A.L. and Myers, R.H. (2007) HaploBuild: an algorithm to construct non-contiguous associated haplotypes in family based genetic studies. Bioinformatics 23, 2190-2192.

Lesk, C., Rowhani, P. and Ramankutty, N. (2016) Influence of extreme weather disasters on global crop production. Nature 529, 84-+.

Li, Y., Sung, W.K. and Liu, J.J. (2007) Association mapping via regularized regression analysis of single-nucleotide-polymorphism haplotypes in variable-sized sliding windows. Am J Hum Genet 80, 705-715.

Lin, S., Chakravarti, A. and Cutler, D.J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nature Genetics 36, 1181-1188.

Liu, F., Schmidt, R.H., Reif, J.C. and Jiang, Y. (2019) Selecting Closely-Linked SNPs Based on Local Epistatic Effects for Haplotype Construction Improves Power of Association Mapping. G3-Genes Genom Genet 9, 4115-4126.

Liu, F., Zhao, Y.S., Beier, S., Jiang, Y., Thorwarth, P., Longin, C.F.H., Ganal, M., Himmelbach, A., Reif, J.C. and Schulthess, A.W. (2020a) Exome association analysis sheds light onto leaf rust (Puccinia triticina) resistance genes currently used in wheat breeding (Triticum aestivum L.). Plant Biotechnol J 18, 1396-1408.

Liu, F., Jiang, Y., Zhao, Y., Schulthess, A.W. and Reif, J.C. (2020b) Haplotype-based genome-wide association increases the predictability of leaf rust (Puccinia triticina) resistance in wheat. Journal of experimental botany.

Liu, N.J., Zhang, K. and Zhao, H.Y. (2008) Haplotype-Association Analysis. Adv Genet 60, 335-405.

Liu, X.L., Huang, M., Fan, B., Buckler, E.S. and Zhang, Z.W. (2016) Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. Plos Genet 12.

Long, A.D. and Langley, C.H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9, 720-731.

Lorenz, A.J., Hamblin, M.T. and Jannink, J.L. (2010) Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. Plos One 5, e14079.

Lowe, I., Cantu, D. and Dubcovsky, J. (2011) Durable resistance to the wheat rusts: integrating systems biology and traditional phenotype-based research methods to guide the deployment of resistance genes. Euphytica 179, 69-79.

Lujan Basile, S.M., Ramirez, I.A., Crescente, J.M., Conde, M.B., Demichelis, M., Abbate, P., Rogers, W.J., Pontaroli, A.C., Helguera, M. and Vanzetti, L.S. (2019) Haplotype block analysis of an Argentinean hexaploid wheat collection and GWAS for yield components and adaptation. BMC Plant Biol 19, 553.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009) Finding the missing heritability of complex diseases. Nature 461, 747-753.

Manolio, T.A. (2013) Bringing genome-wide association findings into clinical use. Nat Rev Genet 14, 549-558.

Mathias, R.A., Gao, P., Goldstein, J.L., Wilson, A.F., Pugh, E.W., Furbert-Harris, P., Dunston, G.M., Malveaux, F.J., Togias, A., Barnes, K.C., et al. (2006) A graphical assessment of p-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. Bmc Genet 7, 38.

McCallum, B.D., Hiebert, C.W., Cloutier, S., Bakkeren, G., Rosa, S.B., Humphreys, D.G., Marais, G.F., McCartney, C.A., Panwar, V., Rampitsch, C., et al. (2016) A review of wheat leaf rust research and the development of resistant cultivars in Canada. Can J Plant Pathol 38, 1-18.

Meng, L.J., Guo, L.B., Ponce, K., Zhao, X.Q. and Ye, G.Y. (2016) Characterization of Three Indica Rice Multiparent Advanced Generation Intercross (MAGIC) Populations for Quantitative Trait Loci Identification. Plant Genome-Us 9.

Meuwissen, T.H.E. and Goddard, M.E. (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155, 421-430.

Mirdita, V., Liu, G.Z., Zhao, Y.S., Miedaner, T., Longin, C.F.H., Gowda, M., Mette, M.F. and Reif, J.C. (2015) Genetic architecture is more complex for resistance to Septoria tritici blotch than to Fusarium head blight in Central European winter wheat. Bmc Genomics 16.

Mo, Y.J., Howell, T., Vasquez-Gross, H., de Haro, L.A., Dubcovsky, J. and Pearce, S. (2018) Mapping causal mutations by exome sequencing in a wheat TILLING population: a tall mutant case study. Mol Genet Genomics 293, 463-477.

Moore, J.W., Herrera-Foessel, S., Lan, C.X., Schnippenkoetter, W., Ayliffe, M., Huerta-Espino, J., Lillemo, M., Viccars, L., Milne, R., Periyannan, S., et al. (2015) A recently evolved hexose transporter variant confers resistance to multiple pathogens in wheat. Nature Genetics 47, 1494-1498.

Morris, R.W. and Kaplan, N.L. (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23, 221-233.

Muhleisen, J., Piepho, H.P., Maurer, H.P., Longin, C.F.H. and Reif, J.C. (2014) Yield stability of hybrids versus lines in wheat, barley, and triticale. Theor Appl Genet 127, 309-316.

Muqaddasi, Q.H., Lohwasser, U., Nagel, M., Borner, A., Pillen, K. and Roder, M.S. (2016) Genome-Wide Association Mapping of Anther Extrusion in Hexaploid Spring Wheat. Plos One 11.

N'Diaye, A., Haile, J.K., Cory, A.T., Clarke, F.R., Clarke, J.M., Knox, R.E. and Pozniak, C.J. (2017) Single Marker and Haplotype-Based Association Analysis of Semolina and Pasta Colour in Elite Durum Wheat Breeding Lines Using a High-Density Consensus Map. Plos One 12.

Navara, S. and Smith, K.P. (2014) Using near-isogenic barley lines to validate deoxynivalenol (DON) QTL previously identified through association analysis. Theor Appl Genet 127, 633-645.

Pan, Y., Chen, J., Guo, H., Ou, J., Peng, Y., Liu, Q., Shen, Y., Shi, L., Liu, Y., Xiong, Z., et al. (2015) Association of genetic variants of GRIN2B with autism. Sci Rep 5, 8296.

Pinnaduwage, D. and Briollais, L. (2005) Comparison of genotype-and haplotype-based approaches for fine-mapping of alcohol dependence using COGA data. Bmc Genet 6.

Pook, T., Schlather, M., de los Campos, G., Mayer, M., Schoen, C.C. and Simianer, H. (2019) HaploBlocker: Creation of Subgroup-Specific Haplotype Blocks and Libraries. Genetics 212, 1045-1061.

Pryce, J.E., Bolormaa, S., Chamberlain, A.J., Bowman, P.J., Savin, K., Goddard, M.E. and Hayes, B.J. (2010) A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. J Dairy Sci 93, 3331-3345.

Ray, D.K., Mueller, N.D., West, P.C. and Foley, J.A. (2013) Yield Trends Are Insufficient to Double Global Crop Production by 2050. Plos One 8.

Sapkota, S., Hao, Y.F., Johnson, J., Buck, J., Aoun, M. and Mergoum, M. (2019) Genome-Wide Association Study of a Worldwide Collection of Wheat Genotypes Reveals Novel Quantitative Trait Loci for Leaf Rust Resistance. Plant Genome-Us 12.

Sato, S., Uemoto, Y., Kikuchi, T., Egawa, S., Kohira, K., Saito, T., Sakuma, H., Miyashita, S., Arata, S., Kojima, T., et al. (2016) SNP- and haplotype-based genome-wide association studies for growth, carcass, and meat quality traits in a Duroc multigenerational population. Bmc Genet 17.

Sato, Y., Yamamoto, E., Shimizu, K.K. and Nagano, A.J. (2019) Neighbor GWAS: incorporating neighbor genotypic identity into genome-wide association studies of field herbivory on Arabidopsis thaliana. BioRxiv, 845735.

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N. and Nelson, A. (2019) The global burden of pathogens and pests on major food crops. Nat Ecol Evol 3, 430-+.

Schneeberger, K. (2014) Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat Rev Genet 15, 662-676.

Segura, V., Vilhjalmsson, B.J., Platt, A., Korte, A., Seren, U., Long, Q. and Nordborg, M. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44, 825-830.

Semagn, K., Bjornstad, A. and Xu, Y.B. (2010) The genetic dissection of quantitative traits in crops. Electron J Biotechn 13.

Seren, U., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K. and Korte, A. (2017) AraPheno: a public database for Arabidopsis thaliana phenotypes. Nucleic Acids Res 45, D1054-D1059.

Shin, J. and Lee, C. (2015) Statistical power for identifying nucleotide markers associated with quantitative traits in genome-wide association analysis using a mixed model. Genomics 105, 1-4.

Slovak, R., Goschl, C., Seren, U. and Busch, W. (2015) Genome-wide association mapping in plants exemplified for root growth in Arabidopsis thaliana. Methods Mol Biol 1284, 343-357.

Song, B., Mott, R. and Gan, X. (2018) Recovery of novel association loci in Arabidopsis thaliana and Drosophila melanogaster through leveraging INDELs association and integrated burden test. Plos Genet 14, e1007699.

Sukumaran, S. and Yu, J. (2014) Association mapping of genetic resources: achievements and future perspectives. In: *Genomics of plant genetic resources* pp. 207-235. Springer.

Templeton, A.R., Boerwinkle, E. and Sing, C.F. (1987) A Cladistic-Analysis of Phenotypic Associations with Haplotypes Inferred from Restriction Endonuclease Mapping .1. Basic Theory and an Analysis of Alcohol-Dehydrogenase Activity in Drosophila. Genetics 117, 343-351.

Thind, A.K., Wicker, T., Simkova, H., Fossati, D., Moullet, O., Brabant, C., Vrana, J., Dolezel, J. and Krattinger, S.G. (2017) Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. Nat Biotechnol 35, 793-796.

Tregouet, D.A., Konig, I.R., Erdmann, J., Munteanu, A., Braund, P.S., Hall, A.S., Grosshennig, A., Linsel-Nitschke, P., Perret, C., DeSuremain, M., et al. (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nature Genetics 41, 283-285.

Tsai, H.Y., Janss, L.L., Andersen, J.R., Orabi, J., Jensen, J.D., Jahoor, A. and Jensen, J. (2020) Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. Sci Rep 10, 3347.

Voss-Fels, K.P., Stahl, A. and Hickey, L.T. (2019) Q&A: modern crop breeding for future food security. Bmc Biol 17.

Wall, J.D. and Pritchard, J.K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4, 587-597.

Wang, B., Lin, Z., Li, X., Zhao, Y., Zhao, B., Wu, G., Ma, X., Wang, H., Xie, Y., Li, Q., et al. (2020) Genome-wide selection and genetic improvement during modern maize breeding. Nat Genet.

Wang, M. and Lin, S.L. (2015) Detecting associations of rare variants with common diseases: collapsing or haplotyping? Brief Bioinform 16, 759-768.

Wang, N., Akey, J.M., Zhang, K., Chakraborty, R. and Jin, L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. Am J Hum Genet 71, 1227-1234.

Wang, X., Xu, Y., Hu, Z.L. and Xu, C.W. (2018) Genomic selection methods for crop improvement: Current status and prospects. Crop J 6, 330-340.

Weigel, D. and Mott, R. (2009) The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol 10.

Whitford, R., Fleury, D., Reif, J.C., Garcia, M., Okada, T., Korzun, V. and Langridge, P. (2013) Hybrid breeding in wheat: technologies to improve hybrid wheat seed production. Journal of experimental botany 64, 5411-5428.

Wurschum, T. (2012) Mapping QTL for agronomic traits in breeding populations. Theor Appl Genet 125, 201-210.

Xu, Y., Li, P.C., Yang, Z.F. and Xu, C.W. (2017) Genetic mapping of quantitative trait loci in crops. Crop J 5, 175-184.

Yang, Q., Li, Z., Li, W.Q., Ku, L.X., Wang, C., Ye, J.R., Li, K., Yang, N., Li, Y.P., Zhong, T., et al. (2013) CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. P Natl Acad Sci USA 110, 16969-16974.

Yang, W. and Liu, D. (2004) Advances in localization and molecular markers of wheat leaf rust resistance genes. Agricultural sciences in China 3, 770-779.

Yang, Y., Li, S.S., Chien, J.W., Andriesen, J. and Zhao, L.P. (2008) A systematic search for SNPs/haplotypes associated with disease phenotypes using a haplotype-based stepwise procedure. Bmc Genet 9.

Yoo, Y.J., Kim, S.A. and Bull, S.B. (2015) Clique-Based Clustering of Correlated SNPs in a Gene Can Improve Performance of Gene-Based Multi-Bin Linear Combination Test. Biomed Res Int 2015.

Yu, J.M., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38, 203-208.

Yu, J.M., Holland, J.B., McMullen, M.D. and Buckler, E.S. (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178, 539-551.

Zhang, Z., Guillaume, F., Sartelet, A., Charlier, C., Georges, M., Farnir, F. and Druet, T. (2012) Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. Bioinformatics 28, 2467-2473.

Zhao, H.H., Fernando, R.L. and Dekkers, J.C.M. (2007) Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. Genetics 175, 1975-1986.

Abbreviations

Abbreviation	Explanation
SNP	single-nucleotide polymorphisms
QTL	quantitative trait locus
MAS	marker-assisted selection
RIL	recombinant inbred line
MAGIC	multiparent advanced generation intercross
NAM	nested association mapping
GWAS	genome-wide association mapping
MLM	mixed linear model
WGS	whole genome sequencing
LD	linkage disequilibrium
MAF	minor allele frequency
Lr	leaf rust resistance genes
FH-based GWAS	functional haplotype-based GWAS
VLMCs	variable-length Markov chains
GC	genomic control
SA	structured association analysis
PCA	principal component analysis
FarmCPU	Fixed and random model Circulating Probability Unification
QTCAT	Quantitative Trait Cluster Association Test

Acknowledgements

It's my pleasure to be a member of Quantitative Genetics Group where I spend a very happy and memorable time. I appreciate all members in our group and all persons who helped me during my PhD study. Specifically, I want to thank my supervisor Prof. Dr. Jochen Reif who supported shape the PhD work and encourage me to do my best, always. I am really grateful to Dr. Yusheng Zhao and Dr. Yong Jiang for sharing their invaluable experience and suggestions. Without their patient explanation of academic question (i.e. the different mathematical equation or model), I could not finish my PhD study smoothly. I'd like to express my gratitude to Dr. Renate Schmidt and Dr. Albert Wilhelm Schulthess for helping me improve academic writing and discussion of biology question. I want to thank Dr. Sang He because he helped me adapt to life in Germany and taught me many other skills. I am also thankful for the helps from my colleagues Dr. Lars-Gernot Otto, Dr. Jianting Chu, Moritz Lell, Maximilian Rembe, Maria Yuli Gonzalez, Jie Zhang and Norman Phillip. They usually give me many good suggestions on both science and life, especially the suggestions on how to give a better presentation. In addition, many thanks to China Scholarship Council (CSC) for funding me to finish my PhD study and Leibniz Institute of Plant Genetics and Crop Plant Research (IPK Gatersleben) for providing the excellent platform.

My family members are always the most important persons for me. Dear parents, I love you both a lot and appreciate your effort and love in bringing me up to be a better individual. That is, I wouldn't be the person I am today and I may not even have been here if I am not with your support, encouragement, guidance and understanding. I also want to express my appreciation to my boyfriend because he is the one who shares all my happiness, depression and discouragement. Last but not least, I want to thank all my friends. Among them, I specially want to thank Dandan Wu, Jinping Cheng and Daiyan Li because we always share the delicious food and talk much, which makes my life more colorful. At the end of this, I appreciate all the persons that I even don't know but they helped me improve the manuscripts or PhD thesis, such as the editors and reviewers. Many thanks for all of your helps.

Curriculum vitae

Personal information

Name	Fang Liu
Gender	Female
Date of birth	November 23th, 1991
Place of birth	Shaodong, Hunan, China
Nationality	Chinese

Education

10.2016 ~	**PhD** Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany Department of Breeding Research Field: Quantitative Genetics Supervisor: Prof. Dr. Jochen C. Reif
09.2013 - 07.2016	**Master degree in Bioinformatics** National Key Laboratory of Crop Genetic Improvement College of Informatics, Huazhong Agricultural University Field: Bioinformatics and Metabolome Analysis Supervisor: Associate Prof. Weibo Xie
09.2010 - 07.2014	**Bachelor of science in Biotechnology** College of Life Science and Technology, Huazhong Agricultural University Wuhan, P.R. China
09.2006 - 07.2010	**High school** Shaodong innovative experimental school, Hunan, China
09.2003 - 07.2006	**Middle school** Huangdiling middle school, Hunan, China
09.1997 - 07.2003	**Primary school** Hanchong primary school, Hunan, China

Research techniques and skills

Computer skills and programming	Proficient in R, familiar with Shell, Perl and Python. Good command of Linux system commands.
Scientific drawing	Statistical graphics by R and Excel, familiar with Adobe Illustrator and Photoshop, and complex figure using R. Good sense of color scheme and graphics design.

List of scientific publications

Research articles of first author or co-first author

Liu, F., Jiang, Y., Zhao, Y., Schulthess, A.W. and Reif, J.C., (2020). Haplotype-based genome-wide association increases the predictability of leaf rust (*Puccinia triticina*) resistance in wheat. *Journal of Experimental Botany*. (doi.org/10.1093/jxb/eraa387)

Liu, F., Zhao, Y., Beier, S., Jiang, Y., Thorwarth, P., H. Longin, C. F., ... & Schulthess, A. W. (2020). Exome association analysis sheds light onto leaf rust (*Puccinia triticina*) resistance genes currently used in wheat breeding (*Triticum aestivum L.*). *Plant Biotechnology Journal*, *18*(6), 1396-1408. (dx.doi.org/10.1111/pbi.13303)

Liu, F., Schmidt, R. H., Reif, J. C., & Jiang, Y. (2019). Selecting Closely-Linked SNPs Based on Local Epistatic Effects for Haplotype Construction Improves Power of Association Mapping. *G3: Genes, Genomes, Genetics*, *9*(12), 4115-4126 (dx.doi.org/10.1534/g3.119.400451)

Research articles of co-author

Boeven, P. H. G., Zhao, Y., Thorwarth, P., **Liu, F**., Maurer, H. P., Gils, M., Schachschneider, R., Schacht, J., Ebmeyer, E., Kazman, E., Mirdita, V., Dörnte, J., Kontowski, S., Horbach, R., Cöster, H., et al. (2020). Negative dominance and dominance-by-dominance epistatic effects reduce grain-yield heterosis in wide crosses in wheat. *Science Advances*, *6*(24), eaay4897.

Pogoda, M., **Liu, F**., Douchkov, D., Djamei, A., Reif, J. C., Schweizer, P., & Schulthess, A. W. (2020). Identification of novel genetic factors underlying the host-pathogen interaction between barley (Hordeum vulgare L.) and powdery mildew (Blumeria graminis f. sp. hordei). PloS one, 15(7), e0235565.

Alomari, D.Z., Eggert, K., Von Wirén, N., Polley, A., Plieske, J., Ganal, M.W., **Liu, F**., Pillen, K. and Röder, M.S., (2019). Whole-

genome association mapping and genomic prediction for iron concentration in wheat grains. *International journal of molecular sciences*, *20*(1), p.76.

Xie, W. B., Wang, G. W., Yuan, M., Yao, W., Lyu, K., Zhao, H., Yang, M., Li, P. B., Zhang, X., Yuan, J., Wang, Q. X., **Liu, F**., Dong, H. X., Zhang, L. J., Li, X. L., et al. (2015). Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proceedings of the National Academy of Sciences of the United States of America, 112*(39), E5411-E5419. doi:10.1073/pnas.1515919112

Eidesstattliche Erklärung / *Declaration under Oath*

Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content.

Datum / Date

Unterschrift des Antragstellers / *Signature of the applicant*

Erklärung über bestehende Vorstrafen und anhängige Ermittlungsverfahren / *Declaration concerning Criminal Record and Pending Investigations*

Hiermit erkläre ich, dass ich weder vorbestraft bin noch dass gegen mich Ermittlungsverfahren anhängig sind.

I hereby declare that I have no criminal record and that no preliminary investigations are pending against me.

Datum / Date	Unterschrift des Antragstellers / *Signature of the applicant*