

# Patterns of genetic diversity and linkage disequilibrium in a highly structured *Hordeum vulgare* association-mapping population for the Mediterranean basin

Jordi Comadran · W. T. B. Thomas · F. Á. van Eeuwijk · S. Ceccarelli ·  
S. Grandó · A. M. Stanca · N. Pecchioni · T. Akar · A. Al-Yassin · A. Benbelkacem ·  
H. Ouabbou · J. Bort · I. Romagosa · C. A. Hackett · J. R. Russell

Received: 11 September 2008 / Accepted: 25 March 2009 / Published online: 5 May 2009  
© Springer-Verlag 2009

**Abstract** Population structure and genome-wide linkage disequilibrium (LD) were investigated in 192 *Hordeum vulgare* accessions providing a comprehensive coverage of past and present barley breeding in the Mediterranean basin, using 50 nuclear microsatellite and 1,130 DArT<sup>®</sup> markers. Both clustering and principal coordinate analyses clearly sub-divided the sample into five distinct groups centred on key ancestors and regions of origin of the germplasm. For given genetic distances, large variation in LD values was observed, ranging from closely linked

markers completely at equilibrium to marker pairs at 50 cM separation still showing significant LD. Mean LD values across the whole population sample decayed below  $r^2$  of 0.15 after 3.2 cM. By assaying 1,130 genome-wide DArT<sup>®</sup> markers, we demonstrated that, after accounting for population substructure, current genome coverage of 1 marker per 1.5 cM except for chromosome 4H with 1 marker per 3.62 cM is sufficient for whole genome association scans. We show, by identifying associations with powdery mildew that map in genomic regions known to have resistance loci, that associations can be detected in strongly stratified samples provided population structure is effectively controlled in the analysis. The population we describe is, therefore, shown to be a valuable resource, which can be used in basic and applied research in barley.

Communicated by A. Kilian.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-009-1027-0) contains supplementary material, which is available to authorized users.

J. Comadran (✉) · W. T. B. Thomas · J. R. Russell  
Genetics Programme, Scottish Crop Research Institute (SCRI),  
Invergowrie, Dundee DD2 5DA, Scotland, UK  
e-mail: Jordi.Comadran@scri.ac.uk; Comadran.Jordi@scri.ac.uk

F. Á. van Eeuwijk  
Biometrics Applied Statistics, Wageningen University,  
Wageningen, The Netherlands

S. Ceccarelli · S. Grandó  
International Center for Agricultural Research in the Dry Areas  
(ICARDA), Aleppo, Syria

A. M. Stanca  
CRA, Genomic Research Centre,  
29017 Fiorenzuola d'Arda (PC), Italy

N. Pecchioni  
Dipartimento di Scienze Agrarie, Università di Modena e Reggio  
Emilia, Reggio Emilia, Italy

T. Akar  
Central Research for Field Crops, Ankara, Turkey

A. Al-Yassin  
NCARTT, Amman, Jordan

A. Benbelkacem  
ITGC, Constantine, Algeria

H. Ouabbou  
INRA Morocco, CRRA, Settat, Morocco

J. Bort  
Department de Biologia Vegetal, Universitat de Barcelona,  
Barcelona, Spain

I. Romagosa  
Centre UdL-IRTA, Universitat de Lleida, Lleida, Spain

C. A. Hackett  
BioSS, SCRI, Invergowrie, Dundee, Scotland, UK

## Introduction

Barley (*Hordeum vulgare* L.) is the fourth main cereal crop worldwide with a long history of breeding and research. It was domesticated from its wild relative *Hordeum vulgare* subsp. *spontaneum* in the Fertile Crescent around 10,000 years ago (Badr et al. 2000). Distinct environmental conditions and breeding paths have shaped its diversity, accumulating a rich reservoir of gene alleles for adaptation and survival in harsh natural environments (Grando et al. 2001). This diversity resource has already contributed many useful alleles to the cultivated gene pool, especially those related to biotic stress tolerance, and will be indispensable in future breeding of crops adapted to climatic change. In spite of the recent advances in biotechnology, wild progenitors and landraces are still, and will remain for the next few decades, the most immediate and acceptable source of useful genes for conventional breeding programs. However, the extended gene pool of barley remains poorly characterised.

Over the last few years, candidate gene studies and high-throughput genotyping platforms have promoted association mapping as a viable alternative for quantitative trait locus (QTL) mapping (Waugh et al. 2009). This strategy differs from meiotic mapping with bi-parental crosses in three key ways. First, the use of large numbers of lines typically provides multiple alleles at each locus. Second, the genotyping of a representative collection of material provides a detailed description of the diversity within the collection and thus an identification of key introgressions if enough markers are used in the introgression areas. Third, QTL allele effects are assessed within different genetic backgrounds, including those which are representative of the present cultivated germplasm, providing more reliable estimates of these effects and higher efficiency in marker-assisted selection procedures.

Non-random association of alleles at different loci, known as linkage disequilibrium (LD), plays a central role in association analysis, because it determines the number and density of markers needed for whole genome scans, and the accuracy of the mapping. Several studies have now been conducted in different plant systems to investigate the extent and patterns of LD across crop plant genomes (Flint-Garcia et al. 2003; Gupta et al. 2005; Mackay and Powell 2007). LD values depend on the current levels and distribution of genetic diversity and the amount of effective recombination events in the ancestry of the lines under study. In consequence, LD decays faster with increasing genetic map distance in allogamous than in autogamous species, in wild relatives or landraces than in modern cultivated genotypes, and in outbred than in inbred genotypes.

In barley, comparison of gene-based studies (Russell et al. 1997; Lin et al. 2002; Morrell et al. 2003, 2005; Caldwell et al. 2006; Stracke et al. 2006) and genome-wide studies (Kraakman et al. 2004; Rostoks et al. 2006; Malysheva-Otto et al. 2006) gives some insights into the strong influence of selection and recent population history on the observed structure of LD. Values of LD are highly variable across the genome. In some cases for *Hordeum spontaneum* populations as well as in landrace collections, LD has low values in genes not subject to the strong directional selection forces of plant breeding (Morrell et al. 2005; Caldwell et al. 2006). In contrast, LD extends over large linkage distances (sometimes greater than 50 cM) in some regions in related elite germplasm (Kraakman et al. 2004; Rostoks et al. 2006). Thus, suggesting that ancestral barley populations, such as *H. spontaneum* populations as well as in old landrace collections, had low LD and the high LD observed in domesticated populations has been recently generated by the breeding process.

To exploit association-mapping strategies in plants and survey the diversity of adaptation, to abiotic stresses, such as drought in barley, we assembled a collection of 192 genotypes embracing much of diversity present in the cultivated gene pool throughout Europe and the Mediterranean basin region. This collection comprises landraces and historically important varieties, as well as modern varieties, together with representative genotypes from Central and Northwest Europe. Genetic variation has been screened in this collection with simple sequence repeat (SSR) and Diversity Array Technology (DArT<sup>®</sup>) molecular markers. SSRs are highly polymorphic single-locus DNA markers that have been shown to remain stable across generations (Russell et al. 1997), which means they are an excellent molecular marker system for population studies. DArT<sup>®</sup> markers (Wenzl et al. 2004) are a promising platform for whole genome profiling of barley as they provide repeatable high-throughput multi-locus dominant bi-allelic markers for medium density genome scans at low cost (Wenzl et al. 2004, 2006).

The detailed aims of this study were (1) to estimate genetic diversity within different barley germplasm groups, (2) to provide whole marker genome coverage of the samples for association-mapping studies, (3) to determine the patterns of population structure between and within the groups detected, (4) to investigate patterns of LD within the collection for further characterisation of germplasm clusters, in light of their ancestral history, and (5) to test the feasibility of a whole genome scan for complex traits with current levels of LD and genome coverage. Ultimately, the goal was to provide the barley research community with a genetically well-characterised population for association-mapping loci controlling adaptation to the drought-prone areas throughout the Mediterranean.

## Materials and methods

### Plant samples and DNA extraction

One hundred and ninety-two accessions of *H. vulgare*, selected to represent the past and present germplasm grown in various regions around the Mediterranean basin (Table 1), were multiplied at the ICARDA field site in Tel Hadya, Syria, for harvest year 2003 to produce sufficient seed for trialing in subsequent years. From this multiplication, a single plant of each genotype was grown in the glasshouse under standard growing conditions. DNA was extracted from seedling leaf material of each genotype using a Mixer Mill MM300 (Retsch) in combination with DNeasy plant extraction kits (Qiagen) according to manufacturer's instructions.

### Genotyping

Genotyping was conducted in two phases. In phase 1, 30 EST-SSR markers and 22 SSRs derived from genomic DNA selected to be reasonably evenly distributed amongst the seven barley linkage groups were used (supplementary Table 1). Marker fragments were generated by performing polymerase chain reactions (PCRs) containing 20 ng genomic DNA, 1× PCR buffer, 0.3 units *Taq* polymerase, 0.3 μM of forward and 0.3 μM 6-FAM fluorescently labelled (Applied Biosystems) reverse primer, and 200 μM dNTPs in a total volume of 10 μl. One microlitre of PCR product was denatured in 8.75 μl of Hi-Di formamide with 0.25 μl of GeneScan Rox500 size standard (Applied Biosystems). PCR products were separated on an ABI3700 capillary sequencer for allele detection and analysed using GENEMAPPER 3.0 software. In phase 2, further samples, containing between 0.5 and 1 μg of total DNA, from each accession were sent to Triticarte™ (<http://www.triticarte.com.au/>) for whole genome profiling with the standard barley DArT® array on a standard contract basis.

### Data analysis

For each of the markers in the set used in phase 1, diversity statistics including total number of alleles and mean number of alleles per locus were calculated using the Microsatellite Toolkit for Excel (Park 2001). Polymorphism information content (PIC) values were calculated in Microsoft Excel using the following formula:  $PIC = 1 - \sum P_i^2$ , where  $P_i$  is the frequency of the  $i$ th SSR allele (Smith et al. 2000).

Based on the data from phase 1 genotyping, population structure was examined using three different methodologies. First, phylogenetic analysis was conducted using an allele sharing distance matrix (Bowcock et al. 1994), computed by the Microsatellite Toolkit (Park 2001), as input for the

neighbour-joining (NJ) tree building clustering algorithm implemented in the PHYLIP package (Felsenstein 1997). The resulting dendrogram was rooted using Jordan landrace genotype entry 194 present in our sample (Table 1) as the out-group; based on its proximity to wild barley (*H. spontaneum*) genotypes from the Fertile Crescent, following genotyping with the same set of SSR markers (data not shown). Moreover, it is likely that most of cultivated barley originated from a single domestication event in the Fertile Crescent (Badr et al. 2000). Second, principal coordinate analysis, PCO, based on simple matching of fragment product sizes was performed with Genstat 9 (Payne et al. 2006). Third, a Bayesian clustering method was applied to identify clusters of genetically similar individuals using STRUCTURE software version 2.1 with admixture (Pritchard et al. 2000b; Pritchard and Donnelly 2001; Falush et al. 2003). STRUCTURE was run three times independently with  $k$  ranging from 1 to 10 in each run in order to verify that the estimates were consistent across runs. To decide on the appropriate number of clusters ( $k$ ), the estimated normal logarithm of the probability of fit, provided in the STRUCTURE output, was plotted against  $k$ . This value reaches a plateau when the minimal number of groups that best describe the population substructure has been reached (Pritchard et al. 2000a) (supplementary Figure 1).

### Linkage disequilibrium analysis

An interim consensus map of barley DArT® markers (A. Kilian, Triticarte, personal communication) was used to study LD amongst the 192 accessions. This map differed slightly from the published DArT® consensus map (Wenzl et al. 2006) in that four markers were not included and the map distances were slightly longer. LD between mapped DArT® loci was calculated by the squared allele frequency correlation coefficient ( $r^2$ ) (Weir 1979) implemented in TASSEL version 1.9.4 (Buckler et al. 2006). In this way,  $P$  values were estimated from all pairs of DArT® bi-allelic markers within the same chromosome after removal of loci with rare alleles ( $f < 0.10$ ). The extent and distribution of LD were visualised by plotting intra-chromosomal  $r^2$  values (significance  $P < 0.001$ ) against the genetic distance in centiMorgans. Unlinked  $r^2$  estimates were square root transformed to approximate a normally distributed random variable and the parametric 95th percentile of that distribution was taken as a critical value of  $r^2$ , beyond which LD is likely to be caused by genetic linkage (Brescaglio and Sorrells 2006).

### Phenotypic and association analysis

This association-mapping population has been phenotyped in 28 trials around the Mediterranean basin for a series of agronomic traits (Comadran et al. 2008). The purpose of this

**Table 1** Genotypes and geographic origin of germplasm sampled and genotyped

Geographical origin		Genotype
A priori groupings	Cultivated region	
Landraces		
East Mediterranean	Jordan	Entry191, entry192, entry193, entry194, Shoubak, Al Mansoura, Mafraq, Jordan1., Kafar Rakeb, Jordan2, Al Rama, That Rass, Al Azraq Al Janoubi, Modeibeia, Bir Al Dabaghat, Al Hisha, Twal Al Janoubi, Al Gdieda village
North Mediterranean	Dalmatian	Ragusa
	France	Hatif de Grignon
	Italy	Agello, Amiterno, Locale Aquila, Locale Leccese, Locale Siciliano, Orazio
	Spain	Albacete, Almunia, Berta, Macotera (SA), Pane, Pamula de Artes (L.), Precoz de Cadreita (NA), 206-33, 206-62, 206-94, 206-44, 206-24, 206-78, 206-99, 206-16
Other	Czech	Hanna, Kneifel, Vollkom
	Denmark	Kenia, Binder
	Germany	Tschemmaks 2-row, Heils Franken, Haisa
	Sweden	Gull
	UK	Pioneer
South Mediterranean	Algeria	Saida, Tichedrett, Algerian Landrace
	Morocco	Sountate 1/1, Sountate 1/31, Sountate 1/64, Sountate 1/61, Sountate 1/167, Taghighachte 8/1, Taghighachte 8/2, Taghighachte 8/9, Taghighachte 8/33, Taghighachte 8/49, Motzli 9/4, Motzli 9/28, Motzli 9/39, Motzli 9/46, Motzli 9/55, Taounate 10/2, Taounate 10/8, Taounate 10/9
Turkey	Turkey	Amb-192, Amb-196, Amb-307, Amb-19, Amb-199, Amb-135, Amb-134, Amb-133, Turkey
Modern cultivars		
East Mediterranean	Jordan	ASCAD 176, Rum
North Mediterranean	France	Manitou, Siberia, Hispanic, Reincte, Tremois
	Italy	Aliseo, Amillis, Dasio, Federal, Gotic, Grecale, Grosso, Kelibia, Mattina, Nure, Naturel, Solen, Sonora, Tea, Tidone, Ultra, Vertige
Other	Spain	Candela, Dobla, Kika, Orria, Zaida
	Germany	Alexis, Igri, Scarlett, Triumph, Barke, Otis, Magda, Regina
	Holland	Atem, Intro, Aramir, Apex
	UK	Chartot, Fanfare, Optic, Graphic, Kym, Nevada, Tipper
	USA	Steptoe
South Mediterranean	Algeria	Alanda-01, Manel
	Morocco	Arig8, Aglou, Amalou, Merzaga-07, Massine, Rabat-01, Oussama
Turkey	Turkey	Efes-98, Anadolu-98, Aydanhanim, Sahin-91, Karatay, Tarm-92, Orza-96, Bulbul-89, Yesevi-93
Old cultivars		
East Mediterranean	Jordan	Entry190
North Mediterranean	France	Plaisant, Ager, Alpha, Astrix, Aurore, Barberousse, Beka, Logra, Monlon, Baraka, Elan
	Italy	Arda, Onice, Vetulio

Table 1 continued

Geographical origin	Genotype	
	Cultivated region	
A priori groupings		
Other	Germany Holland Sweden UK Algeria Turkey	Vogelsanger Gold, Dea, Volla, Union Vada, Emir, Zephyr, Hassan Balder, Ingrid, Pallas, Formula Proctor, Georgie, Maris Otter Martin, Harmal, Assala-04, ER/Apm Tokak, Yesilkoy, Hamidiye, Cumhütiyet, Efes-3, Obruk, Zafer-160, Anadolur-86
South Mediterranean		
Turkey		

paper is to genetically characterise the mapping population and to test the feasibility of a whole genome scan for complex traits with current levels of LD and genome coverage. Barley powdery mildew loci are well described in the literature (Chelkowski et al. 2003) and can be used to assess the suitability of the population for association mapping. Plant lines were phenotyped for powdery mildew in 2004 at trial sites in Italy and Morocco, named ITA\_4D, ITA\_4W, MOR\_4D and MOR\_4W. Two experiments per site were conducted, using a design in which a quarter of the entries were replicated once. Twenty percent of the plots consisted of four checks, which were included repeatedly in a systematic diagonal fashion. Powdery mildew infection scores were recorded for each line at each site on an ordinal scale running from 0, absence of infection, to 10, maximum infection.

Mildew scores were analysed per trial by mixed model analysis using Restricted Maximum Likelihood (REML) in Genstat 9 (Payne et al. 2006), to produce entry means (Best Linear Unbiased Predictors, or BLUPs), taking into account row and column variation (Piepho et al. 2008). TASSEL 2.0 (Yu et al. 2006) was used to derive a relative kinship matrix (K) from the set of DArT<sup>®</sup> marker data. Structure 2.2 (Pritchard et al. 2000a) was used to obtain a population structure probabilities matrix for  $k = 5$  (Q). Marker-trait associations were detected across trials by mixed model analysis with individual DArT<sup>®</sup> markers as fixed effects in Genstat 9. K or Q matrices accounting for the population stratification effects were fitted in the random term of the model to reduce type I errors. *P* value significance thresholds for declaring the presence of positive marker-trait associations were calculated for each individual trial based on a false discovery rate, FDR, which is defined as the expected proportion of true null hypotheses that are rejected (Benjamini and Hochberg 1995). This is implemented as follows: let  $P_{(i)}$  ( $i = 1, \dots, m$ ) be the ordered significances of the linear model relating a phenotypic trait to the marker  $i$ . If  $k$  is the largest value of  $i$  such that

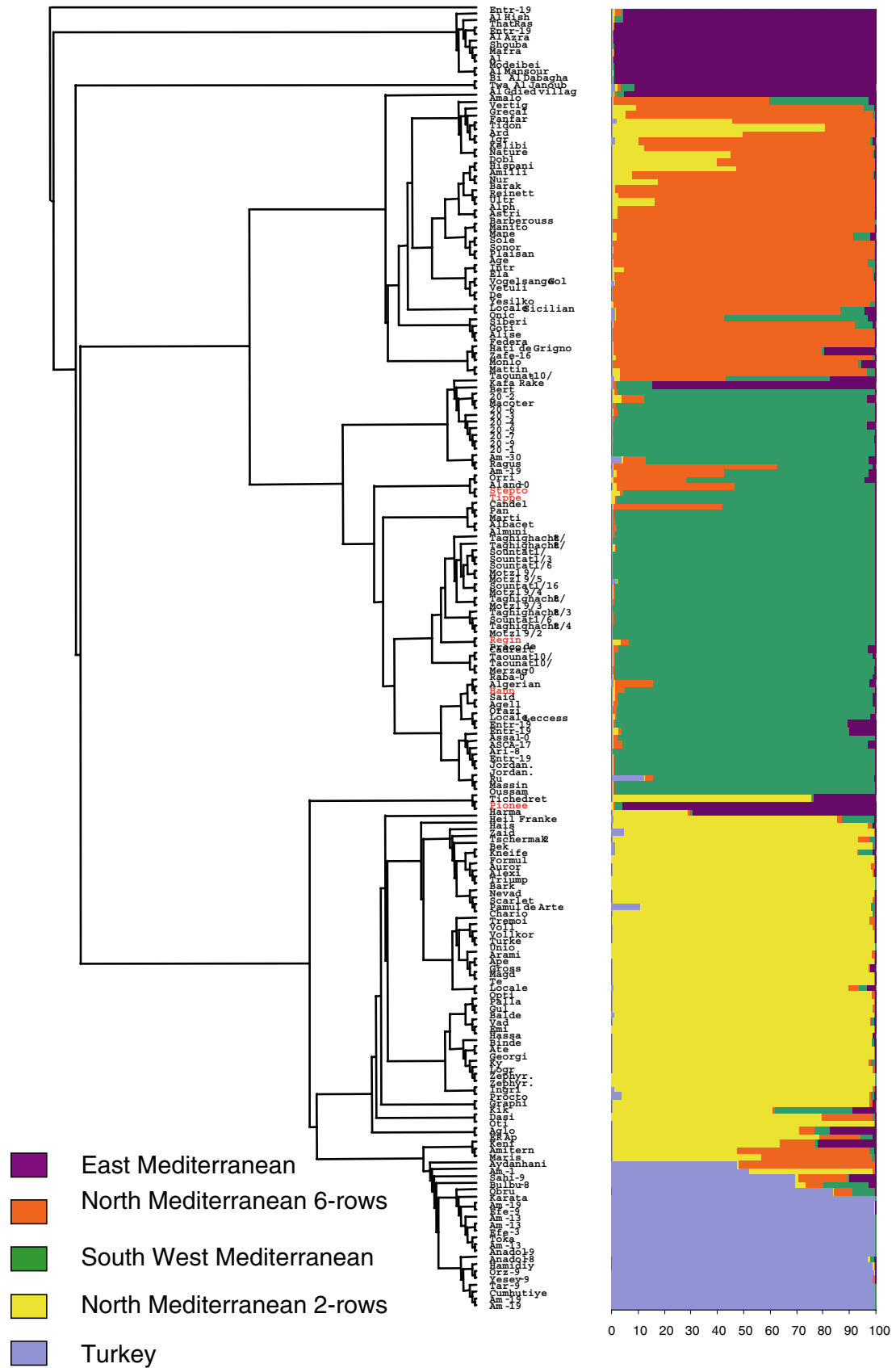
$$P_{(i)} \leq \frac{i\alpha}{m}$$

then the  $k$  values  $P_{(1)}, \dots, P_{(k)}$  are declared significant. This controls the FDR at a level  $\leq \alpha$ . Significance thresholds for a FDR at a level of  $\alpha = 0.05$  as  $-\log_{10}(P \text{ value})$  were set to 4.39, 3.21, 2.45 and 2.87 for ITA\_4D, ITA\_4W, MOR\_4D and MOR\_4W, respectively.

## Results

### Genetic diversity

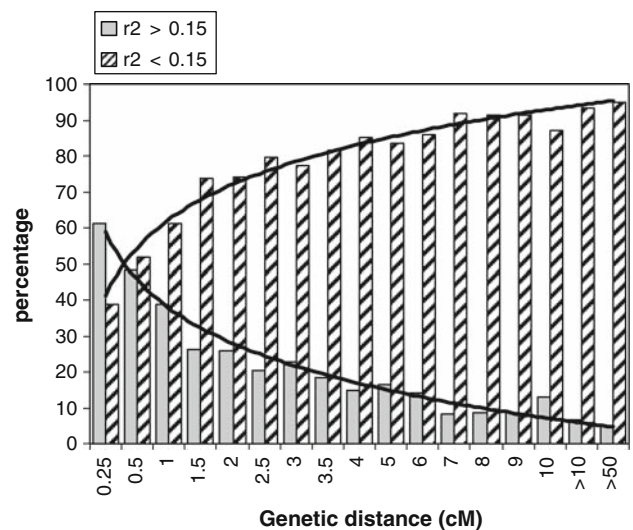
Fifty-two SSR markers used in phase 1 genotyping were able to distinguish all accessions. Diversity statistics for



◀ **Fig. 1** Dendrogram (neighbour joining) and inferred population structure based on 192 individuals and 50 EST and genomic SSR markers using STRUCTURE. Each individual is represented by a *line* partitioned in five coloured segments that represent the individual's estimated membership fractions to each one of the five clusters. Names in *red* refer to aberrant genotypes due to selection of contaminants at some stage in the selection and multiplication process

each phase 1 marker locus are summarised in supplementary Table 1. 474 alleles were detected over the whole sample of 192 accessions with an average of 9.11 alleles per locus and an overall mean diversity value of 0.56, ranging from 0 to 0.909. Comparisons between landraces and cultivars (modern and old pooled together) showed a general decrease in diversity from the former to the latter (mean values of 7.79–5.57 alleles per locus for 84 and 108 accessions for landraces and cultivars, respectively). 124, 11 and 40 rare alleles (frequency < 5%) were specific to landraces, old cultivars and modern cultivars, respectively, indicating a loss of rare alleles due to breeding. In addition, two frequent alleles (frequency > 5%) from SSRs HVM62 and Bmag0353 were specific to landraces. Two EST-SSR markers proved monomorphic across all 192 accessions and were excluded from further analyses. Moreover, barley is predominantly autogamous and as expected no heterozygous plants were observed for the 50 SSR markers assayed. Levels of heterozygosity could not be estimated from DArT<sup>®</sup> markers as they are predominantly dominant.

DArT<sup>®</sup> analysis produced 1,130 bi-allelic markers with corresponding PIC values ranging from 0.126 to 0.500. For a bi-allelic marker, the minimum and maximum PIC values are 0 and 0.5, respectively. Most DArT<sup>®</sup> markers had PIC values greater than 0.450, with an average of 0.407. Mean DArT<sup>®</sup> diversity values for landraces, old cultivars and modern cultivars were 0.390, 0.386 and 0.386, respectively. Only 15 markers had minimum allele frequencies (MAF) lower than 0.05. From the interim consensus map, 811 of the 1,130 DArT<sup>®</sup> markers had a known map location and 105 of the 319 unmapped markers had  $r^2$  LD values higher than 0.95 with mapped markers and could be given a fairly accurate map location. Whilst some DArT<sup>®</sup> markers mapped to the same location, very few produced identical genotypes for the germplasm in our association-mapping panel and thus we consider that redundancy was not a major issue in our data set. DArT<sup>®</sup> markers provided an average genome coverage of 116 markers per chromosome (1 marker per 1.5 cM) with a minimum of 52 markers for chromosome 4H (1 marker per 3.62 cM) and a maximum of 164 for chromosome 3H (1 marker per 1.36 cM). The more sparse coverage of chromosome 4H can also be seen in the published DArT<sup>®</sup> consensus map (Wenzl et al. 2006) and, in our case, resulted in four apparently monomorphic regions of 20 or more cM. It is important to consider that although estimates of marker/cM are given



**Fig. 2** Proportion of marker pairwise  $r^2$  measurements above and below background linkage disequilibrium of  $r^2$  of 0.15 plotted as a function of genetic distance

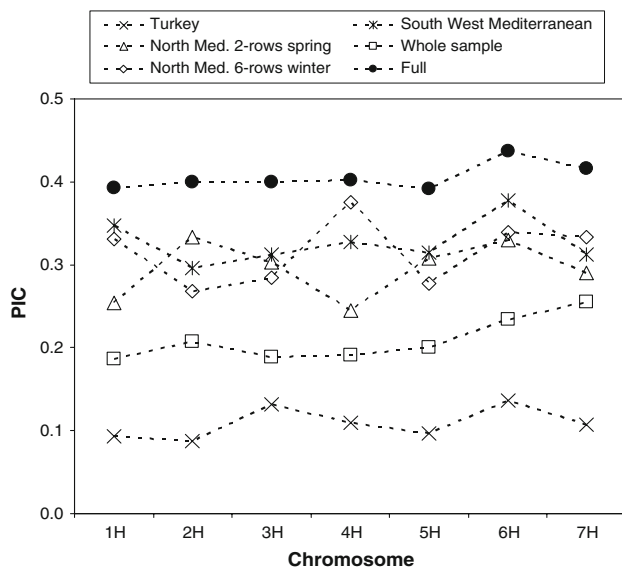
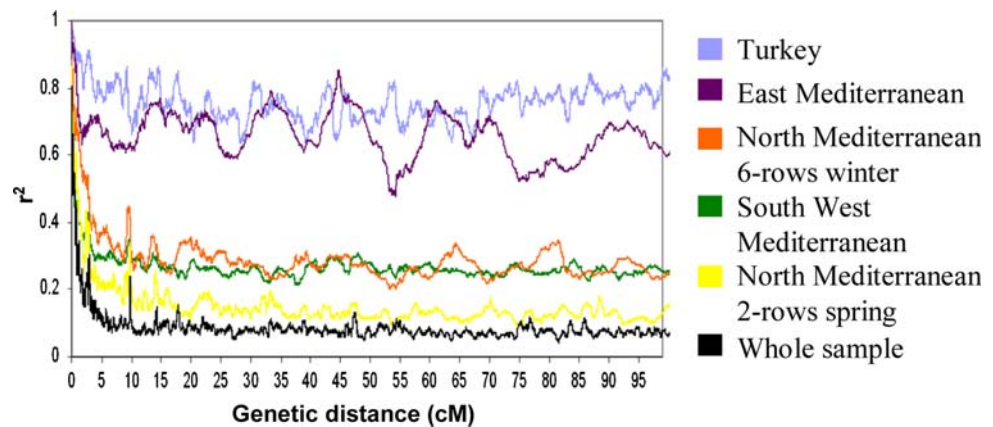
the picture is more complex as in the barley genome we have a very heterogeneous distribution of recombination rates and gene-rich regions amongst the individual barley chromosomes (Kunzel et al. 2000).

#### Population stratification and admixture

Principal coordinate analysis of the polymorphic SSR data showed very similar clustering patterns to NJ with five distinguishable groups (Fig. 1; supplementary Figure 4).

STRUCTURE showed an optimum cluster number ( $k$ ) of 5 (supplementary Figure 1). Growth habit was assessed in a trial experiment at SCRI in 2005 (data not shown). The composition of the clusters at  $k = 5$  splits the sample into classic barley categories relative to growth habit and spike morphology and made geographical and genetical sense so this grouping was utilised for the remainder of the study. The five groups comprised: (a) 61 South West Mediterranean 6-row accessions, (b) 21 Turkish accessions, (c) 16 East Mediterranean accessions, (d) 42 North Mediterranean 6-row winter accessions and (e) 52 North Mediterranean 2-row spring accessions. Introgression episodes (Fig. 1) due to crossing North Mediterranean 6-row winters with North Mediterranean 2-row spring or South West Mediterranean accessions can be clearly observed. For instance, varieties that share ancestors from both groups involved in an introgression event can be identified from the estimated membership probabilities produced by STRUCTURE and we find an observable gradient from North Mediterranean 6-row winter accessions, i.e. those whose ancestors are exclusively drawn from the North Mediterranean 6-row winter gene pool to the South West Mediterranean or 2-row

**Fig. 3** Plots of  $r^2$  measurements as a function of genetic distance (in cM) between informative ( $f > 0.1$ ) polymorphic DArT<sup>®</sup> loci in six sample sets. Intra-chromosomal pairwise DArT<sup>®</sup> markers were sorted as a function of genetic distance and  $r^2$  measurements averaged across a sliding window of 20 adjacent pairwise DArT<sup>®</sup>



**Fig. 4** Genomic diversity: average PIC values of 811 mapped DArT<sup>®</sup> loci in barley germplasm subsets

spring gene pools (Fig. 1). The South West Mediterranean group is more diverse in origin and it is sub-divided when  $k = 6$  (data not shown). Most accessions within this group are 6-row spring types, except for Rum and Massine which are 2-row spring types. Three old Italian varieties, Orazio, Loccale Leccese and Agello, are clustered with South West Mediterranean varieties, but, according to their pedigrees, they originated from selections of a heterogeneous African landrace. In the current study, the group of Turkish accessions appears like a sub-cluster within North Mediterranean 2-row spring accessions with strong genetic identity. For instance, all Turkish accessions are 2-row facultative types except for the 6-row Sahin91.

#### Linkage disequilibrium

Comparing LD between all pairs of DArT<sup>®</sup> markers mapped on the consensus map, the general trend is a rapid

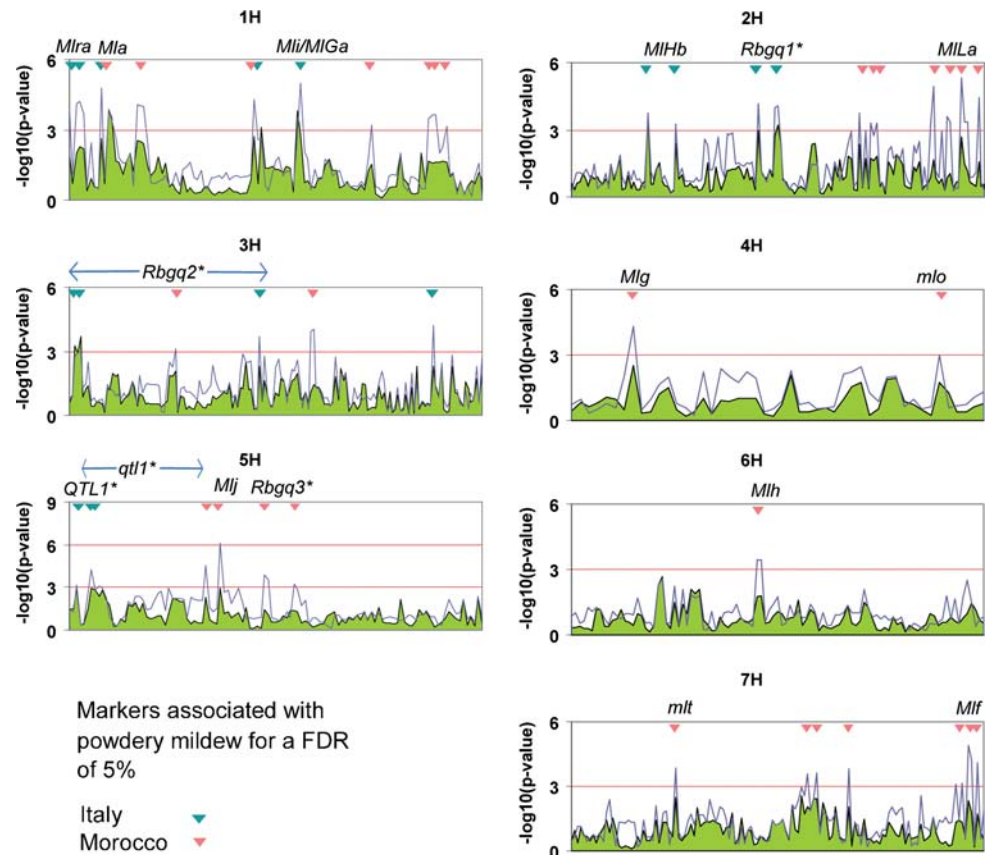
decay with genetic distance (Fig. 2). A critical value of  $r^2$ , beyond which LD is likely to be caused by genetic linkage in the absence of population structure, was fixed at 0.15 (Breseghello and Sorrells 2006). The proportion of markers with  $r^2 > 0.15$  decreased from 60.98 to 8.08% within the first seven cM, although there are still a large number of closely linked markers that show low levels of LD. Conversely, there are also a few contrasting examples where high  $r^2$  values were observed between theoretically unlinked markers ( $>50$  cM) (supplementary Figure 2).

Whilst mean LD across the whole population decayed below the critical  $r^2$  value of 0.15 within 3.2 cM (Fig. 3), three peaks of higher LD at 5.6, 9.9 and 14.3 cM distance were also observed. Further exploration of these spots of high LD revealed that the peak at 9.9 cM was biased by a few markers in strong LD on chromosome 3H and the other two peaks by a few other markers in strong LD on chromosome 4H (data not shown). Considering the distinct genetic clusters, basal mean  $r^2$  values are clearly distinct depending on the group under study. High levels of LD extend across the entire genome in the East Mediterranean and Turkish germplasm groups (Fig. 3). For these two groups, the small sample size and high levels of similarity between accessions are assured to be the main reasons for such high levels of LD. For South West Mediterranean and North Mediterranean 6-row winter groups, basal mean  $r^2$  values were double those of the North Mediterranean 2-row spring group which may reflect the existence of unaccounted population substructure within North Mediterranean 6-row winters and South West Mediterranean 6-row groups as a source of significant LD. We have already observed that both groups are not completely homogeneous in terms of growth habit and Spanish landraces may have two genetic sources for the 6-row spike phenotype (Casas et al. 2005; Yahiaoui et al. 2008).

Comparison of LD (Fig. 3) and diversity values between the STRUCTURE groupings (Fig. 4) offers the opportunity to further explore the population structure of the sample. North Mediterranean 6-row winter and South West



**Fig. 5** Whole genome scans for powdery mildew. Significances of the calculated  $P$  values are shown as  $-\log_{10}(P$  value). *Green* corresponds to  $-\log_{10}(P$  value) scores using a relative kinship matrix (K) and the *blue line* corresponds to  $-\log_{10}(P$  value) scores using STRUCTURE probabilities matrix (Q) for a number of populations of  $k = 5$ . Analysis was performed for each trial independently. Results from the four trials are plotted together showing only maximum  $-\log_{10}(P$  value) for each marker (see supplementary Figure 5 for independent trials plots). See approximate positions of known resistance genes on top of the graph. Resistance genes are described in Chelkowski et al. (2003)). Asterisk (\*) refers to QTLs described in the literature without an associated candidate gene: *Rbgqs* (Shtaya et al. 2006), *qtl* (Thomas et al. 1995) and QTL (Heun 1992)



Mediterranean accessions show levels of diversity similar to North Mediterranean 2-row spring accessions but basal LD values are lower in the latter which indicates that the diversity of the first two groups may be stratified. In fact, the North Mediterranean 6-row winter group, where high levels of admixture are observed, is not completely homogeneous in row number, and the group encompassing South West Mediterranean accessions is heterogeneous in growth habit with some accessions being winter and the rest spring types.

#### Powdery mildew whole genome scans

Correlations of powdery mildew scores with the first PCOs were performed to assess the stratification of powdery mildew within our mapping population. PCO analysis was based on simple matching using 1,130 DArT<sup>®</sup> markers and STRUCTURE groupings proved well represented by the first PCOs (supplementary Figure 4). Powdery mildew scores were significantly correlated with the first five PCOs (supplementary Table 3). Given the strong stratification of mildew resistance within the population, we found that 45.2% of the DArT<sup>®</sup> markers was associated with powdery mildew at  $P < 0.001$  when correction for population structure was not taken into account (supplementary Figure 5). A mixed model framework (Yu et al. 2006) was

also used so that population structure effects could be modelled with a K or Q matrix to reduce the number of spurious associations. Association analyses were performed for each trial independently and a  $-\log_{10}$  transformation was applied to the significance of each DArT<sup>®</sup> marker and the transformed data were plotted together for the purpose of presentation of the data (Fig. 5; see supplementary Figure 5 for results of genome scans for each individual trial).

Fitting K was a more stringent procedure than fitting Q, with 0.9 and 6.5% highly significant marker-trait associations ( $P < 0.001$ ), respectively. For the four trials tested, the K model always gave the best approximation to the expected cumulative distribution of  $P$  values, followed by the Q model and the naïve approach (model without taking population structure into account) (supplementary Figure 6).

When corrected for structure, most significant associations could be related to regions where candidate genes have been reported in the literature (supplementary Table 2). Focusing on the most conservative approach which involved fitting a K matrix, the most significant associations were found with markers located on the short arm of chromosome 1H, which holds a cluster of resistance loci around the *Mla* locus, at the beginning of chromosome 2H coincident with the region where *MIHb* had been

reported and 5H, close to the *Mlj* locus. We also detected three positive associations in genomic regions where QTLs for mildew resistance have been reported, but no major resistance gene has been reported in the literature: 2H, 81.97 cM; 3H, 12.7–13.36 cM; 5H, 33.2 cM (Fig. 5).

## Discussion

By genotyping collections of cultivars with a large number of markers, association studies provide means to improve the genetic characterisation of germplasm. Markers associated with genomic regions controlling traits of interest improve the understanding of the genetic value of the individual genotypes within the germplasm collections. Moreover, as similar studies are completed, the genetic characterisation of many different germplasm sets will provide researchers with a comprehensive global perspective of the barley gene pool.

The current study is amongst the first to deliberately assemble and analyse a specific population representing the cultivated barley germplasm of the Mediterranean basin in order to provide a platform for association mapping in barley. The level of polymorphism observed using bi-allelic markers exceeds those of previous studies, with 1,130 polymorphic DArT<sup>®</sup> markers. Mean  $r^2$  LD values higher than 0.15 extend up to 3.12 cM in our sample, which suggests there is sufficient marker coverage for attempting a whole genome association scan in barley. In fact, previous studies in barley (Kraakman et al. 2004; Rostoks et al. 2006) have reported successful association mapping in barley with average densities of 33.7 and 93.7 markers per chromosome significantly lesser than our average density of 116 (1 marker per 1.5 cM). Moreover, although results are not comparable due to the distinct nature of the germplasm used, similar work with 318 Mediterranean basin *H. spontaneum* accessions demonstrated successful association mapping of rust resistance using 818 mapped DArT<sup>®</sup> markers from 1,090 polymorphic DArT<sup>®</sup> markers (Steffenson et al. 2007).

Kraakman et al. (2004), although detecting two subgroups, concluded that strong population structure was not present in the set of modern European 2-row spring barley cultivars under study. Rostoks et al. (2006) identified winter versus spring growth habit as the primary determinant of population stratification in a sample of elite northwest European barley. In contrast, the germplasm in the current study is considerably more diverse, both genetically and physiologically. This has resulted in having a strong population structure, based on an individual's pedigree and/or origin, which reflects different environmental and selection effects shaping barley diversity. The strong population stratification in Mediterranean cultivated barley, which has

also been reported by Malysheva-Otto et al. (2006) and Yahiaoui et al. (2008), emphasises the extensive use of a few key ancestors well adapted to specific target regions. The old Swedish landrace Gull features strongly in the pedigrees of North Mediterranean 2-row spring accessions, as older accessions developed from Heils Franken, Binder and Kenia (Russell et al. 2000). Similarly, Groninger and Mammot are key ancestral genotypes of North West European 6-row barley, whereas Carstens 2-row and Tschermaks 2-row are key 2-row progenitors (Fischbeck 2002). Such key ancestors are rarely found within Western Mediterranean 6-row accessions, which probably reflect a long history of genetic isolation. A clear genetic divergence between Spanish and European 6-row accessions has already been reported (Yahiaoui et al. 2008) and the observed separate clustering of Spanish and Moroccan 6-row accessions, with occasional cases of admixture, confirms the Western Mediterranean region as a valuable source of diversity for barley breeders working within this gene pool. The Turkish accessions that were sampled in the current study strongly cluster together because most were developed from Tokak. Kilian et al. (2006), analysing haplotype structure at seven barley genes, also reported a strong clustering of the 2-row spring Turkish cultivars. A strong Turkish identity has also been observed in wheat accessions (Kim and Ward 2000), and may reflect singularity of Turkish breeding program strategy and selection for cold hardiness which would have been sufficient to give distinctness from other 2-row barleys.

Previous studies have highlighted contrasting patterns of LD between genes a priori closely linked to a polymorphism of interest, but subjected to distinct selection pressures and recombination histories in the population under study (Caldwell et al. 2006). The study of wild barley by Morrell et al. (2005) demonstrates that intralocus LD decays at rates similar to maize (Remington et al. 2001); however, we are not looking at wild barleys and in a previous study, Caldwell et al. (2006) showed that although a gradual decay of LD was observed with distance, persistence of LD was extensive in elite cultivars compared to landraces. Our results suggest that in barley, whilst LD might decay rapidly in some regions of the genome, the selfing nature of the crop results in the persistence of LD up to 3.2 cM (see Fig. 3). Presumably, regions of interest have already been under stronger natural and breeding selection pressures and will be in higher LD thus facilitating association studies with a medium marker density. For instance, we have observed disparate and contrasting LD values but these may just reflect different recent history of selection and drift in different genomic regions or may be caused by the population substructure (supplementary Figures 2, 3).

In highly structured samples such as the one in the current study, accounting for population structure in

association-mapping studies is essential to minimise the rate of false positive associations. Previously, Rostoks et al. (2006) had demonstrated that a medium density of markers could successfully be used in barley to correctly locate a known character. However, in the current study, we have introduced more complexity by conducting whole genome scans using data from powdery mildew scores gathered from natural infection in a set of four field trials. Powdery mildew resistance genes are well described in the literature and given the strong history of breeding for resistance to this character in certain regions of Europe. We expected powdery mildew resistance to exhibit a strong stratification within the germplasm in the current study. Thus, it is a reasonable choice to investigate the effects of population structure and validate further positive associations. Up to 23 loci were described for barley powdery mildew resistance genes of known map positions (Chelkowski et al. 2003) and we were able to detect resistance effects associated at several.

Known powdery mildew resistance loci have been widely used in breeding programs resulting in a strong stratification of them within the germplasm groups. Accounting for population structure effects by the inclusion of a relative kinship matrix (K) presumably removes most of the variance associated with close relatedness, which suggest that the reduced list of significant associations reported possibly corresponds to those loci or alleles which have not yet been fully exploited in breeding programs or have matching virulences in the pathogen population. The associations detected by including a K matrix match known QTLs where their candidate genes have not yet been described; the multi-allelic nature of *Mla* locus; a locus where a resistance gene derived from *Hordeum bulbosum*, *MiHb* has been reported but where alternative hitherto unreported alleles may be present in cultivated barley; and one association without earlier reports in the literature in 1H around 95 cM. Coincidence of known loci positions with those of reported associations is a positive sign, but caution is the message and several issues have to be taken into consideration (Zhao et al. 2007). Zhao et al. (2007) focused on the difficulties, and the association studies may encounter most of them because of the loss of power due to the use of small sample size and problems arising of mapping approaches involving highly stratified traits in germplasm samples with a strong population structure. In an association-mapping study with a highly structured sample of *Arabidopsis*, Zhao et al. (2007) demonstrated that available methods to account for population structure effects in the analysis may not correct for all the false positives, and that correcting for population structure also introduces false negatives if the QTLs are perfectly correlated with population structure. In barley, although distinct population sub-clusters associate to

distinct geographic regions, breeding practice contributed to gene flow between them. Thus, population structure in barley is strong but not as extreme as in *Arabidopsis* where the samples are completely isolated from each other.

The experimental results of our study allow us to draw some conclusions about the potential of this population for association mapping. First, current genome coverage together with observed LD decay values is sufficient for detecting marker associations in the case of a complex trait provided that population structure is effectively and appropriately controlled within the association analysis. Second, as expected, the number of significant marker-trait associations decreases dramatically when the effects of population structure are taken into account. However, its inclusion as a co-variable in the statistical model seemed to reduce power to detect real effects ( $-\log_{10}(P \text{ value})$  scores were reduced to sixfold). The use of a kinship matrix in the analysis is a more conservative approach than using STRUCTURE estimates as it controls population structure better. Larger population size would be desirable to increase the power of association mapping for highly stratified traits, as the significant correlation of the phenotype with the population structure (supplementary Table 3) will make it difficult to detect alleles of smaller effect using the association-mapping approach (Zhao et al. 2007). This work is part of a series consisting 28 trials across the Mediterranean basin with extensive agronomic and environmental data collected in 2004 and 2005. The analysis of genotype per environment interactions pooling the majority of the environments together will certainly increase power. Third, the mixed model involving a Q matrix identified significant associations with markers mapping close to most of earlier reported powdery mildew QTL or major gene loci (supplementary Table 2), which highlights how broad is the genetic base of the population under study, as a highly diverse and valuable resource for basic and applied research in barley.

In conclusion, this work provides a detailed description of an association-mapping population that represents a historical survey of barley diversity in the Mediterranean rim and encompasses a large proportion of the genetic variation underlying the different mechanisms and strategies for adaptation to drought-prone environments around the region. We have demonstrated that the assembled germplasm set is genetically and physiologically diverse, and strongly structured due to pedigree relationships providing contrasting levels of population stratification. Strongly stratified traits are considered one of the major obstacles of association mapping because extensive LD of those loci across the genome results in a large amount of spurious associations. Although it is important to realise that there are limits in the association-mapping approach, we show that associations related to known candidate/QTL

regions can be detected if the different levels of genetic relatedness are effectively controlled. Future genotyping of this population with sequence-based markers (Rostoks et al. 2005) would increase the genome coverage and provide a rationale for going from any positive associations to the identification of a small set of candidate genes for a given trait.

**Acknowledgments** The above work was funded by the European Union-INCO-MED program (ICA3-CT2002-10026). SCRI received grant in aid from the Scottish Government Rural and Environment Research and Analysis Department.

## References

- Badr A, Muller K, Schafer-Pregl R, El Rabey H, Effgen S et al (2000) On the origin and domestication history of Barley (*Hordeum vulgare*). *Mol Biol Evol* 17:499–510
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* 57:289–300
- Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR et al (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368:455–457
- Breseghele F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172:1165–1177
- Buckler ES, Gaut BS, McMullen MD (2006) Molecular and functional diversity of maize. *Curr Opin Plant Biol* 9:172–176
- Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* 172:557–567
- Casas AM, Yahiaoui S, Ciudad F, Igartua E (2005) Distribution of MWG699 polymorphism in Spanish European barleys. *Genome* 48:41–45
- Chelkowski J, Tyrka M, Sobkiewicz A (2003) Resistance genes in barley (*Hordeum vulgare* L.) and their identification with molecular markers. *J Appl Genet* 44:291–309
- Comadran J, Russell J, Eeuwijk FA, Ceccarelli S, Grando S et al (2008) Mapping adaptation of barley to droughted environments. *Euphytica* 161:35–45
- Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587
- Felsenstein J (1997) An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst Biol* 46:101–111
- Fischbeck G (2002) Contribution of barley to agriculture: a brief overview. In: Slafer GA, Molina-Cano JS, Savin R, Araus JL, Romagosa I (eds) *Barley science: recent advances from molecular biology to agronomy of yield and quality*. Food Products Press, New York, pp 1–29
- Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* 54:357–374
- Grando S, von Bothmer R, Ceccarelli S (2001) Genetic diversity of barley: use of locally adapted germplasm to enhance yield and yield stability of barley in dry areas. In: Cooper HD, Spillane C, Hodgkin T (eds) *Broadening the genetic base of crop production*, pp 351–372
- Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol* 57:461–485
- Heun M (1992) Mapping quantitative powdery mildew resistance of barley using a restriction fragment length polymorphism map. *Genome* 35:1019–1025
- Kilian B, Ozkan H, Kohl J, von Haeseler A, Barale F et al (2006) Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. *Mol Genet Genom* 276:230–241
- Kim HS, Ward RW (2000) Patterns of RFLP-based genetic diversity in germplasm pools of common wheat with different geographical or breeding program origins. *Euphytica* 115:108–197
- Kraakman AT, Niks RE, Van PM, den Berg P, Stam FA, Eeuwijk Van (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* 168:435–446
- Kunzel G, Korzun L, Meister A (2000) Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* 154:397–412
- Lin JZ, Morrell PL, Clegg MT (2002) The influence of linkage and inbreeding on patterns of nucleotide sequence diversity at duplicate alcohol dehydrogenase loci in wild barley (*Hordeum vulgare* ssp. *spontaneum*). *Genetics* 162:2007–2015
- Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* 12:57–63
- Malysheva-Otto LV, Ganai MW, Roder MS (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). *BMC Genet* 24:6–7
- Morrell PL, Lundy KE, Clegg MT (2003) Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite migration. *Proc Natl Acad Sci USA* 100:10812–10817
- Morrell PL, Toleno DM, Lundy KE, Clegg MT (2005) Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. *Proc Natl Acad Sci USA* 102:2442–2447
- Park SDE (2001) Trypano tolerance in West African cattle and the population genetic effects of selection. University of Dublin, Dublin
- Payne RW, Harding SA, Murray DA, Soutar DM, Baird DB et al (2006) *GenStat release 9 reference manual. Part 2. Directives*. VSN International, Hemel Hempstead
- Piepho HP, Mohring J, Melchinger AE, Buchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161:209–228
- Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 60:227–237
- Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. *Am J Hum Genet* 67:170–181
- Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR et al (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 98:11479–11484
- Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L et al (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Mol Genet Genom* 274:515–527

- Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR et al (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proc Natl Acad Sci USA* 103:18656–18661
- Russell J, Fuller J, Young G, Thomas B, Taramino G et al (1997) Discriminating between barley genotypes using microsatellite markers. *Genome* 40:442–450
- Russell J, Ellis RP, Thomas B, Waugh R, Provan J et al (2000) A retrospective analysis of spring barley germplasm development from ‘foundation genotypes’ to currently successful cultivars. *Mol Breed* 6:553–568
- Shtaya MJY, Marcel TC, Sillero JC, Niks RE, Rubiales D (2006) Identification of QTLs for powdery mildew and scald resistance in barley. *Euphytica* 151:421–429
- Smith JSC, Kresovich S, Hopkins MS, Mitchell SE, Dean RE et al (2000) Genetic diversity among elite sorghum inbred lines assessed with simple sequence repeats. *Crop Sci* 40:226–232
- Steffenson BJ, Olivera P, Roy JK, Jin Y, Smith KP et al (2007) A walk on the wild side: mining wild wheat and barley collections for rust resistance genes. *Aust J Agr Res* 58:532–544
- Stracke S, Presterl T, Stein N, Perovic D, Ordon F et al (2006) Effects of introgression and recombination on haplotype structure and linkage disequilibrium surrounding a locus encoding Bymovirus resistance in barley. *Genetics* 175:805–817
- Thomas WTB, Powell W, Waugh R, Chalmers KJ, Barua UM et al (1995) Detection of quantitative trait loci for agronomic, yield, grain and disease characters in spring barley (*Hordeum vulgare* L.). *Theor Appl Genet* 91:1037–1047
- Waugh R, Jannink JL, Muller K, Ramsay L (2009) The emergence of whole genome association scans in barley. *Curr Opin Plant Biol* 12:1–5
- Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics* 35:235–254
- Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E et al (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. *Proc Natl Acad Sci USA* 101:9915–9920
- Wenzl P, Li H, Carling J, Zhou M, Raman H et al (2006) A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. *BMC Genom* 7:206
- Yahiaoui S, Igartua E, Moralejo M, Ramsay L, Molina-Cano JL et al (2008) Patterns of genetic and eco-geographical diversity in Spanish barleys. *Theor Appl Genet* 116:271–282
- Yu J, Pressoir G, Briggs WH, Vroh B, Yamasaki IM et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C et al (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet* 3:e4