Geography and end use drive the diversification of worldwide winter rye populations

FLORENCE PARAT,* GRI T SCHWERTFIRM,† ULR IKER UDOLPH,† THOMAS MIEDANER,‡ VIKTOR KORZUN,§ EVA BAUER,† CHRIS-CAROLIN SCHÖN† and AURÉLIEN TELLIER*
*Population Genetics, TUM School of Life Sciences Weihenstephan, Technische Universität München, Liesel-Beckmann-Str. 2, 85354 Freising, Germany, †Plant Breeding, TUM School of Life Sciences Weihenstephan, Technische Universität München, Liesel-Beckmann-Str. 2, 85354 Freising, Germany, ‡State Plant Breeding Institute, Universität Hohenheim, Fruwirthstr. 21, 70599 Stuttgart, Germany, §KWS LOCHOW GMBH, 29296 Bergen, Germany

Abstract

To meet the current challenges in human food production, improved understanding of the genetic diversity of crop species that maximizes the selection efficacy in breeding programs is needed. The present study offers new insights into the diversity, genetic structure and demographic history of cultivated rye (Secale cereale L.). We genotyped 620 individuals from 14 global rye populations with a different end use (grain or forage) at 32 genome-wide simple sequence repeat markers. We reveal the relationships among these populations, their sizes and the timing of domestication events using population genetics and model-based inference with approximate Bayesian computation. Our main results demonstrate (i) a high within-population variation and genetic diversity, (ii) an unexpected absence of reduction in diversity with an increasing improvement level and (iii) patterns suggestive of multiple domestication events. We suggest that the main drivers of diversification of winter rye are the end use of rye in two early regions of cultivation: rye forage in the Mediterranean area and grain in northeast Europe. The lower diversity and stronger differentiation of eastern European populations were most likely due to more intensive cultivation and breeding of rye in this region, in contrast to the Mediterranean region where it was considered a secondary crop or even a weed. We discuss the relevance of our results for the management of gene bank resources and the pitfalls of inference methods applied to crop domestication due to violation of model assumptions and model complexity.

Keywords: ABC model, domestication, genetic diversity, population structure, SSR marker

Received 26 November 2014; revision received 13 October 2015; accepted 16 November 2015

Introduction

Inferring the genetic structure and evolutionary history of populations is of broad interest not only for evolutionary biologists but also for agricultural scientists. To meet the current and future challenges in human food production, there is a need for science-based population management in breeding programs to maintain the genetic diversity of a species and to exploit the genetic potential of populations. While archaeology can help trace the historical spread of a crop, knowledge of the genetic structure of populations can shed light on other aspects of the domestication history, such as the development of different cultivation methods or crop end use and their impact on genetic diversity. For example, small grain cereals such as rye, wheat, barley and triticale are cultivated as spring or winter growth types that can be clearly separated into two distinct pools at the genetic level (Ma et al. 2004; Mälysheva-Otto et al. 2006; Cockram et al. 2008; Chao et al. 2010; Alheit et al. 2012). Contrasting use of barley varieties is mainly related to differences between spring and winter and two-rowed and six-rowed ear types (Fischbeck 2003). Cultivated rye (Secale cereale L.) is used as a grain crop for bread

© 2015 John Wiley & Sons Ltd
making, brewing, distilling and animal feed and as a forage crop in the form of green chop, pasture, green manure or haylage (Miedaner 2010). Modern rye breeding populations, hereafter referred to as varieties, are adapted to two main end uses, grain or forage, or to a mixture of both; however, it remains unclear where and when the primary use diversification took place.

From an archaeological perspective, there is a general agreement that cultivated rye originated from the Fertile Crescent (Khush 1963). Wild relatives or weedy forms reached Europe probably through a northern route, and remains were present at archaeological sites dating to the late Neolithic age in Poland and Romania, to the Bronze Age in the Czech Republic, Slovakia and Ukraine and to the Iron Age in Germany, Denmark and Crimea (see references in Zohary et al. 2013). Studies on archaeological remains indicated that rye most likely spread as a weed among wheat and barley fields throughout Europe (Behre 1992). Therefore, the first domestication of rye most probably happened through conscious or unconscious selection by early Neolithic farmers, around 4500 BC, leading to a tough rachis, an erect growth habit and large grains (Khush 1963; Behre 1992). The change in harvesting methods during the pre-Roman Iron Age is seen as a principal cause of rye cultivation (Behre 1992). Archaeological records confirm that, from this time on, rye was permanently represented in the seed corn, which are the sets of grains reserved for planting. The cold, harsh conditions of northern Europe were more favourable to rye than to wheat and barley; thus, the weedy rye began to crowd out the main crops and started to be used on its own for grain production. In central and eastern Europe, rye was the main crop from the early 8th century onwards. In some areas of the Netherlands and Northern Germany, rye was even cultivated year after year on the same fields (‘eternal rye cultivation’, Behre 1992).

This traditional importance of grain rye in northern and eastern Europe can be explained by the superiority of rye over other cereals under suboptimal conditions, which lies in its strong winter hardiness, modest nutritional requirement and natural resistance against many pathogens (Miedaner 1997, 2010). Although northeast Europe is still the main rye producing region, its dominance has decreased there since the 1950s, when rapid progress in both breeding and cultivation techniques enabled other crops to cope with suboptimal environmental conditions in these geographic areas (Miedaner 1997). In southern Europe and the Mediterranean, the culture conditions are more favourable to other cereals, and rye was mainly present as a weed or used as forage. As rye was mostly used as forage or for grain and forage (mixed use) in the Mediterranean region and because the exclusive use of rye for grain is more ancient in northeast Europe, we interchangeably use the ancestral main end use and geographical origin to characterize the populations: weedy rye for the uncultivated western Asian populations, forage for populations of southern European ancestry and grain rye for populations that have ancestors in eastern or northern Europe.

A number of research studies have investigated the genetic diversity and structure of rye populations. Based on different marker systems and plant material sets, previous studies identified three factors that led to a clustering of rye populations: spring or winter growth habit (Ma et al. 2004), population history (geographical origin, relatedness and dispersion routes) and level of improvement (Persson et al. 2001). Indeed, because landraces are old, locally adapted populations that underwent little or no mass selection usually present higher levels of diversity than the officially registered varieties that are usually more recent (later than the 1940s for rye) and are products from breeding cycles (Villa et al. 2005). However, the literature also contains conflicting results concerning expected levels of genetic diversity between populations. For example, contrary to a study on Portuguese rye populations by Matos et al. (2001), Ribeiro et al. (2012) suggest that the diversity of Portuguese populations was higher than the diversity of northern European ones (Persson & von Bothmer 2002). Additionally, as expected in most crops but contrary to what has been shown in previous studies (e.g. by Persson et al. 2001), Ribeiro et al. (2012) found higher levels of diversity in rye landraces than in varieties. To address these discrepancies in the literature, we analysed rye population samples using simple sequence repeat (SSR) markers over a wide geographical range. The novelty in our approach lies in assessing the genetic diversity and structure of several populations of grain, forage and weedy rye to infer their history.

We combined a raw analysis of summary statistics with more computationally intensive Bayesian model-based approaches, STRUCTURE (Pritchard et al. 2000) and approximate Bayesian computation (ABC, Beaumont et al. 2002), to investigate the population history and genetic diversity of rye. In particular, we (i) investigated the genetic diversity of weedy rye populations from the centre of origin (Fertile Crescent), as well as landrace populations and modern varieties from eastern and central Europe, the Mediterranean basin and the Americas to test the hypothesis that recurrent crop improvement is related to a reduction in genetic diversity, (ii) sampled grain and forage rye populations to assess the relationship between genetic diversity and end use of rye populations and (iii) inferred the evolutionary development and historical split of grain and forage rye from ancestral weedy populations. Fourteen weedy, grain and forage winter rye populations comprising 37–45
individuals each were selected to represent the global distribution of rye as well as different population improvement levels. We genotyped these populations at 32 genome-wide distributed microsatellite marker loci. Using classic population genetics analyses, we found different levels of genetic diversity across populations, with the weedy populations exhibiting the highest values. Bayesian structure analysis resulted in two main subgroups indicated by a differentiation, which can be described as ‘Mediterranean and weedy rye’ and ‘northeast European rye’. Using ABC, we additionally inferred the likely divergence date of each grain and forage population from weedy rye populations. We finally discuss the usefulness of population genetics methods to retrace the history of crop domestication and new specific information regarding the evolutionary development of rye.

Materials and methods

**Sampling and SSR genotyping**

We chose 14 open-pollinated winter rye populations to represent rye growing areas in Europe and the Americas. The populations were propagated by several cycles of cross-pollination under isolation from seeds obtained from different gene banks or from plant breeders. We randomly sampled between 37 and 45 $S_0$ plants per population, representing a total of 620 $S_0$ individuals (Fig. S1, Supporting information; Table 1). Genomic DNA of the 620 individuals was extracted from leaf samples as described in Rogowsky et al. (1991), and then all were genotyped with 32 unlinked and genome-wide distributed SSR markers following established protocols (Tables S1 and S2, Supporting information). Briefly, separation of fragments by polymerase chain reaction (PCR) was carried out on a 3130xl Genetic Analyzer (Applied Biosystems Inc., Foster City, CA, USA). Alleles were assigned using the software GEMAPPER v. 4.0 (Applied Biosystems Inc.). The software determined fragment lengths using size standards of a known length. We observed stepwise length differences, which would be expected from varying microsatellite repeat numbers, as well as fragment size deviations from the expected due to additional insertions/deletions outside the microsatellite motifs. To improve the SSR marker data quality, allele assignments were checked manually, and ambiguous results were set to ‘missing data’. For the population genetics statistics, we considered null alleles as additional valid alleles for each marker. The global statistical trend remained similar whether null alleles were considered as valid or set to missing, for example, the populations with a higher expected heterozygosity conserved this characteristic (Table S3, Supporting information). Each marker was tested for deviations from the Hardy–Weinberg equilibrium (HWE) within populations using the $\chi^2$ goodness-of-fit test with a Benjamini–Hochberg correction for multiple testing (Table S4, Supporting information), and marker informativeness was measured as the

<table>
<thead>
<tr>
<th>Code</th>
<th>Origin</th>
<th>Usage</th>
<th>Breeding level</th>
<th>Individuals</th>
<th>Alleles</th>
<th>$A_p$</th>
<th>$A_e$</th>
<th>$H_{obs}$</th>
<th>$\bar{H}$</th>
<th>$F_{IS}$</th>
<th>Population name</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR1</td>
<td>Iran</td>
<td>Weedy</td>
<td>Primitive rye</td>
<td>44</td>
<td>161</td>
<td>8</td>
<td>2.99</td>
<td>0.46</td>
<td>0.59</td>
<td>0.23</td>
<td>Altevogt 14160</td>
</tr>
<tr>
<td>IR2</td>
<td>Iran</td>
<td>Weedy</td>
<td>Primitive rye</td>
<td>45</td>
<td>233</td>
<td>17</td>
<td>3.89</td>
<td>0.52</td>
<td>0.67</td>
<td>0.21</td>
<td>IRAN GP.IX</td>
</tr>
<tr>
<td>TUR</td>
<td>Turkey</td>
<td>Weedy</td>
<td>Primitive rye</td>
<td>45</td>
<td>243</td>
<td>41</td>
<td>4.37</td>
<td>0.55</td>
<td>0.69</td>
<td>0.19</td>
<td>Türkischer Unkrautrogen</td>
</tr>
<tr>
<td>ESP</td>
<td>Spain</td>
<td>Forage</td>
<td>n.i.</td>
<td>37</td>
<td>174</td>
<td>8</td>
<td>3.24</td>
<td>0.47</td>
<td>0.61</td>
<td>0.22</td>
<td>R778 (‘Villablancia‘)</td>
</tr>
<tr>
<td>BRA</td>
<td>Brazil</td>
<td>Forage</td>
<td>n.i.</td>
<td>44</td>
<td>140</td>
<td>4</td>
<td>2.58</td>
<td>0.47</td>
<td>0.57</td>
<td>0.18</td>
<td>Cenite Broco</td>
</tr>
<tr>
<td>USA</td>
<td>USA</td>
<td>Forage</td>
<td>Variety</td>
<td>45</td>
<td>132</td>
<td>2</td>
<td>2.48</td>
<td>0.45</td>
<td>0.54</td>
<td>0.16</td>
<td>Florida Black</td>
</tr>
<tr>
<td>PRT</td>
<td>Portugal</td>
<td>Forage</td>
<td>Landrace$^1$</td>
<td>45</td>
<td>199</td>
<td>9</td>
<td>3.22</td>
<td>0.52</td>
<td>0.63</td>
<td>0.15</td>
<td>R1008 (‘Malhadas‘)</td>
</tr>
<tr>
<td>ARG</td>
<td>Argentina</td>
<td>Forage</td>
<td>Landrace$^1$</td>
<td>45</td>
<td>87</td>
<td>0</td>
<td>1.81</td>
<td>0.29</td>
<td>0.45</td>
<td>0.35</td>
<td>Pico Gentejio</td>
</tr>
<tr>
<td>DE1</td>
<td>Germany</td>
<td>Grain</td>
<td>Old variety</td>
<td>45</td>
<td>114</td>
<td>0</td>
<td>2.21</td>
<td>0.39</td>
<td>0.48</td>
<td>0.21</td>
<td>Carokurz</td>
</tr>
<tr>
<td>RU1</td>
<td>Russia</td>
<td>Grain</td>
<td>Landrace</td>
<td>45</td>
<td>125</td>
<td>5</td>
<td>2.62</td>
<td>0.43</td>
<td>0.57</td>
<td>0.23</td>
<td>Karelsiche Landsorte</td>
</tr>
<tr>
<td>RU2</td>
<td>Russia</td>
<td>Grain</td>
<td>Landrace</td>
<td>44</td>
<td>109</td>
<td>0</td>
<td>2.23</td>
<td>0.39</td>
<td>0.48</td>
<td>0.17</td>
<td>Leningrader Landsorte</td>
</tr>
<tr>
<td>BLR</td>
<td>Belarus</td>
<td>Grain</td>
<td>Variety</td>
<td>43</td>
<td>156</td>
<td>3</td>
<td>2.60</td>
<td>0.49</td>
<td>0.56</td>
<td>0.12</td>
<td>Beloruskaja</td>
</tr>
<tr>
<td>DE2</td>
<td>Germany</td>
<td>Grain</td>
<td>Old variety</td>
<td>45</td>
<td>118</td>
<td>2</td>
<td>2.27</td>
<td>0.39</td>
<td>0.48</td>
<td>0.18</td>
<td>Halo</td>
</tr>
<tr>
<td>POL</td>
<td>Poland</td>
<td>Grain</td>
<td>Improved landrace</td>
<td>45</td>
<td>120</td>
<td>0</td>
<td>2.19</td>
<td>0.40</td>
<td>0.49</td>
<td>0.16</td>
<td>Dankowskie Selektyne</td>
</tr>
</tbody>
</table>

n.i., no information.

$^*$Synonymous name.

$^1$Matos et al. (2001).

$^2$Stracke et al. (2003).
polymorphism information content (PIC, Botstein et al. 1980). We defined the following groups of populations by rye main ancestral end use: grain, which assembled the populations of northeast European ancestry; forage, which assembled the populations of Mediterranean ancestry and weedy rye, for the weedy populations from the centre of diversity (Table 1).

Genetic diversity and population structure

We computed for each population the total number of alleles, number of private alleles ($A_p$), number of effective alleles ($A_e$), Kimura & Crow 1964), observed heterozygosity ($H_{obs}$), genetic diversity ($H$, Nei 1987) and Garza-Williamson’s $M$ (Garza & Williamson 2001). The analysis of molecular variance (AMOVA, Excoffier et al. 1992) was computed with ARLEQUIN v. 3.5.1.3 (Excoffier & Lischer 2010). The total variance was partitioned into components due to differences among the three defined groups ($V_a$), differences among populations within those groups ($V_b$) and differences among individuals within populations ($V_c$). Variance components ($V_a$, $V_b$ and $V_c$) were used to calculate the fixation indices ($F$-statistics; $F_{CT}$, $F_{SC}$, $F_{ST}$) according to Weir & Cockerham (1984). $F$-statistics were preferred to allele size-based measures of differentiation based on the results of the allele permutation test proposed by Hardy et al. (2003), which was performed across loci for each pair of populations using the software SPAGEDI 1.5 (Hardy & Vekemans 2002; Fig. S2, Supporting information). However, the allele size data set, without null alleles, was used for the computation of Garza-Williamson’s $M$. This statistic is used to detect recent bottlenecks and is based on the ratio of the number of microsatellite alleles to the observed range in allele size. Although its interpretation can vary depending on mutational models and demography, an $M$ value lower than 0.68 would generally be considered significant (Garza & Williamson 2001). The genetic diversity among weedy, forage and grain groups was denoted as $F_{CT}$, among populations within groups as $F_{SC}$ and among populations denoted as $F_{ST}$. The within-population fixation index $F_{IS}$ was also computed. Jost’s $D$ (Jost 2008) was used as an alternative measure of population differentiation. Measures of genetic diversity and pairwise $F_{ST}$ were aggregated over usage groups and levels of improvement (Table S5, Supporting information). A neighbour joining tree was drawn based on $F_{ST}$ distances using the R-package ‘ape’ v. 3.2 (Paradis et al. 2004). We calculated the pairwise distance between populations based on the proportion of shared alleles (Dps, Fig. S3, Supporting information) as described in Bowcock et al. (1994). Population structure was further investigated by a Bayesian clustering approach implemented in the STRUCTURE software v. 2.2 (Pritchard et al. 2000). Burn-in period and Markov Chain Monte Carlo iterations were both set to 50 000. Ten runs were executed for each number of $K$ assumed subgroups ($K = 1, 2, \ldots, 15$). Resulting membership coefficients from each run were averaged by individual and visualized in a bar chart. The average log-likelihood ($\pm$SD) was calculated for each $K$ to deduce the most probable number of $K$ subgroups. Ten individual runs per $K$ were plotted to check convergence. Additionally, $\Delta K$ was calculated according to Evanno et al. (2005) to determine the optimum $K$ for the uppermost hierarchical level of structure (Fig. S4, Supporting information).

Approximate Bayesian computations

Different scenarios of population split were studied using ABC. This method allowed us to compare demographic scenarios and estimate population parameters using prior knowledge about rye history without evaluating the likelihood function analytically. Due to the large number of populations analysed (11, see below), the number of possible demographic models of population split was very large. Moreover, preliminary ABC analyses demonstrated that models allowing for complex relationships between populations had poor reliability (see Discussion and Fig S8, Supporting information). To maximize the statistical power of the ABC method, we studied a simplified model with three populations defined as one weedy and two derived populations (Fig. 1). Each of these population trios was composed of one of three ancestral weedy populations (IR1, IR2 and TUR), one of the derived grain populations (BLR, DE1, POL, DE2, RU1 and RU2) and one of the forage populations (ARG, BRA, ESP, PRT and USA). Due to features that could lead to unreliable ABC results, we removed ARG, ESP and PRT populations from the ABC analysis (see Results). The weedy population of each trio was characterized by the population size $N_a$, while the derived grain and forage populations were characterized by sizes $N_g$ and $N_f$, respectively. The grain and forage populations were founded from the weedy population at times-to-merger $T_{2Mg}$ and $T_{2MF}$ generations ago, respectively. The founding event was modelled as a bottleneck, characterized by its length $BT_{Lg}$ (or $BT_{Lf}$) and by its strength $BT_{Ng}$ (or $BT_{Nf}$). At times $T_{Sa}$, $T_{Sg}$ and $T_{Sf}$ ago, the weedy, grain and forage populations underwent a short bottleneck of strength $PS_a$, $PS_g$ and $PS_f$, respectively. This recent bottleneck was introduced to mimic the effects of repetitive sampling and gene bank conservation (Fig. 1; Table S6, Supporting information). All bottlenecks were defined as stepwise population size changes. The strength of a bottleneck is the ratio of the
population size before and after the change. Our scenario modelled the origin of rye in the Fertile Crescent that was assumed to be a weedy population, and subsequent bottlenecks associated with the export and spread of rye as grain or forage to diverse parts of the world. Based on archaeological studies that show that rye was brought from the Fertile Crescent to northern Europe through an eastern migration route (via the Caucasus), we excluded models where the northern grain populations came indirectly from the ancestral weedy ones through the southern forage rye lineage (Zohary et al. 2013).

We used ABCtoolbox (Wegmann et al. 2010) to perform the ABC and 2,000,000 simulations were performed with the coalescent simulator fastsimcoal (Excoffier & Foll 2011). The model was defined by the 15 parameters described above for which we assumed prior distributions with wide bounds as information on rye domestication was scarce (Table S6, Supporting information). We simulated 32 independent SSR loci following a generalized stepwise model (GSM) with a mean mutation rate (MU) drawn from a log-uniform distribution between $10^{-5}$ and $10^{-2}$. These priors are consistent with mutation rates obtained from plant species such as wheat, Triticum turgidum (Thuijlet et al. 2002) and maize, Zea mays (Vigouroux et al. 2002). The average proportion ($P_{GSM}$) of mutations that affected the allele size by more than one step was drawn from a uniform distribution between 0.2 and 0.5. The mutation rate at each locus was drawn independently from a gamma distribution of mean MU and shape ALPHAMU, the latter varying between 2 and 30. This allowed for heterogeneity in mutation rates among loci (Excoffier et al. 2005; Xu et al. 2005). Similarly, $P_{GSM}$ per locus was drawn from a gamma distribution of shape ALPHAP with uniform priors between 1 and 4. The range for the prior distribution of the times-to-merger was chosen based on archaeological findings on rye domestication (Behre 1992), namely that forage and grain populations were established no more recently than 1000 but not more than 15,000 years ago. Since rye is an annual plant, years were simply considered equivalent to generations. The simulated data were summarized by the mean and standard deviation of genetic diversity and differentiation statistics over the 32 loci. The statistics for each population are the number of alleles, expected heterozygosity, and private alleles; and over all populations, we used the average, sum and standard deviation of the number of alleles, heterozygosity, Jost's $D$ over all populations and pairwise $F_{ST}$ (details in Table S7, Supporting information). All statistics were calculated with arlsumstat (Excoffier & Lischer 2010) except private alleles and Jost's $D$ that were calculated with a custom Perl script and subsequently compared with the observed values (Table 1).

**‘Scenario’ choice**

For each trio of populations, we performed a ‘model choice’ procedure to distinguish between two scenarios: (i) the forage population split from the weedy population at an earlier time than the grain population so that $T_{2Mf} > T_{2Mg}$ and conversely (ii) the weedy-grain split occurred before the weedy-forage split so that $T_{2Mg} > T_{2Mf}$. We computed acceptance rates for each scenario based on the simulations associated with the smallest Euclidian distances to the observed data (Pritchard et al. 1999). For a given relative distance $\delta$ to the observed data, we computed the Bayes Factor (BF) for scenario 1 against scenario 2 by dividing the number of retained simulations in the respective scenarios. In very rare cases, the two split times were equal, and these were not considered in the model choice. The model choice procedure was done using a range of relative distances $\delta$ (from 0.005% to 1%) as indicated in...
Pritchard et al. (1999). For each trio of populations, we manually checked that the statistical support for one scenario did not depend on $\delta$ (Fig. S5, Supporting information). We summarized the Bayes Factor for scenario 1 against scenario 2 for all trios of populations using the 5000 simulations fitting best (namely, a relative distance of $\delta = 0.25\%$). The dimensionality of the summary statistics was reduced using a logistic regression on 500 000 simulations (Prangle et al. 2014). Statistics included in the regression were chosen based on significance and Bayesian information criterion (Table S7, Supporting information).

Parameter estimation

For each trio of populations, we estimated all parameters of our model but reported the results for the five most important: the population sizes ($N_a, N_f$ and $N_g$) and the times-to-merger for the grain and the forage populations ($T_{2Mg}$ and $T_{2Mf}$). The dimension of the summary statistics was reduced using a linear regression for each of the five estimated parameters as well as for the mutation rate per the method of semi-automatic approximate Bayesian computation proposed by Fearnhead & Prangle (2012). The coefficients of the linear regression were estimated using 500 000 simulations (Table S7, Supporting information). The Leuenberger & Wegmann (2010) postsampling GLM adjustment was used to estimate the posterior probability of the parameters, based on the 0.25% of the simulated data sets (i.e. 5000 simulations) closest to the observed data. Smaller relative distances (0.05% or 0.1%) gave similar results but less smooth densities. The shapes of the posterior distributions for each of these parameters are reported in Fig. S6 (Supporting information), and the mode and credibility intervals are given in Table S8 (Supporting information).

Validation of the ‘scenario’ choice and estimation procedure

We evaluated the power of the ABC method to discriminate between scenarios and to estimate parameters by analysing 1500 pseudo-observed data sets (PODs). The PODs were sampled from the model and prior distribution as described above and transformed using the regression coefficient used for transforming the summary statistics of the observed data. We evaluated the number of times that the correct scenario was found for a trio of populations, the so-called confusion matrix or Type-I error (Bertorelle et al. 2010), and the difference between the estimated and the true parameter value (as the mean percent error and the root relative mean square error).

Results

SSR genotyping

Genotyping of 620 individuals from 14 winter rye populations with 32 genome-wide SSR markers resulted in 374 alleles, including 99 private alleles (Table 1). One individual from each of the populations IR1, BRA and RU2 was excluded from the analyses due to more than 33% of missing data. On average, we observed 11.7 ± 9.2 (range: 2–40) alleles per locus, of which 7.07 ± 10.87 (range: 0–41) were private alleles ($A_{pr}$). Across populations, the observed heterozygosity per SSR was 0.44 ± 0.17, and the expected heterozygosity $H$ was 0.67 ± 0.14. For most of the 32 SSRs, $H_{obs}$ was smaller than $H$; however, the locus rms1218 (5R) showed an opposite pattern, suggesting that this locus might be an outlier (Tables 1 and S4, Supporting information). Deviation from the Hardy–Weinberg equilibrium can be due to multiple factors among which population substructure and sampling bias are the most common. The latter might explain the high number of deviating sites in ARG and PRT.

Genetic diversity of populations

The weedy populations showed, as expected, a high genetic diversity with TUR having the highest number of 7.6 alleles per SSR and the highest heterozygosity ($H_{obs} = 0.55$ and $H = 0.69$; Table 1). Contrary to expectations, we found no pattern of reduced diversity in varieties compared with landraces. Indeed, the five populations showing the lowest diversity ($<2.3$ alleles per marker and $H < 0.5$) included two varieties (DE1 and DE2) and three landraces (ARG, POL and RU2). Interestingly, four of the five populations were of northeast European ancestry and mainly used for grain production. The South American landrace ARG showed an especially low genetic diversity ($H = 0.45$) and had the highest inbreeding coefficient ($F_{IS} = 0.35$), indicating a possible small population size and strong bottleneck during its establishment or sampling bias. The lowest $F_{IS}$ value was found for the grain population BLR ($F_{IS} = 0.12$). AMOVA results showed that molecular variation was mainly (79.70%) found among individuals within populations as expected for cross-pollinated species, whereas variation observed among populations within groups explained 16.39% and the variance among groups only 3.91% of the total genetic variability (Table 2). Although variance among groups was small, permutation tests indicated that both populations and groups explained variance significantly better than random assignments ($P < 0.01$). We further
investigated the differentiation among populations and groups below.

Genetic relationships between the populations

Pairwise $F_{ST}$ and Jost’s $D$ values were calculated to indicate the level of differentiation between populations (Fig. 2). As expected, we found a relationship among $F_{ST}$, Jost’s $D$ and genetic diversity, as pairwise population comparisons containing population PRT achieved lower values and comparisons containing DE1, RU2 or ARG revealed higher values than the other comparisons (Table 1; Fig. 2). This resulted from the enhanced effect of drift in populations with small size that increased allele fixation and differentiation between populations.

Table 2 AMOVA results including fixation indices $F_{CT}$, $F_{SC}$ and $F_{ST}$ for the total population. The genetic differentiation among weedy/forage/grain groups is denoted as $F_{CT}$, among populations within groups as $F_{SC}$ and among populations as $F_{ST}$

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Proportion of explained variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among groups</td>
<td>0.04</td>
</tr>
<tr>
<td>Among populations within groups</td>
<td>0.16</td>
</tr>
<tr>
<td>Within populations</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Fixation indices

$F_{CT}$ 0.04  
$F_{SC}$ 0.17  
$F_{ST}$ 0.20

(D1, D2 or D3 exhibited the smallest expected heterozygosity; Table 1). The relationship between diversity and ancestry appeared to be consistent overall as we observed low $F_{ST}$ and Jost’s $D$ values within weedy and forage groups and among these two groups, while the grain group showed higher values within and among group differentiation. The neighbour joining tree (Fig. 3) confirmed that these results showing a clear group containing the weedy populations, a relatively close but clearly separated group of three forage populations (BRA, ESP and USA) and distinct groups containing the grain populations as well as the two forage populations ARG and PRT. These three clusters could also be observed in the pairwise distance between populations based on the proportion of shared alleles (Fig. S3, Supporting information). As expected for poorly differentiated populations with few private alleles, $F_{ST}$ and Jost’s $D$ showed a similar ordering. However, note that Jost’s $D$ indicated a generally higher differentiation than $F_{ST}$ (Fig. 2) because it was less biased by the high mutation rate of microsatellite markers (Jost 2008).

Population structure analysis

Individual-based grouping was investigated using STRUCTURE for values of $K$ ranging from 2 to 15 (Fig. 4). The log-likelihood curve for different $K$ values did not show a clear plateau (Fig. S4, Supporting information). However, the log-likelihood values started stabilizing around $K = 10$ indicating a possible optimal number of groups between 10 and 12. The differentiation of groups reflected the ancestral main end use and origin of the
rye populations. At $K = 2$, the optimal number of groups based on $\Delta K$, the first subgroup (green) comprised the forage and weedy populations IR1, IR2, TUR, BRA, USA and ESP and was clearly separated from the second subgroup (red) of grain populations BLR, DE1, POL, DE2, RU1 and RU2. The forage population ARG was also found in this group of grain populations. Those subgroups were already observed in the pairwise $F_{ST}$ and Jost’s $D$ statistics. The subgroup comprising forage and weedy populations showed a low level of differentiation, whereas the grain populations were highly differentiated from the forage and weedy as well as among themselves. We found that the forage population PRT could not be assigned to either of the groups as the individuals exhibited mixed membership coefficients. This prompted us to subsequently remove this population from the ABC analysis. When increasing $K$ from five to seven, several geographical subgroups appeared: (1) the three weedy rye populations IR1, IR2 and TUR, (2) the two Russian populations RU1 and RU2, (3) the eastern and central European populations BLR, POL and DE2, (4) the ESP population from Spain with BRA and USA and (5) the German population DE1 with the Argentinian population ARG. These clusters were similar to those observed in the neighbour joining tree (Fig. 3). At $K = 10$, most of the populations appeared separated from each other except IR2 and TUR in the weedy group and BLR, DE2 and POL in the
The high fragmentation of membership coefficients observed for the sampled individuals of IR2 and PRT reflected the very high diversity observed in these populations. At $K = 12$, only the two populations BLR and DE2 clustered together, agreeing with the low $F_{ST}$ and Jost’s $D$ between these populations (Fig. 2).

Approximate Bayesian computation

ABC was performed for 12 of the 14 populations. The Portuguese PRT and Argentinean ARG populations showed atypical characteristics of very high and very low diversity, respectively, suggesting a demographic history that could not be described by our model. As we used a common set of simulations for all population trios, the Spanish forage population ESP was kept in the model choice procedure but was not used for parameter estimation because of its smaller sample size (37 individuals). A smaller sample size could potentially lead to an overestimation of the time-to-merger due to incommensurable summary statistics (e.g. number of alleles, private alleles and $F_{ST}$). Our power analysis of the model choice procedure showed that we had moderate power to discriminate between the two scenarios (percentage of PODs attributed to the wrong scenario was 32%). Additionally, we achieved high power in only some of our parameter estimations as shown by low (8% for population sizes) to high (55% for times) mean percent error (Tables S9 and S10; Fig. S7, Supporting information). The ABC model selection indicated different split orders of the forage and grain populations depending on the considered trios, as indicated by the BF per trio in Fig. 5 (the forage population is on the x-axis). Most of the trios involving the USA forage population had $BF < 1$; therefore, we concluded that the split of this population from the weedy group occurred more recently than splits of the other two forage populations. The grain populations DE1 and RU2 consistently split earlier than the three forage populations from the different weedy populations. In contrast, the grain population RU1 split later than all of the forage populations (Fig. 5). We summarized the estimation results in Fig. 6, by representing each mode of the marginal posterior distributions of the time-to-merger for the forage and grain populations. Table S8 (Supporting information) indicates the minimum and maximum values of highest posterior density (HPD) 50% and 95% calculated over all posterior distributions of a given

Fig. 5 Graphical summary of the ABC results regarding forage and grain time-to-merger. Bayes Factor (BF) for Scenario 1: $T_{2Mf} > T_{2Mg}$ over Scenario 2: $T_{2Mf} < T_{2Mg}$ for each trio of populations (weedy, grain and forage). The BF has been estimated as the ratio of the number of accepted simulation for each scenario when taking 5000 simulations fitting best ($\delta = 0.25\%$). The dots above the horizontal (red) line represent a BF $> 1$, indicating that Scenario 1 (i.e. the split of the forage population is older than split of the grain population) is more strongly supported by the trio under consideration than Scenario 2 (i.e. the split of the grain population is older than split of the forage population). The distance to the BF $= 1$ horizontal line indicates the strength of evidence for the given scenario. Dots on or close to the BF $= 1$ line indicate a lack of support for one or the other order of split.

© 2015 John Wiley & Sons Ltd
time split. Note that, as expected, the similar estimated values of times-of-split of the grain populations BLR, POL and DE2 were in line with the pairwise $F_{ST}$ and Jost’s $D$ values and the STRUCTURE analysis. Indeed, two populations showing a recent split from the ancestral weedy population would exhibit lower genetic differentiation between them than if they had diverged a longer time ago due to the action of genetic drift. The ancestral population sizes of the weedy group were estimated to be higher than those of the incipient populations. IR1 was estimated to be the smallest weedy population but with a higher effective population size of $N = 23\,000$ than all grain and forage populations, whose effective population sizes were found to be in the range of 8700 and 10 100. The population sizes matched the overall expectations based on heterozygosity, except for RU1 that had a lower population size than expected (Tables 1 and S8, Supporting information).

**Discussion**

In this study, 14 winter rye populations were genotyped with SSRs. These populations had different within population levels of diversity (number of alleles and heterozygosity) while $F_{ST}$ and Jost’s $D$ values revealed differentiation among the 14 populations and confirmed the assumption of considerable population structure within cultivated rye. The STRUCTURE analysis revealed two main subgroups indicating a differentiation according to both geography and end use, which can be described as ‘southern European forage rye’ vs. ‘northern European grain rye’. The observed genetic diversity and population structure of a global collection of rye populations with different end uses and improvement levels suggest that (i) forage populations have reached a lower diversification level than grain ryes and (ii) the strong structuration of populations according to geography and usage may be the root of disagreements among previous studies. In addition, the ABC results suggest that differences in time-of-splits might be attributed to different domestication events, possibly explaining the high genetic distances between geographically close populations.

*Cultivated ryes are at different diversification stages*

For any crop plant, it is assumed that genetic diversity decreases from its wild form over landraces to modern varieties (Yamasaki et al. 2005; Feuillet et al. 2008). Following this hypothesis, we tested whether the present sets of rye populations showed a decrease in genetic diversity concurrent with an increase in improvement status. We confirmed that primitive rye populations had a significantly higher level of genetic diversity than the landraces and varieties. Noticeably, the genetic diversity levels of landrace and variety populations were comparable; whereas differences in genetic diversity could be seen between grain and forage rye populations independently of the improvement status. Despite the high number of populations in this study, our statistical power to distinguish clearly between these two patterns was somewhat limited because the landraces and varieties were unequally distributed between grain and forage populations.

There was a higher pairwise differentiation among grain than among forage rye populations ($F_{ST}$ and Jost’s $D$), which indicated that these groups probably followed different domestication and/or artificial selection paths. The lower diversity found in grain populations suggested that the differentiation might have been accelerated by successive bottlenecks that go hand in hand with domestication and selective breeding, especially in cases of adaptation to diverse environments as
was the case between ARG and RU2 ($F_{ST} = 0.35$). Therefore, we suggest that the early end use of rye explains its peculiarity compared with most other cereals. In northeast Europe, rye was used exclusively for grain early on and was, therefore, submitted to a longer and more intensive selection consisting of successive bottlenecks and diversification of populations for local adaptation. Conversely, in southern Europe, where wheat performed better, rye cultivation was neglected, and it was mainly used as forage or remained as a weed among other crops and selection was less intensive. This divergent end use and breeding is much more ancient than the distinction between landraces and varieties and had a much stronger impact on diversity patterns and among population differentiation. Our results might explain the discrepancies found in earlier studies on rye as highlighted in the introduction. In fact, studies sampling from either forage or grain populations or across both usages would obtain very different results when comparing genetic diversity. This is especially true when stratifying the data based on the landrace or variety status of the populations (Table S5, Supporting information).

**Several domestication events**

The ABC results showed that some of the grain and forage populations may have split at different times from the weedy population. The estimated split times of the weedy populations were spread over 760 years from the Bronze Age (RU2) to the Iron Age (RU1), a period for which there is supporting archaeological evidence of rye populations in Europe. These results agree with the hypothesis of multiple domestication events proposed by archaeological studies in rye (Khush 1963; Sencer & Hawkes 1980; Behre 1992; Burger et al. 2008; Zohary et al. 2013). ABC performed using a single domestication event model with one forage and one grain ancestor splitting from the weedy population giving rise to our various rye populations, with or without migration between them (Fig. S8, Supporting information), were also tested, but we could not reproduce the data (results not shown), corroborating the previous results. Different waves of breeding or domestication in northern European populations would explain the patterns of population differentiation and the STRUCTURE results (cf. $K = 5$). For instance, despite their spatial proximity, the two Russian landraces RU1 and RU2 are differentiated from the eastern and central European populations BLR, POL and DE2 that cluster together. Ma et al. (2004) reported a similar distinction between Russian cultivars and those from Norway, Finland, Estonia, Ukraine and Poland. In contrast, we found similar times-of-split (360 years apart) for the two forage populations: USA and BRA from the American continent. USA is one of several of the southern North American varieties that originated from the Italian cultivar ‘Abruzzi’, imported by the US Department of Agriculture in the early 1900s to be used for pasture or as a cover crop (Briggle 1920). This explains why, in our study, USA genetically clustered with the southern European population ESP. The low differentiation between USA and BRA and their high shared population membership at $K = 7$ were also expected as BRA (‘Centeio Branco’, which means white rye) was one of the first rye populations brought from the USA and introduced to Brazil in 1984 (De Mori et al. 2013).

However, we cannot draw any conclusions on the multiplicity of domestication events because of the variance in split time estimates from different trios and the fact that, despite their likely common origin, the difference between time-of-split estimates for USA and BRA is of the same order of magnitude as the differences among grain populations (Fig. 6). Our ABC results highlight the difficulty of capturing complex demographic domestication events using simplified modelling and summary statistics based on microsatellite data. As very little is known about the history of the populations studied here, we have built a simple model of a population split using the scarce information obtained from archaeological studies, our analyses of genetic diversity and preliminary ABC simulations. We have modelled strong bottlenecks associated with population splits (founding events) as well as very recent bottlenecks to mimic the effect of conservation in gene banks. However, several other bottlenecks may have occurred in the history of the populations explaining low values of the Garza–Williamson statistics (between 0.35 and 0.41; Table S3, Supporting information) and in some cases, admixture events that would explain the mixed membership coefficients obtained from STRUCTURE analyses for PRT and POL. The bottlenecks detected mostly occurred in the last 500 generations, i.e. more recently than domestication, otherwise the populations would have recovered sufficiently and the GW statistics would be comparable to that of a population of constant size (Garza & Williamson 2001). Many cultivated species are known to have repeatedly experienced such complex demographic events in their history (Glemin & Bataillon 2009). Ultimately, the power of the ABC to study domestication (e.g. Cornille et al. 2012) will depend on the system studied (population history, type of markers and generation time), prior knowledge and violation of the model hypotheses. This calls for caution when performing an ABC. While it is possible to use more complex models such as many populations with bottlenecks, splits and introgression/gene flow, statistical power for the estimation of parameters will be low.

© 2015 John Wiley & Sons Ltd
Typically, the model choice procedure will lead to a low rate of correct assignment and parameter estimation does not give credible posterior distributions. In this study, we chose an alternative to derive sets of very simple models, each consisting of a limited number of populations and parameters. While it is clear that not accounting for these more complex features (i.e. repeated bottlenecks, admixture, substructure and migration) might bias our results, our model fitted the data reasonably well while giving sufficient power to compare scenarios (as shown in the moderate error rates in the confusion matrix; Table S9, Supporting information) and to infer the parameters of interest. Therefore, we concluded that the ABC can be assumed to give realistic estimates for relative split times between the grain/forage populations and the weedy populations, providing new insights into the history of these populations.

Impact of population management and the signature of bottlenecks

Our analyses highlighted possible issues concerning the sampling and management of these various populations. Despite generally high levels of diversity, all populations, including weedy ones, exhibited signs of recent bottlenecks with low GW values. The fact that the three groups have the same average value for the GW statistic (0.38) suggests that the detected bottlenecks might be due to the maintenance of accessions in seed banks rather than previous population history. Our ARG sample showed a relatively low genetic diversity and comparatively high inbreeding. Both facts might indicate that a recent bottleneck effect occurred in the original production area of South America. We hypothesize, however, that the genetic bottleneck is an artefact, caused by small and biased sampling of the original South American population, or because of recurrent bottlenecks during the maintenance of accessions in seed banks. Similar conclusions could be drawn for DE1, an old German variety, and RU2, a landrace of northwest Russia, although the effects on genetic diversity were less severe than for ARG. Another point of interest in our study is the unexpected shared ancestry between populations. For example, the grouping of ARG with DE1 among the European grain rye populations in the STRUCTURE analysis was unexpected, as DE1 originates from northern Germany and ARG from Argentina. It is noteworthy that a common ancestry is not known between these populations. In addition, STRUCTURE suggests that a high admixture occurred between PRT and other populations possibly resulting from severe seed and/or pollen contamination. In summary, genetic bottlenecks can occur in population management due to few seeds collected from the original population, limitations during seed propagation in gene banks (Chebotar et al. 2003; Börner et al. 2005) or because of the small number of seeds distributed by gene banks and made available for research. Moreover, introgression/contamination likely explains unexpected genetic similarities among populations although this cannot be confirmed as the breeding history is often unknown.

High diversity in weedy ryes

The three weedy rye populations in our study represented an area of Turkey and Iran, a region considered to be within the cereals’ domestication centre (Khush 1963; Behre 1992; Nesbitt & Samuel 1998; Badr et al. 2000) and the centre of maximum diversity for rye with regard to cytological and morphological aspects (Khush 1963). We found that, as expected, the three populations of weedy rye IR1, IR2 and TUR showed diversity parameters higher than the cultivated populations in this study since we used wild grown populations, which are more closely related to the wild rye ancestor and have most probably not been subjected to strong bottlenecks. The comparison of membership coefficients from the STRUCTURE analysis among IR2 individuals revealed high heterogeneity within this population. This observation is not surprising since the original seed collection from 1930 was composed of seeds from individuals growing in a wide area around Elburz-Karaj in Iran, and these were taken together as one population (Kuckuck 1956; Kranz 1957). In contrast, the weedy population TUR was collected from a single wheat field (Hartwig H. Geiger, personal communication). We expected individuals distributed over a wider area to be more genetically diverse and to deviate more from HWE compared with individuals from a population in close spatial proximity. Despite the sampling differences, IR2 and TUR showed comparable levels of diversity although IR2 exhibited an excess of rare alleles.

As we independently considered each trio of populations in the ABC, the parameter estimates for a given population could vary among trios. Some trios showed a high variance or outlier estimates for some parameters. For example, two trios of POL and DE2 had a very ancient split time compared with other estimates for these populations (Fig. 6). In several of these high variance cases, the outlier trios all contained the same TUR weedy population. We conclude that the three weedy populations are not identical and do not represent the true ancestor of cultivated populations. They exhibit an unknown history and possibly complex relationships with each other and to cultivated rye.
Conclusions

The present study used microsatellite markers, classical population genetics analyses and ABC modelling to offer new insights into the diversity and genetic structure of cultivated winter rye populations. Our results can be used as an ‘anchor’ for describing the population structure of cultivated rye. We conclude that two main roads of breeding occurred corresponding to the history of the spread of rye and the end use. For the first time, we provide time estimates for these events based on genetic data. Our results provide valuable data for effective population management and a rational use of the diversity present in genetic resources for the improvement of modern cultivars and hybrids.

Acknowledgements

We thank Dr. Peer Wilde, KWS LOCHOW GMBH (Bergen, Germany), for helpful discussions about rye history and Dr. Bernd Hackauf, Julius Kühn-Institut (Groß Lüsewitz, Germany), for making SSR primer sequences available. GS thanks the Deutsche Forschungsgemeinschaft (DFG) for financial support (grant no. HA 6185/1-1). FP and AT acknowledge support from the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr ‘Synbreed – Synergistic plant and animal breeding’ (grant no. 0315528I).

References

Alheit KV, Maurer HP, Reif JC et al. (2012) Genome-wide evaluation of genetic diversity and linkage disequilibrium in winter and spring triticale (xa Triticeaeae Wittmack). BMC Genomics, 13, 239.


© 2015 John Wiley & Sons Ltd


Yamasaki M, Tenaillon MI, Bi IV et al. (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell, 17, 2859–2872.


Data accessibility

All SSR-marker data necessary to reproduce the analysis presented in the article have been archived on Dryad (doi: http://dx.doi.org/10.5061/dryad.q0694).
Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Origins of the rye population samples.

Table S1 PCR primers and PCR conditions.

Table S2 Overview of SSR markers used for genotyping 14 winter rye populations.

Table S3 Populations under study and their genetic diversity after removing null alleles.

Table S4 $P$ values of $\chi^2$ goodness-of-fit test for HWE within each population.

Fig. S2 $P$ values of the allele permutation test across loci for each population pair (after Hardy et al.).

Table S5 Genetic diversity and FST aggregated (A) over usage groups and (B) over levels of improvement.

Fig. S3 Pairwise distance between populations based on the proportion of shared alleles (Dps).

Fig. S4 Graphical analyses of structure’s log-likelihood for each $K$ assumed subgroup.

Table S6 Definition and prior model parameters.

Table S7 Summary statistics.

Fig. S5 Cumulative counts of accepted simulations for each scenario depending on the total number of accepted simulations.

Fig. S6 Prior, truncated prior (posterior before GLM) and posterior (after GLM) distributions for the five estimated parameters and the mutation rate for one representative population trio.

Table S8 Average marginal mode of estimation values, range of estimations, HPD50 and HPD95 ranges.

Table S9 Confusion matrix.

Fig. S7 Model choice error rate as a function of the absolute ‘times of split’ difference.

Table S10 Error measures.

Fig. S8 Models with a single domestication event for all forage population and one for all grain populations with or without migration within groups.