Population structure and genetic diversity in red clover (Trifolium pratense L.) germplasm

Jones, Charlotte; de Vega, Jose; Lloyd, David; Hegarty, Matthew; Ayling, Sarah; Powell, Wayne; Skot, Leif

Published in:
Scientific Reports

DOI:
10.1038/s41598-020-64989-z

Publication date:
2020

Citation for published version (APA):
Jones, C., de Vega, J., Lloyd, D., Hegarty, M., Ayling, S., Powell, W., & Skot, L. (2020). Population structure and genetic diversity in red clover (Trifolium pratense L.) germplasm. *Scientific Reports*, *10*(8364), [8364]. https://doi.org/10.1038/s41598-020-64989-z

SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Population structure and genetic diversity in red clover (*Trifolium pratense* L.) germplasm

Charlotte Jones[1], Jose De Vega[2], David Lloyd[1], Matthew Hegarty[1], Sarah Ayling[2,3], Wayne Powell[1,4] & Leif Skøt[1]✉

Red clover (*Trifolium pratense* L.) is a highly adaptable forage crop for temperate livestock agriculture. Genetic variation can be identified, via molecular techniques, and used to assess diversity among populations that may otherwise be indistinguishable. Here we have used genotyping by sequencing (GBS) to determine the genetic variation and population structure in red clover natural populations from Europe and Asia, and varieties or synthetic populations. Cluster analysis differentiated the collection into four large regional groups: Asia, Iberia, UK, and Central Europe. The five varieties clustered with the geographical area from which they were derived. Two methods (BayeScan and Samβada) were used to search for outlier loci indicating signatures of selection. A total of 60 loci were identified by both methods, but no specific genomic region was highlighted. The rate of decay in linkage disequilibrium was fast, and no significant evidence of any bottlenecks was found. Phenotypic analysis showed that a more prostrate and spreading growth habit was predominantly found among populations from Iberia and the UK. A genome wide association study identified a single nucleotide polymorphism (SNP) located in a homologue of the *VEG2* gene from pea, associated with flowering time. The identification of genetic variation within the natural populations is likely to be useful for enhancing the breeding of red clover in the future.

Red clover is a forage legume, which is used primarily in temperate livestock agriculture. It declined in importance when industrially produced nitrogen fertiliser became available in the early to mid-20th century; but before that, it and other forage legumes were key to maintaining soil fertility, and providing high protein forage for ruminant livestock[1]. However, as we move towards a more sustainable agriculture policy, increased legume use is being encouraged, due to the recognition of their utility[2]. This includes increased production potential, especially in mixtures with grasses; environmentally friendly N input into grassland by virtue of their symbiotic N_2 fixation; high protein content, and higher voluntary intake and thus better livestock performance[2].

Red clover is a relatively newly domesticated species, and it was not until around 1000 years ago, that it was first intentionally grown in Southern Spain[1]. Many of the varieties currently on the market in Europe originate from a Mattenklee type of plant, with their growth habit characterised by tall upright stems. However, some natural populations show varied growth habits, characteristically with less dependency on the central crown and a more prostrate nature to stem growth. Under certain temperature and moisture conditions, stems of these prostrate plants are able to produce nodal root growth[3]. The genetics underlying the prostrate growth habit of such red clover ecotypes, as well as the latent disease resistance often associated with natural populations, is of considerable interest for breeding programmes, in terms of furthering the range of climatic and agricultural conditions suited to red clover growth and use[4]. When plants expand into new ecosystems they encounter new and diverse selective pressures[5,6], which may lead to changes in growth habit, morphology, the partitioning of metabolic resources and disease resistance. This diversity in growth habit has enabled red clover to adapt to varied environmental conditions[7].

The genetic diversity of natural and cultivated populations of red clover has previously been studied using methods of isoenzymes[8], AFLP[9,10] RAPD[11,12] and SSR markers[13]. All of these studies reported high genetic diversity, and that within population diversity was larger than between populations. It is now possible to obtain

[1]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, Ceredigion, United Kingdom. [2]The Earlham Institute, Norwich Research Park, Norwich, Norfolk, United Kingdom. [3]Present address: Sarah Dyer, National Institute of Agricultural Botany, Huntingdon Road, Cambridge, CB3 0LE, United Kingdom. [4]Present address: Scotland Rural College (SRUC), Edinburgh, United Kingdom. ✉e-mail: lfs@aber.ac.uk

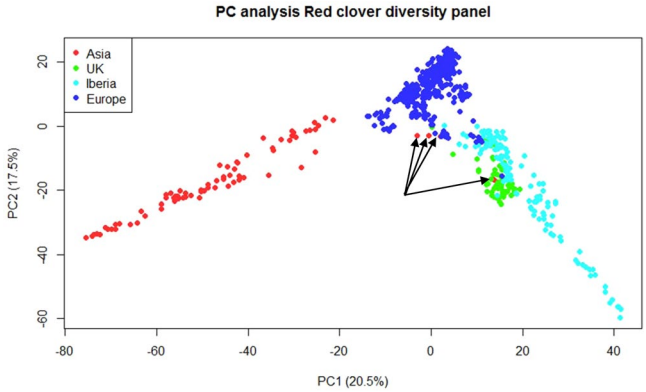

Figure 1. Population structure of the red clover accessions described by Principal Component (PC) analysis. The 4 red dots highlighted by arrows present in the centre of the UK and the Central Europe cluster represent the anomalous Italian accession Aa4445. This accession was described as being part of the Asian group by UPGMA.

large-scale genome wide variation in populations by using next generation sequencing (NGS) technologies. Genotyping by sequencing (GBS)[14] has become a popular method with which to identify large scale variation in species both with and without a reference genome. The use of GBS has allowed researchers to identify thousands to millions of single nucleotide polymorphism (SNP) markers, which can be used to analyse genetic variation within and between populations, and facilitate the analysis and dissection of complex traits, especially those involved in adaptive selection. While the bioinformatics analysis is more challenging compared to arrays, the cost per SNP is much lower, at least in less intensely studied crop species such as red clover.

The aim of the present work was to assess and analyse the genetic variation present in the accessions of the Genetic Resources Unit at the Institute of Biological, Environmental and Rural Sciences (IBERS). The relatively recent history of red clover cultivation and breeding, would lead to the expectation that differentiation between natural populations and varieties is likely to be small. We report here on the genotyping by GBS of 640 plants of red clover representing 70 natural populations from Europe and Asia and 5 modern varieties. These populations may represent a potential source of novel alleles, which could be used in breeding programmes to alter growth morphology and enhance traits such as disease resistance.

Results

GBS data and SNP detection. The sequencing produced an average of 1,700,000 reads per sample (range 203,996 to 5,897,687 of good barcoded reads). From these reads a total of 1,804,668 tags across all 640 plant samples were identified. The SNP analysis in Tassel v5.2, using the red clover genome assembly[15] as a reference, resulted in 264,927 SNPs across all 640 plants. Substantial filtering of the data (see SNP discovery in the Methods section), resulted in a panel of 12,577 robust polymorphic SNPs. In that process 10 samples were lost due to poor sequencing results. Of the 12,577 SNPs, 8,118 were mapped to the seven chromosomes, and the remaining 4,459 were located in unmapped scaffolds of the assembly. The data presented in the rest of this paper refer to the 8,118 SNPs mapped to the seven chromosomes. The SNPs were identified as transition or transversion SNPs (Supplementary Table S1).

Phylogenetic relationship. The 8,118 SNPs were used to assess the genetic relationship between the accessions according to the unweighted pair-group method of arithmetic mean (UPGMA) using the package Cluster. The change in slope angle (Supplementary Fig. S1a) indicated that the accessions split into four groups, and the hierarchical UPGMA tree confirmed this (Supplementary Fig. S1b). They consisted of accessions from Asia (including Iran), Central Europe, UK, and the Iberian Peninsula, respectively. Three varieties, Milvus, Britta and AberRuby all belonged to group 2 (Central Europe), and Grasslands Broadway and Crossway to group 4 (Iberian Peninsula), which is consistent with their derivation. The clustering of the accessions revealed two anomalies: the Italian populations Aa4445 and Aa4441 were placed in the clusters represented by Asia and UK, respectively, while all the remaining Italian ecotypes belonged in the Central European group (Supplementary Fig S1).

A principal component analysis (PCA) of the data was performed. The PCA plot (Fig. 1) illustrates the contribution of the first two principal components to the variation. It demonstrates that the Asian accessions are separate to the European ones, and that the Iberian and UK populations are close to each other, but distinct from those of Central Europe. The PCA indicated that four genotypes of the Italian accession Aa4445 are not part of the Asian cluster as indicated by the UPGMA analysis, but are genetically related to the Central Europe and UK groups (Fig. 1). This is not consistent with the result of the UPGMA analysis (Supplementary Fig. S1). The geographic location of where all the accessions were collected is shown in Fig. 2.

STRUCTURE analysis identified nine ancestral groups in the data as inferred by ΔK (Supplementary Fig. S2). STRUCTURE also identified a secondary peak at $K = 2$, subdividing the population into European and Asian accessions. The nine groups population (Supplementary Fig. S2) could also be classified into four sets, of Asia (group 3), Iberia (group 2 and 8), the UK (group 5), and more loosely Central Europe (group 1, 4, 6, 7, 9). Group 1 consisted mainly of Croatian and Bosnian accessions; group 4 of Italian and Slovakian accessions. Group

Figure 2. Map of collection sites of the ecotype accessions used for the red clover genetic diversity analysis. The data was accessed through Genesys Global Portal on Plant Genetic Resources, www.genesys-pgr.org, 2018-07-2, filtering for red clover and the holding institute GBR140. The plant genetic resources accession level data was provided by Aberystwyth University Plant Genetic Resources Unit.

	Asia	Central Europe	UK	Iberia	
H_O	0.208	0.258	0.256	0.251	
H_E	0.234	0.266	0.246	0.264	
F_{IS}	0.104**	0.055	−0.010	0.055	
Overall					
H_S	H_T	D_{ST}	F_{IS}	F_{IT}	F_{ST}
0.260	0.275	0.015	0.066	0.136	0.076***

Table 1. Basic population parameters generated by genetic diversity analysis calculated according to Nei (1977)[61]. The statistical significance of F_{IS} and F_{ST} were derived from χ^2 tests with one degree of freedom as $\chi^2 = NF_{IS}^2$, and $\chi^2 = 2NF_{ST}$, respectively; **($P < 0.01$); ***($P < 0.001$).

7 contained only one accession, that of the Italian Aa4445. Group 9 was a mix of accessions from Hungary, Poland, The Czech Republic and Slovakia. These four sub-groups appeared to be related by geographic distribution. However, group 6, which contained three of the varieties (Milvus, Britta and AberRuby), also consisted of accessions from Northern, Eastern Europe and Italy. There were four oddly placed accessions: Aa4441 from Italy in group 5 (UK), Aa4480 and Aa4487 from Spain and Aa4406 from Pakistan in group 6 (Central Europe). There were four accessions, from Bosnia (Aa4280), Bulgaria (Aa4355, Aa4358) and The Czech Republic (Aa4304), which appeared to consist of mixed populations. Individuals from these accessions belonged to groups 1, 4, 6 and 9 (Supplementary Fig. S2).

Genetic diversity of the red clover panel. Genetic diversity was evaluated according to the four groups identified by the hierarchical UPGMA analysis. This analysis showed that at the population level, there was more genetic diversity within the accessions than between them, as indicated by the H_S and D_{ST} values (Table 1). In the Asian subpopulation the H_O was lower than H_E, while in the three other populations the difference between the two parameters was smaller, or in the case of the UK population, H_O was higher than H_E (Table 1). This was reflected in the subpopulation inbreeding coefficient, F_{IS}, which was only significantly different from zero in the Asian population. The overall F_{ST} value indicated a low to moderate gene differentiation. The F_{IT} value indicated

	Asia	Europe	UK	Iberia
Asia	—	—	—	—
Europe	0.111	—	—	—
UK	0.171	0.067	—	—
Iberia	0.166	0.062	0.083	—

Table 2. Pairwise F_{ST} values calculated in Stampp, between the four groups as defined by UPGMA.

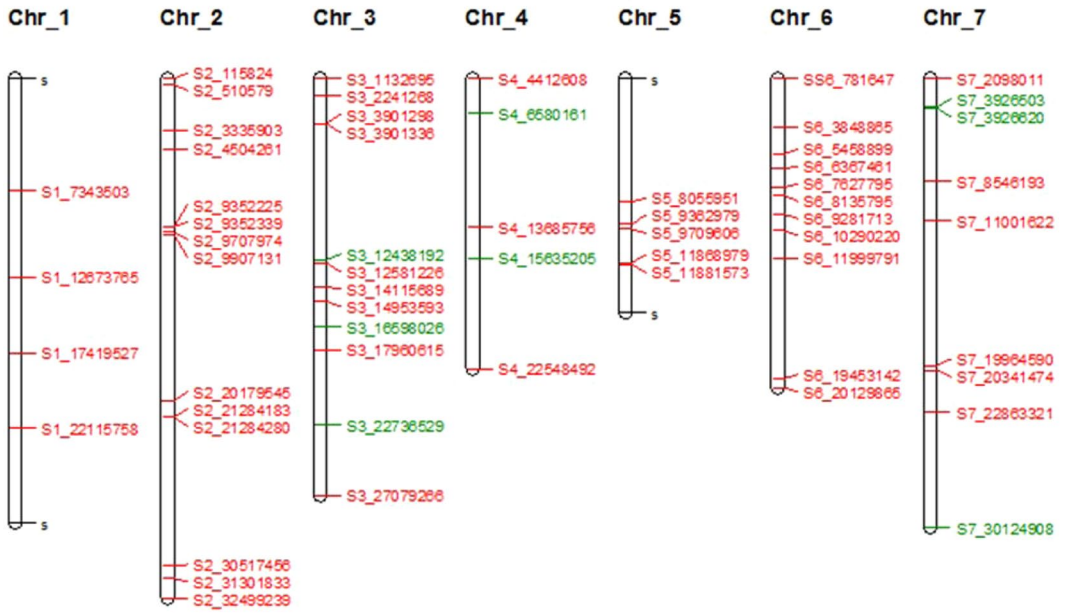

Figure 3. Molecular map showing position of loci potentially under selection as measured by two methods (BayeScan and Samβada). The red SNPs are those that have been identified by both methods. The green SNPs are unique to BayeScan.

larger deviation from Hardy-Weinberg equilibrium in the total population, than in the individual subpopulations (F_{IS}).

Allele frequencies, H_O and H_E were also calculated for individual accessions. This data revealed no significant differences between the observed and expected levels of heterozygosity, as determined by the Goodness of Fit test (Supplementary Table S2). However, each accession consists of a small number of plants (6-16), so these allele frequencies are not known with high accuracy.

Pairwise F_{ST} values (Table 2) showed that the Asian population was highly differentiated from the UK and Iberian populations, but only moderately differentiated from the Central European population, thus confirming the Asian population as most separated from the three other subpopulations. The UK and Iberian populations were moderately differentiated from each other and to the Central European population. The overall level of gene flow (Nm) was 3.04, and gene flow among accessions within each of the four groups were 1.43 (Asia), 2.13 (UK), 6 (Iberia) and 7.39 (Central Europe).

Analysis of molecular variance (AMOVA) was performed to assess how much of the total genetic variance could be attributed to different hierarchical levels of division of the total population. The results are summarised in Supplementary Table S3. The first level was among the four groups, Asia, UK, Iberia and Europe. The second level was among accessions within groups, and the third level was individuals within populations. It demonstrates that among group variation accounted for approximately 20% of the total variance, among accessions within groups accounted for 22.9%, and the largest contribution to the variance was within accessions (56.3%).

Two main methods were used to identify outlier markers in genome scans for selection. BayeScan identified 74 loci and Samβada identified 1,020 loci. They had 60 loci in common spread across the seven chromosomes (Fig. 3). Those loci were found in gene models that have been identified as, amongst others, transcription factors, which are key regulators of gene expression and may have a direct effect on the plants ability to respond to its environment. Others included "household" genes, and some involved in growth responses, flowering time and disease resistance (Supplementary Table S4). In Samβada the 1,020 SNP loci significantly correlated to the geography of the collection sites, after Bonferroni correction for both the Wald and G Score. The majority of the SNPs (85.6% of the 1,020 outliers) were found to be correlated to longitude to some extent (Supplementary Table S5). Regression analysis was performed using the principal components against the geographic co-ordinates. This analysis used all of the SNP data not just the outliers, and a highly significant correlation ($r = 0.93$) between PC1

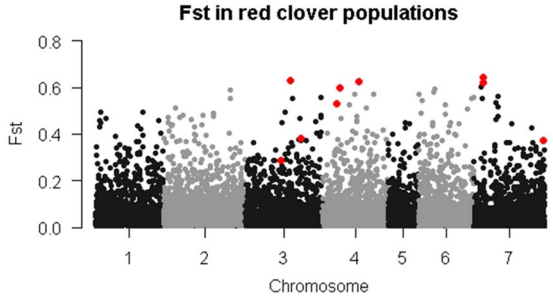

Figure 4. Manhattan plot of average F_{ST} values over the four subpopulations identified in the UPGMA analysis. The markers highlighted in green represent markers with high F_{ST} values and/or of interest because of their putative function. These markers are also highlighted in Supplementary Table S4.

Chromosome	No of markers	Average r^2	No of markers	Average r^2
	Varieties		Ecotypes	
1	978	0.026	1197	0.007
2	1212	0.029	1433	0.008
3	1163	0.025	1367	0.007
4	934	0.028	1150	0.008
5	427	0.032	536	0.008
6	820	0.027	963	0.008
7	1004	0.027	1246	0.008

Table 3. Summary of LD decay in the varieties and ecotypes. Data collected from LDMap data and Synbreed was used to obtain the average r^2.

and longitude of the collection sites was found (Supplementary Fig. S3). The other two components of latitude and altitude were not significantly correlated.

The 60 common loci represent five from intergenic regions, and 55 genic SNPs, of which 51 were assigned gene ontology (GO) terms. GO terms were collected from the genome browser (https://legumeinfo.org) and GO terms analysed in Revigo[16] to assign biological process, cellular component and molecular function. Over half of these GO terms were assigned to a molecular function (Supplementary Fig. S4). Supplementary Table S4 also highlights 8 markers from 7 genes, five of which have the highest F_{ST} values (Fig. 4). Two of those markers are located on Chromosome 7 in a gene model for a RING/FYVE/PHD type Zinc finger family protein, involved in plant development. The three others are models for a putative methionyl tRNA synthetase on Chromosome 4, an S-adenosyl methionine dependent methyltransferase and a phosphoglycerate kinase protein on Chromosome 3. Three other markers are highlighted because of their potentially interesting function with respect to plant breeding. On Chromosome 3 at position 12,438,192 there is a TIR-NBS-LRR type disease resistance gene, and at position 22,736,529 there is a basic leucine zipper transcription factor, which is a homologue of the flowering time gene *VEG2* in pea[17]. On Chromosome 7 at position 30,124,908 there is a β-D-glucan synthase like gene. Glycans are known to be involved in plant reproduction, especially male gametophyte development[18].

Linkage disequilibrium. The average r^2 for each chromosome was calculated. For the five varieties this value ranged from 0.026 – 0.032 per chromosome, however in the ecotypes this value was lower at 0.007–0.008 (Table 3).

We also compared the decay in LD between the four main groups, and between varieties and ecotypes according to the method of Wang *et al.*[19], with modifications described[15]. The model excludes values where r^2 is 0, so the absolute value of the rate of decay is therefore underestimated. In all the comparisons the rate of decay was proportional to population size (Supplementary Fig. S5). Thus, LD in the European population decayed most rapidly, and most slowly in the UK and Asian populations. When we sampled 120 genotypes randomly from the European population of 364, the rate of decay increased to the same level as that of the Asian population, which is of similar size.

Phenotypic assessment. During the experiment, there was considerable loss of plant material from the first full year to the second year (Supplementary Table S6). This was due to poor winter survival and the inability of many of the ecotypes to survive the cutting regime. The plants were given a lax cut at the end of each year. During the second year, two harvests were collected. The number of plants that survived into year three (2017) was little over 30%, so no further data were collected. This meant that phenotype measurements were collected in one full year only. However, two sets of measurements were collected. From the 514 plants that survived into year two, flowering times were recorded. Of these plants, 415 flowered within the time constraint and a further 99 had

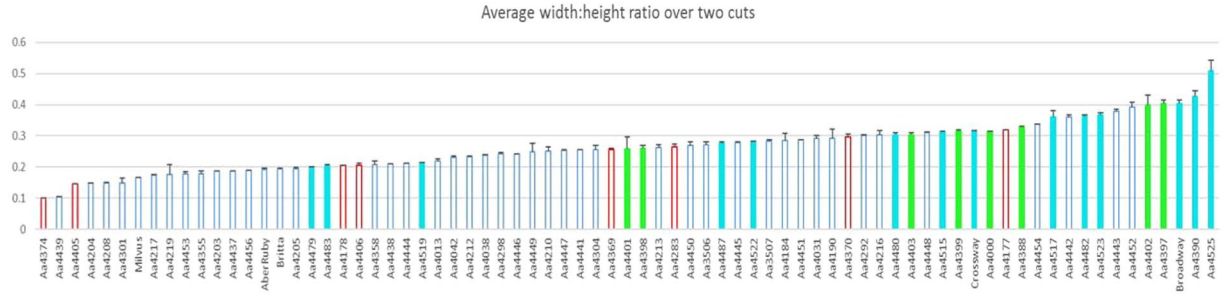

Figure 5. Average plant width to height ratio, with error of the ratio, of phenotypic data collected in year two, and averaged over two cuts. Measurements are in cm, and the graph is coloured in the same way as in the cluster and PC analyses. The varieties are labelled AberRuby, Britta, Grasslands Broadway, Crossway, and Milvus. Average plant heights ranged from 10–153 cm. Average plant widths ranged from 10–44 cm.

flowered by day 100 or later, the day of harvest. There was a large range of variation in flowering across the ecotype panel, with the average being at day 70. Three accessions (Britta, Aa4013 (Denmark) and Aa4038 (Finland)) did not flower before the cut off day. These were all Scandinavian in origin, and Britta was developed as a late flowering variety to maximise vegetative growth[20]. As the data were not normally distributed, it was $\log_{10}$ transformed, and subjected to a GWAS analysis using GAPIT (Supplementary Fig. S6). The analysis produced one significant SNP located at Chr3_22736556, which lies within gene 10558. This is a basic leucine zipper transcription factor. A BLAST[21] search showed it to be homologous to the *VEG2* gene in pea, which is involved in flowering time and inflorescence development[17]. This SNP was also in the list of loci identified as potentially under selection. An investigation into the effects of the alleles at this locus showed that the minor allele reduced the time to flowering by 11 days. The average flowering time for the homozygous major allele was 73.3 days, for heterozygotes it was 66.6 days and for the homozygous minor allele it was 62.4 days. The late flowering accessions Britta, Aa4013 (Denmark) and Aa4038 (Finland) were all homozygous for the major allele for this SNP.

In year two of the experiment two sets of measurements were taken for plant height, plant width, and stem number. Over the two collection dates, average plant height per accession ranged between 10–153 cm; the five tallest accessions were in descending order Aa4217 (Slovenia), Aa4444 (Italy), Milvus, Aa4203 (Sweden) and Britta. Average plant width per accession ranged between 10–44 cm; the five widest accessions were in descending order Grasslands Broadway, Aa4443 (Italy), Aa4451 (Italy), Aa4525 (Iberia), and Aa4442 (Italy). To estimate the degree of prostrateness in the population a width:height ratio was calculated. A high ratio described those accessions as being wider than tall, and more prostrate. The ratio ranged between 0.09 – 0.57. The five accessions with the highest ratio were in descending order Aa4525 (Iberia), Aa4402 (UK), Aa4390 (Iberia), Grasslands Broadway, and Aa4523 (Iberia) (Fig. 5, Supplementary Table S7).

The average number of stems per accession was recorded. The values ranged from 1 to 97. The six accessions with the most stems were in descending order Aa4397 (UK), Grasslands Broadway, Aa4399 (UK), Aa4403 (UK), Crossway, and Aa4525 (Iberia) (Supplementary Fig. S7). No significant marker-trait associations were found with the width:height ratio or the stem number.

Discussion

We present here a study of the genetic diversity and population structure in a collection of *Trifolium pratense* using GBS for the molecular marker data. *Ape*KI was selected as the restriction enzyme, as it is partially methylation sensitive and rarely cuts in retrotransposons. Therefore, *Ape*KI digestion products are fragments preferentially from low-copy genomic regions[14,22], and are more likely to be genic in origin. This could explain why most of the 60 SNPs potentially under selection were located in transcribed regions (Supplementary Table S4). Red clover is an allogamous plant, and this was reflected in the number of heterozygous biallelic SNPs identified. The 8,118 SNPs were predominantly transitions as opposed to transversions (Supplementary Table S1). This phenomenon has also been reported amongst others in chickpea[23]. The bias towards transitional SNPs is advantageous during natural selection as these SNPs are more likely to conserve protein structure than are transversion SNPs[24].

The results show a strong relationship between geography and accessions. The germplasm was a collection of material from ecotypes or natural populations from Europe and Asia and five varieties bred from germplasm originating in Europe. There were some anomalies in the genetic structure analysis, and there is some ambiguity as to how many groupings best explain the diversity. According to UPGMA clustering, the change in slope identified most prominently a four group structure, with a minor change in slope angle at two groups. This also was reflected in the PCA (Fig. 1). STRUCTURE identified two groups, but a more likely ancestral origin of nine groups (Supplementary Fig. S2). In all cases, the two group structure separated Asia from Europe; the four group structure differentiated Asia, UK, Iberia and Central Europe, at least partially. The position of the varieties reflected the geographic origin of the germplasm they were bred from. The nine groups identified using STRUCTURE resulted from a subdivision of the Iberian subpopulation into two subgroups, and the Asian subpopulation appeared to be an outgroup from the Central European subpopulation rather than a founder (Supplementary Fig. S2). This subpopulation had the largest heterozygosity deficit of the four, and the highest apparent inbreeding coefficient (Table 1). Given the geographical distribution of the Asian subpopulation in two separate regions, it could be a consequence of the Wahlund effect of "lumping" the two geographically separate groups together. If they do

not intercross, it would result in higher H_E than H_O. However, the STRUCTURE and the hierarchical analysis (Supplementary Fig. S2) did not reveal strong evidence of subdivision of the Asian subpopulation, despite the geographically distinct location of the two Iranian and the other Asian populations (Fig. 2).

The samples from the Iberian Peninsula consisted of a moderately unrelated set of genotypes, and the population as a whole was moderately differentiated from the UK and Central Europe (Table 2). At both the local and pan European level, these ecotypes are geographically separated, which may have reduced the potential for gene flow between the populations. However, the Nm value within this group was 6, which suggests some gene flow. The increase in genetic differentiation may be a result of some selection pressure. It is interesting to note that most of the more prostrate or spreading accessions in the panel were from Iberia or the UK (Fig. 5). The potential of prostrate, even stoloniferous red clover for increased grazing tolerance was discussed by Taylor[3]. Likewise, Pecetti[25,26] reports on the grazing tolerant nature of prostrate alfalfa. Whether grazing pressure has contributed to the prostrate nature of these Iberian accessions is speculative.

The majority of the natural populations, especially from Central Europe and Asia, studied here had an upright nature consistent with the varieties Milvus, Britta and AberRuby (Fig. 5). This growth habit, in natural populations, is likely to occur in habitats with low levels of disturbance[10]. It is also possible that the populations in Central Europe either contain varietal escapes or have undergone considerable gene flow as there was little differentiation between the varieties or the natural populations in either phenotype or genetic diversity (Figs. 1 and 5). This is consistent with the fact that gene flow among the accessions within the European group was highest at 7.39, the largest among the four groups.

The UK population showed a low overall genetic differentiation and a moderate one from Iberia and Central Europe. Island populations have lower genetic diversity when compared to continental populations[27]. This is most likely due to reduced genetic variation in the initial population. The PCA analysis (Fig. 1) and the width to height ratio of plants (Fig. 5) also indicate genetic and phenotypic similarities between the UK and the Iberian subpopulations. To what extent this is a consequence of germplasm exchange between the UK and Iberia, response to similar selection pressures in terms of climatic conditions is difficult to say. The range in geneflow estimates from 1.43 (Asia) to 7.39 (Central Europe) should be enough to prevent genetic drift having an effect[28]. The distribution of allele frequencies among all 75 accessions has a narrow range between 0.6 and 0.78 with a peak around 0.74–0.76. This is consistent with a relatively high Nm value[28]. The only outlier at 0.434 was one accession from the Asian population (Supplementary Table S2), which is also the only population with an inbreeding coefficient (F_{IS}) significantly higher than zero (Table 1). The majority of the Asian accessions are from a continental climate with cold winters and dry hot summers (Supplementary Table S8). The climatic and geographic isolation have probably both contributed to the relative genetic differentiation from the other populations (Table 2). The relative importance of climatic conditions and geographical distance in explaining genetic differentiation is difficult to disentangle. The longitudinal correlation is also associated with a climatic gradient ranging from a continental climate in Asia towards a temperate and humid Atlantic climate in Western Europe (Supplementary Table S8).

The overall genetic diversity of this red clover germplasm ($H_T = 0.275$) (Table 1) is comparable to that found in studies using RAPD ($H_T = 0.29$) and isoenzymes ($H_T = 0.29$) in South American, Swiss, and American cultivars respectively[29,30]. Higher values were found with AFLP in Italian Red clover ($H_T = 0.43$)[10], and allo-enzyme analysis in Caucasian natural populations ($H_T = 0.35$)[8]. Although the data indicated that there was no significant differences between the observed and expected levels of heterozygosity at the individual accession level, they were higher than the subpopulations in this study and H_O was higher than H_E (Supplementary Table 2). Such excess heterozygosity could be explained by small population size[31], and is consistent with outbreeding species and indicates a high level of genetic variation within each individual plant and accession, and a negative F_{IS}. Red clover has a one locus, gametophytic S-allele system of self-incompatibility, which prevents self-fertilization cross-fertilization by plants that have the same S-allele genotype. It is worth noting that within the Iberian accessions the two varieties Crossway and Broadway had the lowest H_O and the 1st and 3rd lowest H_E. Similarly within the European accessions the three varieties, Britta, Milvus and AberRuby had the 2nd, 3rd and 5th lowest H_O, and the 1st, 2nd and 5th lowest H_E, respectively (Supplementary Table S2). This indicates a slight narrowing of diversity, and selection pressure within the varieties. It should, however be noted that allele frequencies estimated from such small numbers of individuals (6-16) are less accurate.

Red clover is native to Europe, Western Asia and northwest Africa, and is adapted to many edaphic and climatic conditions[7]. The environment, including longitude, latitude and altitude, may result in physiological challenges that in turn may lead to plant morphological and molecular adaptations[32]. Samβada indicated a strong correlation between longitude and adaptive SNPs, and indeed a regression of the first principal coordinate identified the same correlation (Supplementary Fig. S3). These correlations also reflected the population structure as defined by cluster analysis in UPGMA and PCA. The preponderance of prostrate accessions in the Iberian and the UK populations (Fig. 5) meant that there was a weak correlation between longitude and the width:height ratio, but it was not significant (data not shown). Figure 3 shows an outline molecular map of the SNPs potentially under selection, as identified by both BayeScan and Samβada. Although the SNPs cover all seven chromosomes there would appear to be no significant clustering of SNP in any region, especially the 60 under selection in both analyses (red SNPs in Fig. 3). Similar results were reported for winter survivor populations of red clover in Scandinavia[33].

In addition to BayeScan and Samβada, we also applied a sliding windows Fst scan[34] to each individual accession. Only one region of interest was identified: In the two Iranian accessions Aa3506 and Aa3507 the same region on Chromosome 7 was identified as potentially under selection (Supplementary Fig. S8). Outlier SNPs of interest in this region included one at position 5,219,303. This SNP is within gene number 38,211, a 7 transmembrane MLO family protein that is involved in mildew resistance[35]. There was another 7 transmembrane MLO family gene very close by. Whether these genes are important for mildew resistance selection remains to be seen.

The AMOVA analysis demonstrated that within population variation (Asia, UK, Iberia, Central Europe grouping) explained most of the variance. This was also evident in the genetic diversity analysis, in which H_S was significantly higher than D_{ST}. The high heterogeneity and heterozygosity of red clover is expected as already explained by its self incompatibility[36]. Previous studies into the genetic diversity of red clover, using SSR and AFLP markers, have shown that the majority of the diversity is at the within-population level[37–40]. Both the genetic diversity and the AMOVA analyses are thus consistent with previous results.

Domestication and selective breeding would be expected to reduce allelic variation and genetic diversity[41]. Breeding populations are typically small, and by their nature selected for uniformity for traits such as flowering time. It should be borne in mind that red clover breeding is a recent occurrence[7], and is thus less likely to have had a major effect on allelic variation. LD is strongly dependent on recombination frequency and effective population size[42,43], so the difference in rate of LD decay seen with the varieties versus the natural populations (Table 3), could be explained by the difference in size of the two populations, rather than any reduction of diversity in the varieties. Supplementary Fig. S6 lends further support to this explanation, as the rate of decay is proportional to the size of the UK, Asian, Iberian and European populations. Obligate outcrossing gives rise to many effective recombination events which causes LD to decay rapidly[44]. The results presented here is consistent with that notion (Supplementary Fig. S5, Table 3). Other outbreeding crop species have also been reported to have low LD, for example in cauliflower[45] ($r^2 = 0.06$) and maize ($r^2 = 0.07$)[46]. The results are also in line with what was reported previously for red clover[15]. However, as the extent of LD is very low in the whole panel of accessions used here, interpretation of differences should be treated with caution.

Phenotypic data from all accessions were obtained from the first full season only, due to the high mortality rate in some accessions (Table S6). Mortality could be an interesting phenotype if it was due to lack of adaptation to the climate at IBERS (temperate without dry season, warm summer). However, many confounding factors could have played a role, such as soil and local rhizobia etc. Furthermore, since this work was aimed at identifying promising breeding material, mortality would limit their usefulness in such work. Flowering time is known to be a highly heritable trait, so the single year data are likely to be indicative of the ranking of genotypes. The SNP identified in this analysis was found with high homology to the *VEG2* gene in pea (3^{e-61}, 76% identity over 964 bases). The *VEG2* gene is an *FD* homolog that is essential for flowering and compound floral development[17], and as such is a good candidate for further study and verification in other populations. This gene was also identified as one of the candidates potentially under selection (Supplementary Table S4). The Samβada analysis showed that latitude rather than longitude was the main factor in its variation. This has also been found in other species with genes involved in flowering time responding to environmental cues[47]. It should be noted that although flowering in red clover is promoted by long days, it has no requirement for vernalization[48].

Vegetative growth analysis were the result of data from two time-points. They showed a variation in growth habit across the panel, especially in terms of width:height ratio (Fig. 5). The Portuguese accession and one of the UK accessions were also among the more prostrate ones (Supplementary Table S7). The prostrate varieties Grasslands Broadway and Crossway are also derived from Iberian populations[49]. They were developed as an alternative to the upright type varieties that dominate the market. Grasslands Broadway was defined as prostrate due to it having the largest width, but in comparison to the two most prostrate accessions included here (Aa4525 and Aa4402) it is twice as tall.

Conclusions

Based on over 8000 SNP markers, a panel of European and Asian red clover accessions from the IBERS Genetic Resources Unit was divided into four groups according to UPGMA and structure analysis, and the diversity was strongly correlated to longitude. The varieties included in the panel were not distinguishable from the ecotypes in terms of their genetic diversity, and there was no strong evidence for a bottleneck during the breeding process. However, within two Iranian accessions the same region on LG7 was clearly identified as being under selection. Two other methods identified 60 outlier loci indicating signs of selection, but no single chromosomal region was highlighted. The high mortality rate we observed in many of the natural populations is most likely a result of them being unimproved and un-adapted to the growing conditions at IBERS. Nevertheless, this study has shown that some contain novel allelic diversity that could be a source of new variation with potential use in breeding programmes.

Materials and methods

Plant material. A total of 75 accessions were used in this work. Of these, 70 were characterised as natural populations or ecotypes, originating from 16 countries in Europe, three from Asia and one from the Middle East (Fig. 2). A further five accessions were commercially available varieties. There were 640 plants in the panel; eight from each ecotype accession and 16 from each of the varieties. The clovers were planted as a spaced plant experiment at IBERS, Gogerddan (52.43° latitude, −4.02° longitude) in 2015 in a randomized block design with 8 replicate plants arranged in 2 blocks with rows of 4 plants per accession in each block.

The countries were allocated to their geographic regions according to The World Factbook and Eurovoc Table (Supplementary Table S8). The climate zones were based on the Köppen-Geiger classification[50]. The five varieties (Supplementary Table S9) were chosen for certain characteristics. AberRuby is an old IBERS variety. It has a lax growth habit, with few stems and its agronomic use has diminished. Crossway and Grasslands Broadway are early generation varieties developed at AgResearch in New Zealand from Portuguese and Spanish ecotype collections. The plants have a creeping growth habit and may under certain damp conditions produce runners that root from the nodes[49]. Milvus and Britta are tall MattenKlee type varieties that were developed in Switzerland and Sweden, respectively. Milvus is an early flowering variety, which has many high yielding stems and was bred for dry hay production. It is also reported to show resistance to *Sclerotinia trifoliorum* (crown rot)[51]. Britta is a late flowering variety[20] and has a degree of resistance to stem nematode[7].

DNA extraction and genotyping by sequencing. DNA was extracted from 100 mg of fresh young leaf tissue using a Qiagen DNeasy extraction kit in a 96 well format. The DNA concentration was measured using a Qubit™2.0, and normalized to 10 ng μl^{-1} with sterile TE buffer. The DNA was prepared for sequencing following the published GBS protocol[14], with modifications. *Ape*KI was used as the restriction enzyme, and bar coded adapters were annealed to each genotype. 16 of the annealed DNA samples were pooled at a time across the plate, and cleaned with magnetic beads. The concentration of the 16 pooled genotypes (6 per plate) was measured by Qubit™2.0, and 40 ng was used for PCR amplification. The PCR product was cleaned with magnetic beads, and the concentration measured by Qubit™2.0. The PCR amplification was repeated 3 more times. All of the PCR products were then mixed at equal concentration (10 ng μl^{-1}) to form the final library, which was analysed in a 96-plex format with 125 bp single end sequencing using an Illumina HiSeq. 2000 NGS platform.

SNP discovery. TASSEL5v2[52] and BWA[53] were used in conjunction with the red clover genome[15] in a reference based GBS pipeline to identify high quality SNPs within the populations. In the TASSEL5v2 pipeline, the quality of the SNP calling was set with the following parameters: 20 bp minimum and 64 bp maximum length; a Phred quality score of 20 was used ensuring a base call accuracy of 99%, or a 1:100 probability of incorrect base calling. All Phred scores in the sequences were at 30 (99.9%). There was a minimum cut off of 10 reads per SNP site, and the minor allele frequency was set to 0.05 to allow for heterozygote calling. For a SNP to pass into the final count, it had to be present in at least 80% of the genotypes. Finally, missingness was set to 0.1 per genotype. This ensured that genotypes with less than 10% of the SNP were excluded from the dataset. The data were analysed in R studio using a variety of packages[54]. Data re-coding and imputation of missing markers was implemented in the R package Synbreed[55]. The imputation was performed by random sampling from the allele distributions, with the minor allele frequency set to 0.05, and data missingness threshold to 0.2.

Data analysis. *Population structure and statistics.* Using all of the markers, the population was grouped into respective clusters by using UPGMA in the program Cluster[56]. This package was also used to produce a dendrogram with phylogenetic relationships (termed relationship tree in the text). A principal component analysis (PCA) was performed using R, and coloured according to the derived clusters from the above analysis. To further identify delimitations in the population structure and to infer ancestry, STRUCTURE v2.3[57] was used with an unbiased Bayesian approach with Markov chain Monte Carlo (MCMC) clustering of samples. The data were assessed for K values ranging from 1 to 15 with burn-in and MCMC iterations set to 20,000 each. For each value of K, three replications were made. STRUCTURE harvester web v0.6.94[58] was used to find the optimum K value for the population using the ΔK. This assigned the accessions to their genetic group. A total of 250 SNP were used per chromosome.

An analysis of molecular variance (AMOVA)[59] was performed on the data using the R package Pegas[60]. A hierarchical analysis was carried out with three levels: among the four groups identified by the UPGMA analysis, and among accessions within groups, and within accessions.

Once a population structure had been defined according to the UPGMA analysis, the populations were assessed for genetic diversity. The measurements included observed (H_O) and expected (H_E) heterozygosity, and the inbreeding coefficient over loci (F_{IS}). The genetic differentiation among populations was estimated from the fixation index (F_{ST}), the between groups diversity by F_{IT}, and the overall genetic diversity H_T. D_{ST} is the between population genetic diversity, and was obtained by subtracting the average of individual population H_E from the total genetic diversity, H_T. The parameters were obtained according to Nei[61]. The pairwise comparison of F_{ST} was obtained using the R programme StAMPP[62].

Outlier detection. To identify candidate loci potentially influenced by selection, three methods were used: F_{ST} genome scan[34], BayeScan v2.0[63] and Samβada v0.5.3[64–66]. To test the effect of environment (*i.e.* geographic origin) on genetic diversity, the geographical coordinates were used for each sampling site. Samβada is a spatial correlation method, which uses non-random associations between cause and effect. It uses Moran's I correlation[67] which assesses the overall clustering of the data, and this is related to the first law of geography "everything is related to everything else"[68]. Parameters were set to spatial for the autocorrelation, and nearest neighbours 20 for the weighting. Significant loci were identified after Wald and G tests following Bonferoni correction at a 99% confidence level. Significant outlier SNPs that were common to all three analyses were identified in the genome assembly, and gene models assessed using BLAST2GO[69] for biological and molecular processes, and cellular component, and to assign gene ontology (GO term).

Linkage disequilibrium. LD was calculated as the squared allele frequency (r^2) between each pair of SNP loci. Alleles were only considered if the minor allele frequency was above 0.05, as r^2 has large variances if rare alleles are considered[70]. Modelling of LD decay was performed using a custom R script derived from the method described in Wang *et al.*[19], and modified as described[71].

Phenotype measurements. Flowering time was recorded per plant from the initial start point of 1st April until 21st June in 2016 (day 1–81) and a further date recorded as day 100 (10th July 2016) as day of harvest. Measurements for plant height, width, number of stems and two harvest wet weights were also recorded. Any data that were not normally distributed were log transformed to improve homogeneity and analysed in Genome Association and Prediction Integrated Tool (GAPIT) for GWAS. The analysis was performed with the compressed mixed linear model[72] implemented in the GAPIT R package[73].

Data availability

The genotypic data have been deposited as raw sequence reads in the NCBI database under BioProject PRJEB30826 and phenotypic data are available upon request from the corresponding author.

Received: 29 January 2019; Accepted: 22 April 2020;
Published online: 20 May 2020

References

1. Kjærgaard, T. A Plant that Changed the World: The rise and fall of clover 1000–2000. *Landscape Research* **28**, 41–49, https://doi.org/10.1080/01426390306531 (2003).
2. Lüscher, A., Mueller-Harvey, I., Soussana, J. F., Rees, R. M. & Peyraud, J. L. Potential of legume-based grassland-livestock systems in Europe: a review. *Grass and Forage Science* **69**, 206–228, https://doi.org/10.1111/gfs.12124 (2014).
3. Taylor, N. L. A century of clover breeding developments in the United States. *Crop Science* **48**, 1–13, https://doi.org/10.2135/cropsci2007.08.0446 (2008).
4. Van Minnebruggen, A., Roldán-Ruiz, I., Van Bockstaele, E., Haesaert, G. & Cnops, G. The relationship between architectural characteristics and regrowth in Trifolium pratense (red clover). *Grass and Forage Science* **70**, 507-518, https://doi.org/10.1111/gfs.12138 (2014).
5. Vannier, J. The Cambrian explosion and the emergence of modern ecosystems. *Comptes Rendus Palevol* **8**, 133–154, https://doi.org/10.1016/j.crpv.2008.10.006 (2009).
6. Perez, J. E., Nirchio, M., Alfonsi, C. & Munoz, C. The biology of invasions: The genetic adaptation paradox. *Biological Invasions* **8**, 1115–1121, https://doi.org/10.1007/s10530-005-8281-0 (2006).
7. Taylor, N. L. & Quesenberry, K. H. *Red Clover Science*. (Kluwer Academic Publishers, 1996).
8. Mosjidis, J. A., Greene, S. L., Klingler, K. A. & Afonin, A. Isozyme Diversity in Wild Red Clover Populations from the Caucasus. *Crop Science* **44**, 665–670, https://doi.org/10.2135/cropsci2004.6650 (2004).
9. Collins, R. P. *et al*. Temporal changes in population genetic diversity and structure in red and white clover grown in three contrasting environments in northern Europe. *Annals of Botany* **110**, 1341–1350, https://doi.org/10.1093/aob/mcs058 (2012).
10. Pagnotta, M. A., Annicchiarico, P., Farina, A. & Proietti, S. Characterizing the molecular and morphophysiological diversity of Italian red clover. *Euphytica* **179**, 393–404, https://doi.org/10.1007/s10681-010-0333-6 (2011).
11. Dias, P. M. B., Julier, B., Sampoux, J.-P. & Dall'Agnol, M. Genetic diversity in red clover (Trifolium pratense L.) revealed by morphological and microsatellite (SSR) markers. *Euphytica* **160**, 189–205, https://doi.org/10.1007/s10681-007-9534-z (2008).
12. Dias, P. M. B., Pretz, V. F., Dall'Agnol, M., Schifino-Wittmann, M. T. & Zuanazzi, J. A. Analysis of genetic diversity in the core collection of red clover (*Trifolium pratense* L.) with isoenzyme and RAPD markers. Analysis of genetic diversity in the core collection of red clover (Trifolium pratense L.) with isoenzyme and RAPD markers. *Crop Breeding and Applied Biotechnology* **8**, 202–211 (2008).
13. Ahsyee, R. S. *et al*. Genetic diversity in red clover (Trifolium pratense L.) using SSR markers. *Genetika-Belgrade* **46**, 949–961, https://doi.org/10.2298/gensr1403949a (2014).
14. Elshire, R. J. *et al*. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *Plos One* **6**, https://doi.org/10.1371/journal.pone.0019379 (2011).
15. De Vega, J. J. *et al*. Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement. *Scientific Reports* **5**, https://doi.org/10.1038/srep17394 (2015).
16. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *Plos One* **6**, 9, https://doi.org/10.1371/journal.pone.0021800 (2011).
17. Sussmilch, F. C. *et al*. Pea VEGETATIVE2 Is an FD Homolog That Is Essential for Flowering and Compound Inflorescence Development. *Plant Cell* **27**, 1046–1060, https://doi.org/10.1105/tpc.115.136150 (2015).
18. Töller, A., Brownfield, L., Neu, C., Twell, D. & Schulze-Lefert, P. Dual function of Arabidopsis glucan synthase-like genes GSL8 and GSL10 in male gametophyte development and plant growth. *The Plant Journal* **54**, 911–923, https://doi.org/10.1111/j.1365-313X.2008.03462.x (2008).
19. Wang, L., Sorensen, P., Janss, L., Ostersen, T. & Edwards, D. Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. *BMC Genetics* **14**, 115, http://www.biomedcentral.com/1471-2156/14/115 (2013).
20. Lundin, P. & Jönsson, H. A. Weibull's Britta - a new medium late diploid red clover with a high resistance to clover rot. *Agriculture Hortique Genetica* **32**, 44–54 (1974).
21. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**, 3389–3402, https://doi.org/10.1093/nar/25.17.3389 (1997).
22. Sonah, H. *et al*. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *Plos One* **8**, https://doi.org/10.1371/journal.pone.0054603 (2013).
23. Kujur, A. *et al*. Employing genome-wide SNP discovery and genotyping strategy to extrapolate the natural allelic diversity and domestication patterns in chickpea. *Frontiers in Plant Science* **6**, https://doi.org/10.3389/fpls.2015.00162 (2015).
24. Wakeley, J. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution* **11**, 158–163, https://doi.org/10.1016/0169-5347(96)10009-4 (1996).
25. Pecetti, L., Romani, M., De Rosa, L. & Piano, E. Selection of grazing-tolerant lucerne cultivars. *Grass and Forage Science* **63**, 360–368, https://doi.org/10.1111/j.1365-2494.2008.00640.x (2008).
26. Pecetti, L., Annicchiarico, P., Battini, F. & Cappelli, S. Adaptation of forage legume species and cultivars under grazing in two extensive livestock systems in Italy. *European Journal of Agronomy* **30**, 199–204, https://doi.org/10.1016/j.eja.2008.10.001 (2009).
27. Crawford, D. J. *et al*. Allozyme diversity within and divergence among 4 species of Robinsonia (*Asteraceae, Senecioneae*), a genus endemic to the Juan Fernandez Islands, Chile. *American Journal of Botany* **79**, 962–966, https://doi.org/10.2307/2445008 (1992).
28. Hedrick, P. W. *Genetics of Populations*. (Jones and Bartlett Publishers, 2000).
29. Ulloa, O., Ortega, F. & Campos, H. Analysis of genetic diversity in red clover (Trifolium pratense L.) breeding populations as revealed by RAPD genetic markers. *Genome* **46**, 529–535, https://doi.org/10.1139/g03-030 (2003).
30. Yu, J., Mosjidis, J. A., Klingler, K. A. & Woods, F. M. Isozyme diversity in North American cultivated red clover. *Crop Science* **41**, 1625–1628, https://doi.org/10.2135/cropsci2001.4151625x (2001).
31. Balloux, F. Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution* **58**, 1891–1900, https://doi.org/10.1554/03-692 (2004).
32. Storz, J. F. Evolution. Genes for high altitudes. *Science* **329**, 40–41, https://doi.org/10.1111/j.1365-294X.2005.02437.x (2010).
33. Ergon, A., Skot, L., Saether, V. E. & Rognli, O. A. Allele Frequency Changes Provide Evidence for Selection and Identification of Candidate Loci for Survival in Red Clover (Trifolium pratense L.). *Frontiers in Plant Science* **10**, https://doi.org/10.3389/fpls.2019.00718 (2019).
34. Goncho, C. *Primer to Analysis of Genomic Data Using R*. (Springer International Publishing AG, 2015).
35. Kusch, S. & Panstruga, R. mlo-Based Resistance: An Apparently Universal "Weapon" to Defeat Powdery Mildew Disease. *Molecular Plant-Microbe Interactions®* **30**, 179–189, https://doi.org/10.1094/mpmi-12-16-0255-cr (2017).

36. Riday, H. & Krohn, A. L. Genetic map-based location of the red clover (Trifolium pratense L.) gametophytic self-incompatibility locus. *Theoretical and Applied Genetics* **121**, 761–767, https://doi.org/10.1007/s00122-010-1347-0 (2010).
37. Dugar, Y. N. & Popov, V. N. Genetic structure and diversity of Ukrainian red clover cultivars revealed by microsatellite markers. *Open Journal of Genetics* **3**, 235–242, https://doi.org/10.4236/ojgen.2013.34026 (2013).
38. Gupta, M., Sharma, V., Singh, S. K., Chahota, R. K. & Sharma, T. R. Analysis of genetic diversity and structure in a genebank collection of red clover (*Trifolium pratense* L.) using SSR markers. *Plant Genetic Resources*, 1–4. https://doi.org/10.1017/S1479262116000034. (2016).
39. Berzina, I., Zhuk, A., Veinberga, I., Rasha, I. & Rungis, D. D. Genetic fingerprinting of Latvian red clover (*Trifolium pratense* L.) varieties using simple sequence repeat (SSR) markers: comparisons over time and space. *Latvian. Journal of Agronomy* **11**, 28–32 (2008).
40. Annicchiarico, P. & Pagnotta, M. A. Agronomic value and adaptation across climatically contrasting environments of Italian red clover landraces and natural populations. *Grass and Forage Science* **67**, 597–605, https://doi.org/10.1111/j.1365-2494.2012.00887.x (2012).
41. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321, https://doi.org/10.1016/j.cell.2006.12.006 (2006).
42. Rogers, A. R. How population growth affects linkage disequilibrium. *Genetics* **197**, 1329–1341, https://doi.org/10.1534/genetics.114.166454 (2014).
43. Sved, J. A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**, 125–141, https://doi.org/10.1016/0040-5809(71)90011-6 (1971).
44. Flint-Garcia, S. A., Thornsberry, J. M. & Buckler, E. S. Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357–374 (2003).
45. Matschegewski, C. *et al*. Genetic variation of temperature-regulated curd induction in cauliflower: elucidation of floral transition by genome-wide association mapping and gene expression analysis. *Frontiers in Plant Science* **6**, https://doi.org/10.3389/fpls.2015.00720 (2015).
46. Hoyle, M., Hayter, K. & Cresswell, J. E. Effect of pollinator abundance on self-fertilization and gene flow: Application to GM canola. *Ecological Applications* **17**, 2123–2135, https://doi.org/10.1890/06-1972.1 (2007).
47. Stinchcombe, J. R. *et al*. A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4712–4717, https://doi.org/10.1073/pnas.0306401101 (2004).
48. Bowley, S. R., Taylor, N. L. & Dougherty, C. T. Photoperiodic response and heritability of the pre-flowering interval of two red clover (Trifolium pratense) populations. *Annals of Applied Biology* **111**, 455–461, https://doi.org/10.1111/j.1744-7348.1987.tb01474.x (1987).
49. Rumball, W., Keogh, R. G. & Miller, J. E. 'Crossway' and 'Grasslands Broadway' red clovers (Trifolium pratense L.). *New Zealand Journal of Agricultural Research* **46**, 57–59 (2003).
50. Beck, H. E. *et al*. Present and future Koppen-Geiger climate classification maps at 1-km resolution. *Sci Data* **5**, 180214, https://doi.org/10.1038/sdata.2018.214 (2018).
51. Boller, B., Tanner, P. & Schubinger, F. Merula und Pavo: neue, ausdauernde Mattenkleesorten. *AGRARForschung* **11**, 162–167 (2004).
52. Glaubitz, J. C. *et al*. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *Plos One* **9**, https://doi.org/10.1371/journal.pone.0090346 (2014).
53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
54. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2017).
55. Wimmer, V., Albrecht, T., Auinger, H. J. & Schon, C. C. synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* **28**, 2086–2087, https://doi.org/10.1093/bioinformatics/bts335 (2012).
56. Cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6. (2017).
57. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
58. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359–361 (2012).
59. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992).
60. Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–420, https://doi.org/10.1093/bioinformatics/btp696 (2010).
61. Nei, M. F-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.* **41**, 225–233 (1977).
62. Pembleton, L. W., Cogan, N. O. I. & Forster, J. W. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources* **13**, 946–952, https://doi.org/10.1111/1755-0998.12129 (2013).
63. Foll, M. & Gaggiotti, O. E. A genome scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180**, 977–993, https://doi.org/10.1534/genetics.108.092221 (2008).
64. Anselin, L. Local Indicators of Spatial Association—LISA. *Geographical Analysis* **27**, 93–115, https://doi.org/10.1111/j.1538-4632.1995.tb00338.x (1995).
65. Joost, S. *et al*. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology* **16**, 3955–3969, https://doi.org/10.1111/j.1365-294X.2007.03442.x (2007).
66. Stucki, S. *et al*. High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources* **17**, 1072–1089, https://doi.org/10.1111/1755-0998.12629 (2017).
67. Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
68. Tobler, W. A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46**, 234–240 (1970).
69. Conesa, A. *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676, https://doi.org/10.1093/bioinformatics/bti610 (2005).
70. Wen, W. W. *et al*. Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (Oryza sativa L.). *Theoretical and Applied Genetics* **119**, 459–470, https://doi.org/10.1007/s00122-009-1052-z (2009).
71. Grinberg, N. F. *et al*. Implementation of genomic prediction in Lolium perenne (L.) breeding populations. *Frontiers in Plant Science* **7**, https://doi.org/10.3389/fpls.2016.00133 (2016).
72. Zhang, Z. W. *et al*. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355–U118, https://doi.org/10.1038/ng.546 (2010).
73. Lipka, A. E. *et al*. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399, https://doi.org/10.1093/bioinformatics/bts444 (2012).

Acknowledgements
This work was supported by an Industrial Partnership Award from the Biotechnology and Biological Sciences Research Council (BBSRC) to IBERS (BB/L023563/1) and the Earlham Institute (BB/L022257/1). We are grateful for the support from our industrial partner Germinal Holdings Ltd. We wish to thank Jim Vale and his team for assistance with the field trial.

Author contributions
C.J. planned the experiment, performed the library preparations for sequencing, and the bioinformatics analysis, collected and analysed the phenotypic data, and drafted the paper. J.d.V. contributed to the conception of the project, performed some of the bioinformatics analysis, and contributed to the drafting of the manuscript. D.L. contributed to the conception and supervision of the project, the phenotypic data analysis and drafting of the manuscript. M.H. contributed to the conception of the project, supervised the next generation library production, and performed the sequencing runs and some of the bioinformatics analysis. S.A. and W.P. conceived and supervised the project. L.S. conceived and co-ordinated the project, contributed to the data analysis and interpretation, and drafted the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-64989-z.

Correspondence and requests for materials should be addressed to L.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.