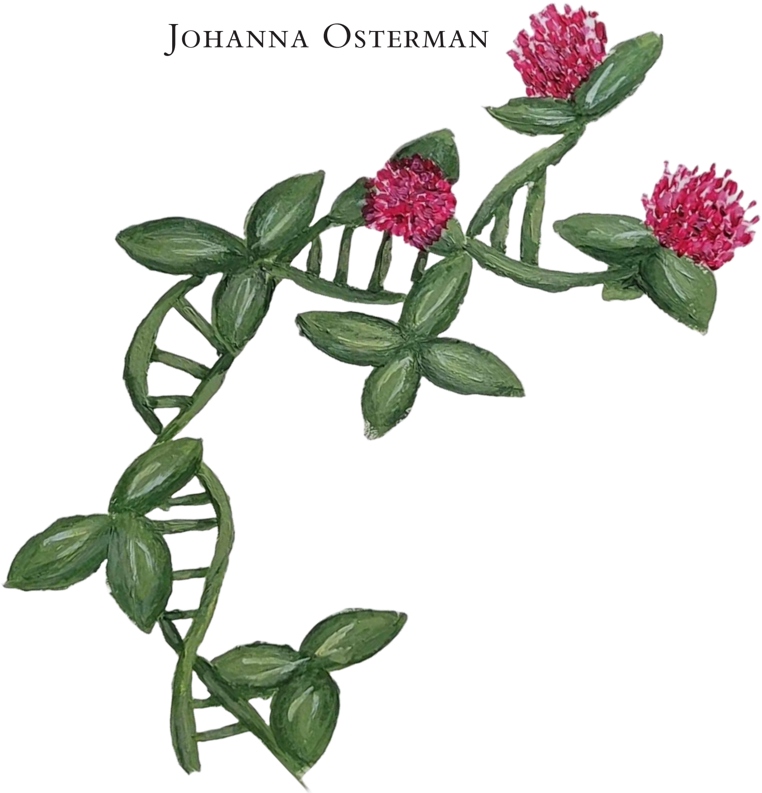
Acta Universitatis Agriculturae Sueciae

SLU

Doctoral Thesis No. 2024:33
Faculty of Landscape Architecture, Horticulture
and Crop Production Science

Advancing red clover breeding through genomic selection methods

Johanna Osterman



Doctoral Thesis No. 2024:33 • Advancing red clover breeding through genomic selection methods • Johanna Osterman

Advancing red clover breeding through genomic selection methods

Johanna Osterman

Faculty of Landscape Agriculture, Horticulture and Crop Production Science

Department of Plant breeding

Alnarp

SLU

SWEDISH UNIVERSITY
OF AGRICULTURAL
SCIENCES

DOCTORAL THESIS

Alnarp 2024

Acta Universitatis Agriculturae Sueciae
2024:33

Cover: Red clover with stems mirroring the DNA double helix. Painted by Johanna Osterman.

ISSN 1652-6880
ISBN (print version) 978-91-8046-330-0
ISBN (electronic version) 978-91-8046-331-7
https://doi.org/10.54612/a.70avv46bjm

Swedish University of Agricultural Sciences, Department of Plant Breeding, Alnarp, Sweden

Print: SLU Grafisk service, Alnarp 2024

Advancing red clover breeding through genomic selection methods

Abstract

Red clover is a major forage legume and a highly valuable crop for Northern Europe due to its high protein value and multiple ecological services. It is an important crop for both the ruminant industry and ecological farming. As growing conditions change due to rapid climate change, the demand for red clover breeding has increased. In this thesis, the potential for accelerating genetic gain through improved red clover breeding was studied.

First, since the response to selection is a function of genetic variation, the success of genomic selection depends on available genetic resources. Hence, red clover genetic resources available at the Nordic Genetic Resource Center (NordGen) gene bank and the Swedish seed company Lantmännen were used to evaluate the crop's genetic diversity and population structure. Red clover accessions currently used for breeding have low values for measures of inbreeding, which suggests a lower risk of inbreeding depression. However, their genetic diversity was low, relative to available wild populations and landraces, which can increase the risk of inbreeding depression. Hence, the progression of breeding could be limited by the gene pool. In this thesis, red clover populations with the potential to be used in red clover breeding to increase genetic diversity were identified.

Second, genetic gain in red clover can be rapidly increased by the introduction of genomic prediction models that minimize the need for time-consuming field trials. Both genome-wide association study (GWAS) and genomic prediction (GP) were tested for dry matter yield and forage quality traits based on data generated through multi-environment field trials and genotyping-by-sequencing. The results showed that dry matter yield and forage quality are affected by genes regulating responses to environmental inputs and stresses. This thesis showed that, by increasing genetic diversity and implementing GP in red clover breeding, genetic gain can be accelerated.

Keywords: Red clover, genetic diversity, population structure, GWAS, genomic prediction, genomic selection

Advancing red clover breeding through genomic selection methods

Abstract

Rödklöver är en foderbaljväxt som är viktig i norra Europa på grund av dess höga proteinhalt och många ekosystemtjänster. Rödklövern är betydelsefull för både de djur som lever av hö och ensilage samt för det ekologiska jordbruket. Som en effekt av klimatförändringarna ökar behovet av effektiv och snabb rödklöverförädling. I denna avhandling har potentialen av genomisk selektion för ökad avkastning och foderkvalitet i rödklöver studerats i två steg genom att undersöka den genetiska variationen och att utvärdera genomassociationstudier (GWAS) och genomisk prediktion (GP).

Eftersom förädlingsresponsen är beroende av genetisk variation utvärderades genetiska diversitet och populationsstruktur hos det material som är tillgängligt hos genbanken Nordiskt Genresurscenter (NordGen) samt hos växtförädlingsföretaget Lantmännen. Resultatet från studien visar att de rödklöveraccessioner som idag används för växtförädling har låga värden för inavel vilket minskar risken för inavelsdepression. Dock var den genetiska diversiteten i förädlingsmaterialet relativt låg i jämförelse med vild rödklöver och lantsorter. Således, kan effekten av förädlingsinsatser vara begränsad. Dock identifierades rödklöveraccessioner som kan vara möjliga kandidater bland lantraser vars inkludering kan höja den genetiska variationen.

Genom att prediktera utfallet av en korsning istället för att testa det i fällt förkortas tiden för en förädling cykel vilket ökar den genetiska vinsten över tid. Både GWAS och GP testades för att prediktera biomassans torrsubstans samt foderkvalitet. Resultaten visar att påverkan av genetiska effekter gentemot miljö på torrsubstans och foderkvalitet varierar över säsongen. Där de underliggande generna tyckts vara involverade i respons på miljö och stress. Resultaten av den här avhandlingen visar att rödklöverförädling kan effektiviseras genom att öka den genetiska variationen och implementera GP. Implementationen av dessa metoder kan hjälpa förädlingen att möta de utmaningarna från klimatförändringarna och det ökade ekonomiska trycket.

Dedication

To my parents, you've raised me to be confident, stubborn and curious
– the hallmarks of a good scientist.

To Jacob, without your love and support I'd be a mad scientist.

To Alice, a great listener and tough code reviewer – a top-notch co-scientist.

"Yeah buddy – lightweight!"
Ronnie Coleman

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Osterman, J., Hammenhag, C., Ortiz, R., and Geleta, M. (2021). Insights Into the Genetic Diversity of Nordic Red Clover (*Trifolium pratense*) Revealed by SeqSNP-Based Genic Markers. *Frontiers in Plant Science* 12, 2402. doi: 10.3389/fpls.2021.748750

II. Osterman, J., Hammenhag, C., Ortiz, R., and Geleta, M. (2022). Discovering candidate SNPs for resilience breeding of red clover. *Frontiers in Plant Science* 13. doi: 10.3389/fpls.2022.997860

III. Osterman, J., Gutierrez L., Öhlund L., Hammenhag, C., Ortiz, R., Parsons, D., and Geleta, M. Comparison of single-trait and multi-trait models for genomic prediction in red clover (manuscript)

IV. Osterman, J., Gutierrez L., Öhlund L., Hammenhag, C., Ortiz, R., Parsons, D., and Geleta, M. Comparing longitudinal multi-trait GWAS to single-trait GWAS for yield and forage quality in red clover (manuscript)

Papers I and II are reproduced with the permission of the publishers.

The contribution of Johanna Osterman to the papers included in this thesis was as follows:

I. Participated in the collection of the accessions and planted, cared for and sampled the plants for genotyping. Planed and performed the data analysis and wrote the manuscript together with the co-authors.

II. Participated in the collection of the accessions and planted, cared for and sampled the plants for genotyping. Planed and performed the data analysis and wrote the manuscript together with co-authors.

III. Contributed to the sampling of the populations for genotyping. Planed and performed the data analysis and wrote the manuscript together with co-authors.

IV. Contributed to the sampling of the populations for genotyping. Planed and performed the data analysis and wrote the manuscript together with co-authors.

List of figures

Abbreviations

BLUE	Best linear unbiased estimate
BLUP	Best linear unbiased prediction
GEBV	Genetically estimated breeding value
GxE	Genotype by environment interaction
GS	Genomic selection
GP	Genomic prediction
GWAS	Genome wide association study
HWE	Hardy-Weinberg equilibrium
LASSO	Least-absolute shrinkage and selection operator
LMM	Linear mixed model
MT	Multi-trait
NEL	Net energy lactation
NDF	Neutral detergent fiber
NJ	Neighbour-joining
QTL	Quantitative trait loci
SNP	Single nucleotide polymorphism
ST	Single-trait

1. Background

In Sweden, over a quarter of arable land is used for the production of green biomass from forages. The biomass is used as feed for ruminants in the form of silage or through grazing. Hence, forage crops have been highly important for Swedish agriculture for a long period of time. Forages are used as a mixture of legumes and grasses where red clover is a key forage legume.

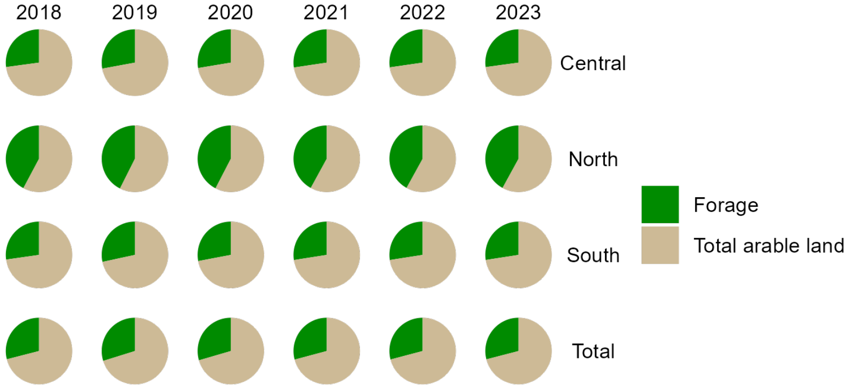


Figure 1 Pie charts showing the proportions of arable land covered by forages grown for biomass in the central, southern, and northern parts of Sweden as well as across the whole country from 2018 to 2023. The total arable land covers all agricultural land used for crops excluding forages grown for biomass. Data source: Jordbruksverket https://jordbruksverket.se/om-jordbruksverket/jordbruksverkets-officiella-statistik/statistikdatabasen.

1.1 Red clover and its cultivation history

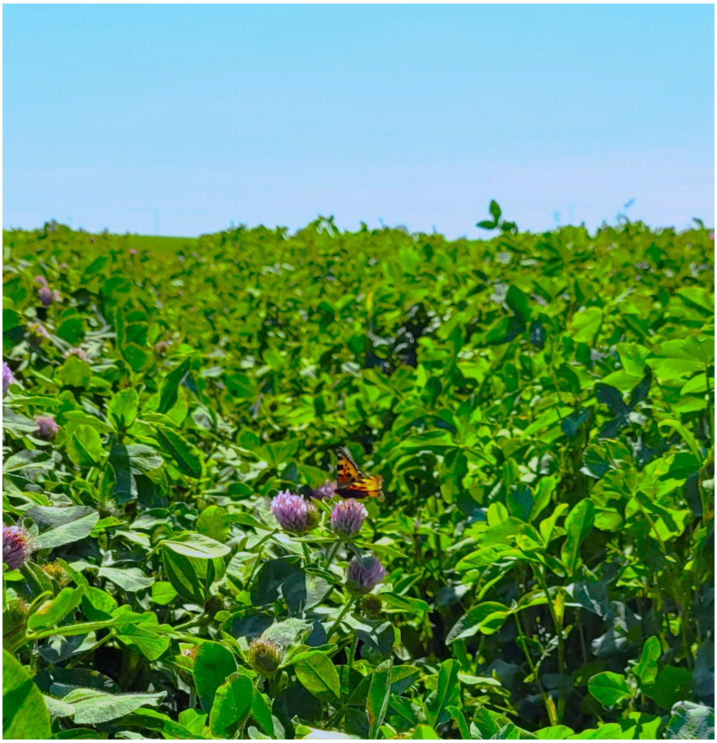


Figure 2 Picture from first year field trials in June 2021, taken by Johanna Osterman

Trifolium pratense L. (red clover) belongs to the largest genus in the Fabaceae family with about 255 species. It is one of the 16 *Trifolium* species that are actively cultivated for forage or pasture, additional native species occur in forage or pasture, which are not actively bred (Gillett *et al.*, 2001). Red clover is distinguished by its purple-red flowers (Figure 2), however not to be confused with the true red flowers of the crimson clover, and its taller statue compared to its close relatives, white clover and alsike clover. Wild red clover grows close to the ground, has longer stems and smaller leaves, and carries more flower heads than cultivated clover. Cultivated red clover, on the other hand, grows taller, and has more and larger leaves and fewer flower heads than wild red clover. This corresponds to the breeding goals of high green biomass yields with high protein content and easily digestible fiber, which are more abundant in leaves (Abd El Moneim, Khair and

Rihawi, 1990). Furthermore, the reduced number of flower heads is a consequence of breeding for increased green biomass yield, as the breeders-eye favors plant size over flower development. Records of red clover cultivation date back to the 13th century in Spain where red clover is first thought to be cultivated. A phylogenetic analysis of the *Trifolium* genus by Ellison *et al.,* (2006) based on nucleic and chloroplast DNA data supports the Mediterranean origin of the genus. The cultivated red clover then spread to other southern European countries with records from the 1560s and to the north with records from the 1770s (Merkenshlager, 1934). Red clover varieties are now cultivated in all temperate regions as a cover crop, in pastures or for silage (Riday, 2010). Red clover is not only cultivated for its properties as a high-quality forage, especially when grown as a cover crop, but contributes with several ecosystem services. Being a perennial crop, it covers the soil during the winter months, reducing soil erosion due to wind and rainfall (Glover *et al.*, 2010). The other benefit of perennial crops is the reduced need for plowing as it is re-sown less frequently than biannual and annual crops, which reduces soil compression and is consequently gentler to organisms living in the soil (Glover *et al.*, 2010). The added root mass of perennial crops increase the level of soil carbon, which benefits the microbiome of the soil and the other crops in the crop rotation (Glover *et al.*, 2010).

The most known ecological service of red clover is its role in increasing the bioavailable nitrogen in the soil through symbiosis with rhizobacteria (Thilakarathna *et al.*, 2017). Until the 1900s and the discovery of industrial fertilizers, red clover was a vital crop in the crop rotation as it decrease the usage of manure in the fields, which was considered quite gross (Kjærgaard, 2003). However, with the negative effects of long-term usage of industrial fertilizers on ecosystems, crop rotations including red clover and other legumes has been revitalized, especially in organic agriculture. The rhizobacteria that live in symbiosis with the red clover's root system fix atmospheric nitrogen (N_2) to bioavailable nitrogen in the soil. The bacteria infect the roots and, as a response, the plants form nodules in the roots where the bacteria grow (Sturz *et al.*, 1997). All legumes have the ability to form such symbiotic relationship, however, compared to other clover varieties red clover is the most efficient (Steinshamn and Thuen, 2008). The surge in bioavailable nitrogen due the symbiosis with rhizobacteria increases the protein content of the clover.

Red clover is a diploid species but in the 1970s, Swedish researchers developed a successful program for developing autotetraploids via chromosome duplication (Sjödin and Ellerström, 1986). Chromosome duplication is achieved by a chemical process that prevents the separation of chromosome pairs during meiosis, resulting in gametic cells with twice the number of chromosomes as normal gametic cells. Plants are then developed from these cells using tissue culture techniques. Therefore, as the additional chromosomes result from a duplication event within the same species, the tetraploid is an autopolyploid. The duplicated genome results in increased persistence and resilience (Öhberg, 2008). Additionally, tetraploid plants grow larger, hence, have higher biomass yield. However, their fertility is lower thus tetraploids have reduced seed yield compared to diploids (Amdahl *et al.*, 2016). This is due to differences in contributing traits, such as the number of flower heads per plant and number of seeds per flower head. The lower number of seeds per flower head in tetraploids is possibly due to lower pollen quality and a higher risk of embryo abortions during seed development (Vleugels, Roldán-Ruiz and Cnops, 2015).

1.2 Red clover genetics

Diploid red clover has 2n = 2x = 14 chromosomes and a genome size of approximately 420 Mb (Sato *et al.*, 2005) with a reference genome showing a highly heterozygous genome with low LD (De Vega *et al.*, 2015). Studies on elite cultivars (Campos-de Quiroz and Ortega-Klose, 2001), breeding populations (Ulloa, Ortega and Campos, 2003), wild populations and landraces (Herrmann *et al.*, 2005; Jones *et al.*, 2020) found high within-population genetic variance, and genetic differentiation found between red clover accessions was most often due to restricted gene flow due to natural barriers (Greene, Gritsenko and Vandemark, 2004; Nay *et al.*, 2023).

The high genetic diversity within populations is partly attributable to the self-incompatibility of red clover. Red clover is governed by a gametophytic self-incompatibility (GSI) system that is activated when pollens arrive on the stigma and the pollen tube growth begins. Pollen tubes grow rapidly but, as they approach the ovules, the growth of the tubes from incompatible pollens are slowed down to a halt (Taylor and Quesenberry, 1996). Hence, there is a genetic factor that determines which pollen is accepted by the ovule. The locus that determines compatibility between pollens and ovules is called S-locus. Red clover has a uniquely large number of self-incompatibility alleles, with 143 and 193 alleles reported for two commercial varieties, respectively (Williams and Williams, 1947). A very low level of self-pollination may occur in diploid red clover as a 0.10% of selfing was reported by Williams and Silow (1933) . A high level of population heterogeneity and self-incompatibility pose difficulties in plant breeding. Strict self-incompatibility prohibits the development of inbred lines and makes it difficult to fix favorable alleles in a population. Additionally, crossbreeding closely related genotypes leads to strong inbreeding depression, hence it is difficult to maintain stable breeding programs, especially for small markets. The high population heterogeneity in red clover makes genetic mapping studies and phenotypic measurements costly as many individuals needs to be sampled to represent a single population.

1.3 Conventional and modern plant breeding methods in red clover

A plants ability to self-fertilize is a key aspect in plant breeding schemes as it allows for the creation of lines where all individuals are genetically identical. Even in outcrossing crops self-fertilization is applied in hybrid breeding, where maize is an example of the significant increase in gains (Li *et al.*, 2022). Many have tried to implement inbreed strategies in red clover breeding with small success at high cost (Taylor, 1982). Any inbred lines achieved are difficult to keep, both in seed storage, due to the low amount of seed, or vegetative, due to the population's lack of regeneration. Another approach is using clonal breeding as in potatoes. While red clover can be cloned by propagation, the success is dependent on the specific root structure of the individual plant (Dawson and Street, 1959). Due to the many obstacles of either hybrid or clonal breeding red clover is bred by populations of full sibling (sib) or half sib families. Current red clover breeding programs are based on phenotypic recurrent selection where crosses are made both as bi-parental crosses (between two parents) or multi-parental crosses (between 4-8 parents), which yield full-sib or half-sib families, respectively. To accomplish this, potential parents are selected from nurseries, where individual genotypes from elite populations are planted and evaluated. Then, prospective parents are uprooted and transferred to glasshouse chambers where bees perform the crosses. The seeds (F_1) from each cross are then multiplied in a field, where F_2 seeds are produced. This is followed by a three-year field trial. Each F_2 population is evaluated based on breeding goals, such as improved biomass yield, forage quality, resistance, tolerance, and persistence traits. Given that such populations could be affected by inbreeding depression multiple F_2 families are combined in to synthetic populations (Busbice, 1969). Breeders select seeds from a number of populations, relative to the size of their gene pool, based on trait similarity or a specific trait goal. The synthetic populations are then evaluated in a field trial for another three years, and the best performing populations are selected as candidates and sent to Value for Cultivation and Use (VCU) and distinctness, uniformity and stability (DUS) trials.

The time between the initial cross of the parents and market listing of a cultivar or synthetic population today takes over 20 years (Jordbruksaktuellt, 2020). However, this time can potentially be reduced significantly with the implementation of genomic selection (GS) (Meuwissen, Hayes and Goddard, 2001). The major time-consuming aspect of red clover breeding is field evaluation across years, as it is a perennial crop. Hence, replicating red clover evaluation over time is challenging due to time constraints. Additionally, seed trait evaluation is often disregarded as it interferes with forage biomass yield evaluation. Introducing genomic selection has several advantages such as time reduction by predicting performance of offspring based on parents, to combining partially overlapping data sets thru genetic correlation for prediction of costly traits. Simulations have been done for forages and shown the possible gain (Simeão Resende, Casler and de Resende, 2014). The major points of implementing GS was based on number of markers and size of the training population.

2. Aim and objectives

2.1 General

The aim of this thesis is divided into two parts (i) a description of the genetic variation and population structure of red clover in Northern Europe, and (ii) evaluation of different genomic prediction models as well as finding Single Nucleotide Polymorphisms (SNP) markers of interest using marker-trait association models, which can be used to improve the current breeding scheme.

2.2 Specific

The specific aim of the population genetic analyses of red clover was to determine the levels and distribution of genetic variation within and among populations, and describe the variation based on the environmental profile of regions within Northern Europe, and subsequently find genes under selection based on target environments.

The goal of designing models for both genomic prediction and genome-wide association studies was to gain insight into the relationship between genomic markers and key traits related to yield and quality. This information can aid breeders in making a more efficient selection of parents to develop new varieties with improved forage yield and forage quality.

3. Population genetics

Population genetics describe genetic diversity and population structure using the distribution of genotypic and phenotypic frequencies of populations, and how they change over time due to processes like natural selection, genetic drift, mutation, and gene flow to. The markers can be isoforms of proteins or DNA markers. DNA markers are more common in the era of genome sequencing and each marker is associated with a locus in the genome. The effects of evolutionary forces, such as selection and migration can be estimated by observing the changes in allele frequencies at marker loci within and across populations. Genetic variation with individuals or sub-populations is measured relative to variation in their corresponding total population. The total population can span a given geographic region or contain germplasm from the whole world and is based on the specific research question. The research questions in this thesis revolve around red clover breeding and production in the Northern Europe. Hence, the germplasm used in this thesis was selected based on red clover genetic resources available in this region. This study aimed to determine the population structure and diversity of wild populations and compare them with landraces, old cultivars, and modern breeding populations. This is to gain a better understanding of how and the extent to which the genetic diversity of the red clover gene pool of northern Europe was utilized in breeding.

3.1 Genetic diversity and population structure

The genetic diversity of a species can be described using various parameters. F-statistics, also called fixation indices (F_{IT}, F_{IS}, and F_{ST}), are measures of genetic correlation of allelic states in the total population, sub-population and individual (Wright, 1949, 1965). The F_{IS} and F_{ST} were used to compare the different sup-populations. The F_{ST} value is a ratio between the variance of the sub population to the variance of the total population. Thus, at a value of zero the variance within a sub population is the same as the total variance and there is no population structure. Consequently, values closer to a F_{ST} of one indicates that there is a population structure as the variance within subpopulations is smaller than the variance of the total population (Wright, 1949, 1965). The F_{IS} measure levels of inbreeding by comparing the correlation of alleles between individuals and corresponding sub-population. If all alleles are homozygote in all loci of all individuals in the sub-population the correlation is one, on the other hand if the individuals are heterozygote so that the sub population has an excess of heterozygosity the correlation will be negative (Wright, 1949, 1965).

If the heterozygosity is in excess or not, can also be determined using the Hardy-Weinberg equilibrium (HWE), which is the frequency of alleles given the assumptions of random mating, no mutations, no selection, and no migrations. The HWE equations are

$$p^2 + 2pq + q^2 = 1$$

for diploids and

$$p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4 = 1$$

for tetraploids. Any significant deviation from this equilibrium indicates that one or more of the assumptions were not met, and may indicate inbreeding, self-incompatibility, or natural or artificial selection. Another interesting aspect of population genetics and population structure is to identify specific loci that deviate from Hardy-Weinberg equilibrium as they could be causal for trait differences that distinguish sub-populations. However, such deviation may also be the result of a genetic phenomenon where the change in allele frequencies at certain loci has more to do with the transfer of genetic material across generation than the natural selection of a fitness trait. This effect is called genetic draft where the allele hitchhike on the trait effect. Thus, any marker that significantly deviates from the HWE needs to be further studied based on its location relative to genes of interest. Another measurement of the effect of allele frequencies is Tajima's D, which describes the presence of rare alleles in populations. The presence of rare alleles means that there still is a potential of a response on selection. If all alleles were fixed, there would be no difference between individuals, and consequently no gain in crossing two parents. Tajima's D utilizes the pairwise comparison of genetic variation at a marker site to find alleles that are fixed in some populations but rare in others (Tajima, 1989). The idea is that if the sample size is small the correlation is high, and vice versa. As the sample size is the number of alleles the correlation indicates a rare allele if it is high and common allele if it is low. Hence, a breeder could check if candidate germplasm has desirable alleles that are fixed, but rare in a target breeding population. This could help make more informed decisions regarding the introgression of alleles from new material into a target population.

3.2 Pooled sequencing (pool-seq)

A major issue with the introduction of genomic selection in red clover breeding is the amount of genotyping required. As there is a large genetic variation within a red clover population, multiple individuals need to be genotyped and costs can get out of hand. However, Futschik and Schlötterer, (2010) showed that it's possible to pool genetic material from individual plants and sequence one population as a single sample for estimating population genetics. They calculate allele frequencies at each locus based on the number of reads per allele. Furthermore, they showed that using pool-seq allows more individuals to be sampled and sequenced at the same cost. Hence, the estimated allele frequency is closer to the true frequency compared with sampling and sequencing each individual separately. This technique makes it possible to perform genetic variation and population structure studies on a large number of populations and facilitates future GS in outcrossing species. The method have also been used in other fields of research such as mapping induced mutations (Schneeberger *et al.*, 2009), identifying local adaptation in *Arabidopsis lyrata* (Turner *et al.*, 2010) and identifying transposable elements in *Drosophila melanogaster* (Kofler, Betancourt and Schlötterer, 2012). However, one major drawback of pool-seq is the loss of information about loci that do not conform to the HWE, as the genotypes of individuals cannot be determined. Research on imputation methods for identifying individual genotypes in non-inbreed wheat show promising results (Clouard and Nettelblad, 2024). In this thesis the evaluation of pool-seq was based on the difference in information and resulting conclusions between individual and pooled results.

4. Genetic diversity and population structure of red clover in Northern Europe

The accessions used for assessing genetic diversity and population structure were obtained from the Nordic Genetic Resources Center (NordGen, Alnarp Skåne, Sweden) and the seed company Lantmännen (Svalöv, Skåne, southern Sweden). The accessions include cultivars, breeding populations, landraces and wild populations gathered from Denmark, Sweden, Norway and Finland. Lantmännen contributed with cultivars derived from F_2 populations and synthetic populations generated through their breeding program, as well as DLF and Graminor varieties available on the market. A total of 29 accessions were assessed at the individual genotype level. In addition, using pool-seq, 382 accessions were genotyped as pools with 10 individuals representing each accession. Both individual genotyping and pool-seq were performed using a targeted genotyping-by-sequencing (targeted GBS) method, SeqSNP, and resulted in 623 and 661 SNPs, respectively.

4.1 Genetic diversity in cultivated populations vs wild populations

The breeding material from Lantmännen showed high heterozygosity but with little population structure for both cultivars and synthetic populations (Figure 3). Synthetic populations are a mixture of F_2 families and its purpose is to buffer any inbreeding depression that could occur within F_2 families (Busbice, 1969). Inbreeding depression is the result of a homozygote loci for recessive deleterious alleles that affect the fitness of the plant negatively. The loss of fitness can also be due to decreased heterozygosity in loci which have a heterozygote advantage (Charlesworth and Charlesworth, 1999). The Lantmännen populations had an excess of heterozygotes, shown by the negative F_{IS} values, compared with the wild populations which had a mean value around zero (Figure 2). Thus, the risk of accessions suffering from inbreeding depression is minimalized. However, the low levels of F_{ST} values indicate low population differentiation, which can be a result of repeated crossing of closely related accessions. The response to selection is based on the genetic diversity, hence, low genetic diversity will limit the possible genetic gain. The loss of response on selection is most notable on traits with large additive effects, i.e. a linear relationship of the allele frequency and the expression of phenotype. However, for recessive alleles the response on selection is notable even under weak population structure (Whitlock, 2002). Even though synthetic populations have higher levels of heterozygotes they can still show symptoms of inbreeding depression due to the low population structure. In contrast, wild red clover had higher F_{ST} values and a mean F_{IS} value of around zero (Figure 3). This could be due to restricted gene flow between the populations possibly due to natural barriers in the environments where wild red clover populations grow. Thus, inbreeding depression could be more common within wild populations then cultivated varieties, however, crossing between to wild populations could have a higher response in

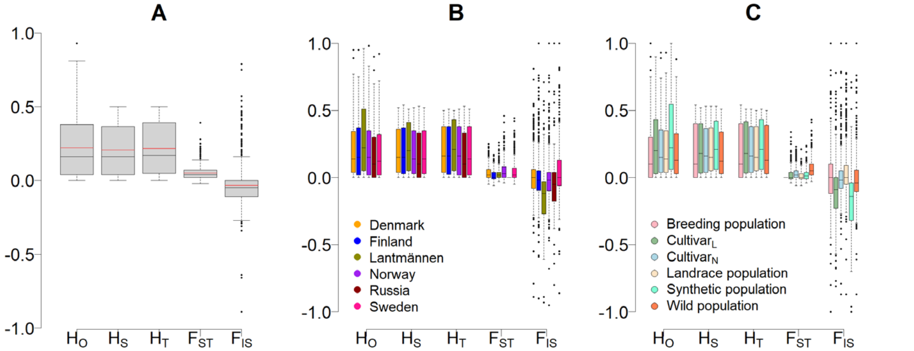


Figure 3 Boxplots visualizing the summary of different population genetics parameters for (A) all accessions, together with a red line marking the mean values, (B) each group of accessions grouped according to their origin, (C) each group of accessions grouped according to their type. The y-axis refers to the range of values for the parameters whereas the different parameters were given along the x-axis. H_O = observed heterozygosity; H_S = within population gene diversity; H_T = overall gene diversity; F_{ST} = population differentiation; F_{IS} = inbreeding coefficient.

additive traits than crossing between cultivated populations while the opposite would be the case for non-additive effects.

4.2 Sources of genetic variation that can contribute to increased genetic gain

One way of reducing the effect of inbreeding depression and increase the potential for genetic gain in a breeding population is to increase the genetic variance by introducing new germplasm. New germplasm additions can come from released cultivars or gene bank accessions such as wild populations, landraces or old cultivars. The cultivars from the NordGen gene bank included in this study showed high genetic similarities to the Lantmännen material. They had similar values for F_{ST} but higher F_{IS} values (Figure 3), hence, low population structure and relatively high inbreeding. In contrast, by comparing the Lantmännen populations with landraces and wild populations, which did not have significant inbreeding, higher genetic difference was observed between them (Figure 4). Thus, if a breeder would

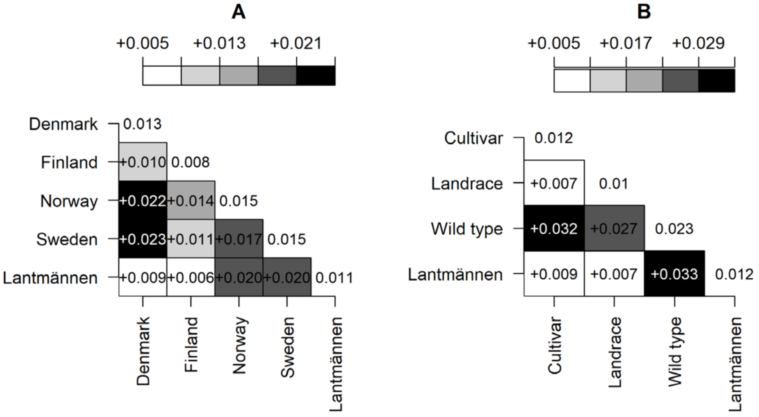


Figure 4 A graphical illustration of pairwise F_{ST} between groups of accessions representing different (A) country of origin: Sweden, Norway, Denmark, Finland and Lantmännen, and (B) population type: cultivar, landrace and wild populations from NordGen, and Lantmännen accessions. The gradient colour intensity corresponds to low to high F_{ST} values (the darker the colour the higher the F_{ST} value). All values are significant at a threshold value of 0.05. F_{ST} value of each pair is shown in the corresponding square. The diagonal values are mean F_{ST} values of each group. Lantmännen is represented as a separate group under country of origin.

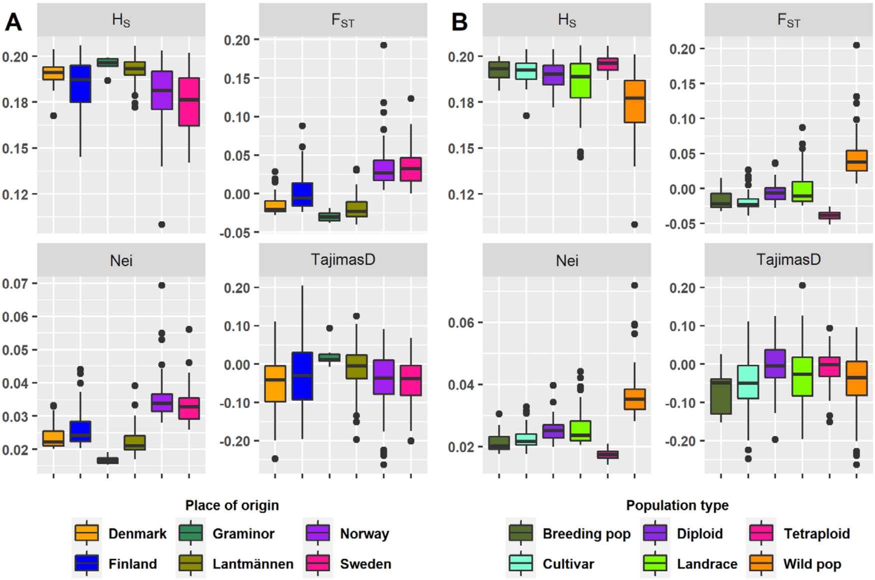


Figure 5 A box plot depicting the range and median for the genetic parameters on each group according to (A) Origin and (B) Type. The genetic parameters were H_S, mean expected heterozygosity; Nei, Nei's standard genetic distance; F_{ST}, mean fixation index; Tajima's D, Tajima's population genetic test statistic.

like to expand the genetic diversity of their breeding populations it would be recommended to explore the landraces and wild populations available at NordGen. However, the large genetic distance could entail high phenotypic differences with possible undesirable trait characteristics, hence, wild populations are not always desirable for introgression. Landraces from Finland exhibited genetic profiles that could constitute a good middle ground and should be considered first for further evaluation (Figure 4).

4.3 Estimating genetic diversity based on pool-seq data

In paper II, a pool-seq genotyping approach was used, utilizing the same target marker set as in paper I. Through the pool-seq mehtod information about observed heterozygosity is lost, and therefore the degree of inbreeding could not be estimated. In paper II, F_{ST}, Tajima's D as well as neighbour-joining (NJ) based clustering patterns of accessions were analyzed and discussed. The expected heterozygosity F_{ST} values were quite similar to those in paper I, hence, the use of pool-seq to estimate allele frequencies in

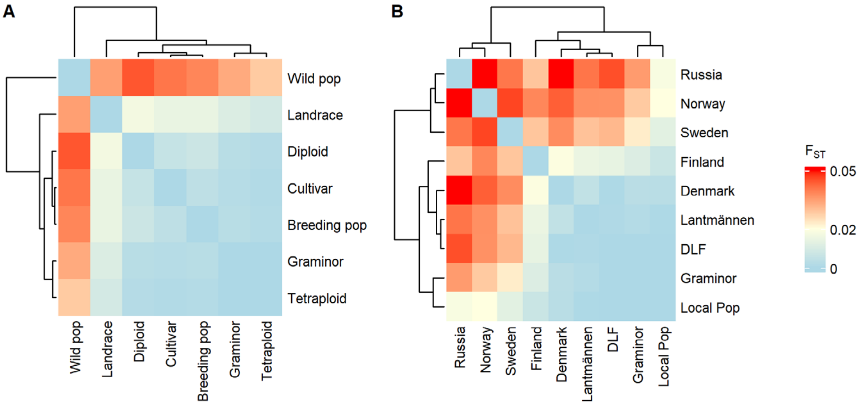


Figure 6 Heatmaps depicting the pairwise F_{ST} values between groups of red clover populations based on population type (A) and origin (B).

a population was shown to be a successful strategy. With the additional information from Tajima's D, it was possible to show that half of the Lantmännen accessions contained few rare alleles (Figure 5). Thus, can be considered to be in balancing selection. The same observation was made for the Graminor accessions, however the sample size was very small. The presence of rare alleles in the current breeding material indicates that there is potential for further genetic gain, given that those alleles carry genetic effects in line with breeding goals. The F_{ST} showed high similarities between NordGen cultivars and breeding populations, Graminor cultivars and Lantmännen cultivars (Figure 6). Thus, even though there were rare alleles within each group, the overall similarities were high. Consequently, introgression from these groups might not provide the desired effects on genetic variation and, as a result, genetic gain. However, as in paper I, landraces and wild populations had higher genetic diversity and could increase genetic variation in a gene pool for the breeding program.

4.4 Using LASSO models on allele frequency and bioclimatic variables to find candidate genes

As genetic variation is a key factor in a successful breeding program, the specific genes underlying the genetic variation is of high interest. Hence, the meta-data of each accession was compared using Nei's genetic distance-based NJ cluster analysis to see if estimated similarities and differences had any reflection on the characteristics of each accession (Figure 7). The meta-data includes maturity classes for Lantmännen accessions and geographic coordinates of collection sites for wild populations. The maturity classes are either late or middle-late, where the late types are more suitable for northern parts of Sweden with harsher winters and shorter growing seasons, while middle-late types are more suitable for milder climates and longer growing seasons. The idea is that the wild populations have, by natural selection, adapted to the climate where they grow, thus, any genetically similar cultivated accession ought to have the genetic makeup to grow in those environments. This study showed that cultivars included in this study were genetically more similar to wild populations from areas along the coastline or next to larger lakes than the inland wild populations, which were clustered separately. As red clover is also grown inland it is of interest to study the differences between inland wild populations and landraces compared to those grown close to large bodies of water. For example, if better winter hardiness is associated with smaller plants with low leaf biomass then that genetic variation will probably not benefit the breeding program. To find any differences in the genetic makeup that were not contributed by population structure a novel approach was tested. A machine learning method, least-absolute shrinkage and selection operator (LASSO) regression, was used to identify SNPs that could explain the differences in climatic variables across wild populations. The results were sparse but informative, where genes controlling the stomata showed predictive power in the models for annual precipitation and range of temperature during the year. Similarly genes coding for proteins with a kinas-binding function had good predictive performance in estimating annual mean temperature. These are proteins known to regulate responses to changes in temperature (Praat, De Smet and van Zanten, 2021). These results need to be further verified through controlled experiments such as classification of stomata count between accessions and identifying kinas-dependent pathways in response to temperature.

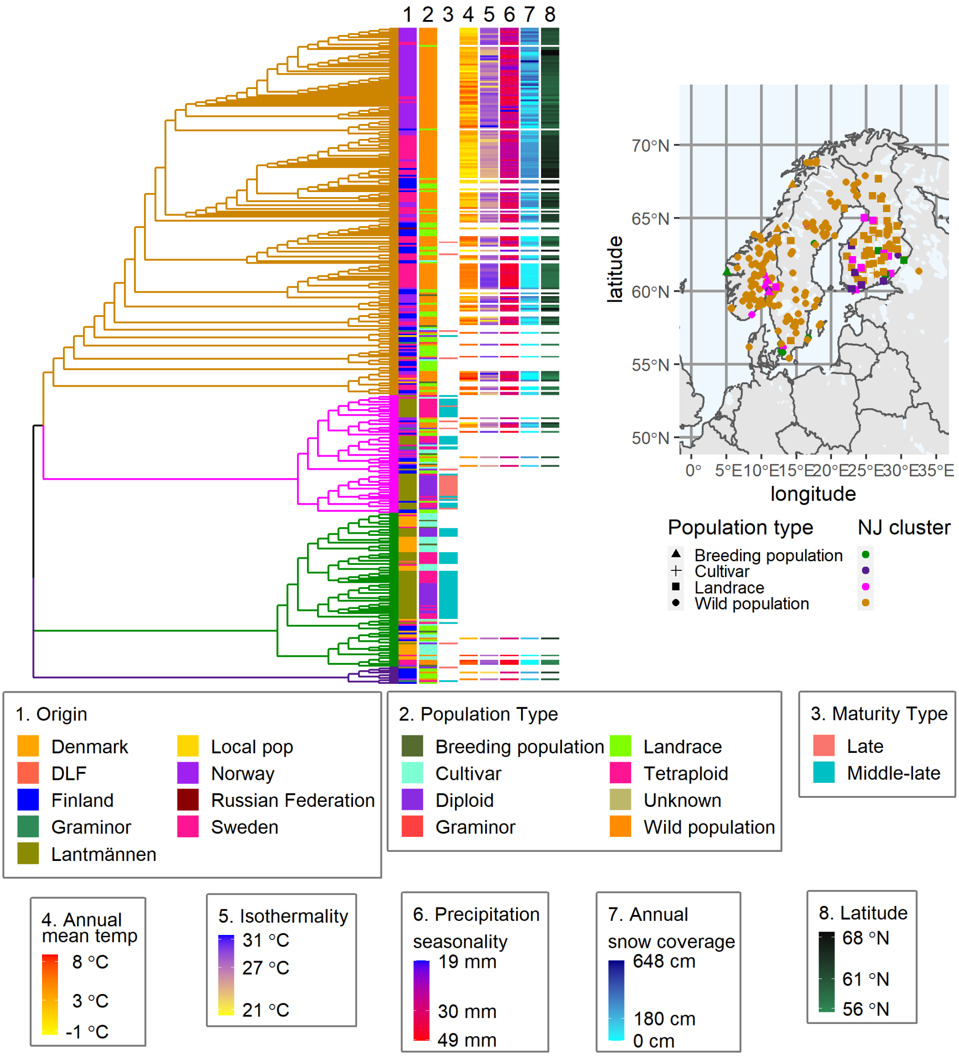


Figure 7 Nei's standard genetic distance-based neighbour-joining tree of 382 red clover populations showing four major clusters. Each numbered column (1-8) is linked to population descriptions or bioclimatic variables: 1-2, origin and type; 3 maturity groups according to their available records; 4-8, bioclimatic variables describing the collection sites of each wild population. The geographical map shows the collection site coordinates for breeding populations, cultivars, landraces, and wild populations obtained from NordGen.

5. Genomic selection

Genomic selection (GS) is an advanced plant breeding method that utilizes genome-wide genetic markers to predict the performance of untested individuals in terms of a particular trait. The effect of an individual's genotype on the population mean is called breeding value (BV). By estimating the BV through genetic markers, an estimated genetic breeding value (EGBV) is obtained. GS models can either be based on a few markers with large contributions to the phenotype or on many markers with small contributions spread throughout the genome. Markers associated with GEBVs could be identified through genome-wide association studies (GWAS). When multiple markers distributed across the genome are used to estimate GEBVs, genomic prediction is used (GP). The development of GS models requires a training population, a test population and a validation population. These can either be designated populations of breeding lines or varieties, where the training population contains all the breeding material and the test population is a subset of the material grown in a different year or location. The validation population is the offspring of the training population tested in the same environments. This is the preferred approach but it can be costly and time-consuming, hence, models can also be validated through cross-validation where a set of lines or varieties are masked and then predicted.

5.1 The fundamentals of genomic selection – summary of Falconer and Mackay (1996)

The idea of GS is to predict the performance of an individual using the performance of its relatives. By mapping the genetic effect on traits, the performance of untested populations can be predicted. The genetic effect is based on genetic variance, which is a term in the equation of phenotypic variance

$$V_P = V_G + V_E + V_{G\times E}$$

where the variance of the phenotype (V_P) is the sum of the variances of genotype (V_G), environment (V_E) and any interactions between genotype and environment ($V_{G\times E}$). Hence, the goal of genomic selection studies is to capture as much environmental variation as possible in the designed experiment, as it facilitates accurate estimation of V_G and $V_{G\times E}$.

For a trait to be a suitable target for genomic selection, it needs to have good heritability. The heritability (narrow-sense) is calculated as the ratio between the variance due to additive gene effects and phenotypic variance as

$$h^2 = \frac{V_A}{V_P}$$

Heritability (h^2) ranges from 0 to 1, where 0 indicates insignificant genetic effects on the phenotypic variation of a trait while 1 indicates the phenotypic variance is fully explained by additive effects. Heritability is a major factor in breeding, as trait inheritance levels play a major role in response (R) to selection

$$R = h^2 S$$

where S is the selection differential, i.e. the difference of mean phenotype of the parental generation and the offspring of the selected parents. If the following two conditions hold: only top-performing genotypes are selected as candidate parents and the phenotypic values are normally distributed and the selection differential can then be estimated as

$$S = i\sigma_P$$

where i is the selection intensity, i.e. the percentage of potential parents selected and σ_P is the square root of phenotypic variance

$$\sqrt{V_p} = \sqrt{\sigma_P^2} = \sigma_P$$

Since,

$$h = \frac{\sigma_A}{\sigma_P}$$

R can be written as

$$R = ih\sigma_A$$

Since the correlation (r^2) between a true value and an estimated value is h^2, the equation can also be rewritten as

$$R = ir\sigma_A$$

where r is selection accuracy. The response of selection over time (*t*) is a measurement of progress, where the genetic gain ΔG is the increase if beneficial genetic variance

$$\Delta G = \frac{ir\sigma_A}{t}$$

This equation is called the breeders equation and is fundamental to breeding. The faster a selection can be made without losing too much genetic variance the faster a breeding program can advance and the more genetic gain can be achieved in a given time. This increase of progress rate gives genomic selection an advantage over phenotype-based selection. Werner *et al.* (2023) showed that even at low accuracy, the gain was higher compared to conventional breeding due to shorter time needed per breeding cycle. These results were dependent on the levels of dominance effects and inbreeding. Hence, the change in time has the largest effect on gain, however, caution is advised in the presence of dominance effects.

5.2 The math and the solution for forages

There are different model frameworks for GP and GWAS, but the most widely used in genomic selection are linear mixed models (LMM). The LMM used in genomic selection is BLUP, Best Linear Unbiased Prediction, which introduces the ability to estimate fixed effects and breeding values simultaneously (Henderson, 1975). The name "BLUP" describes its most important features:

- Best – minimizes the prediction error, hence, maximizes the correlation between true and estimated breeding values.
- Linear – it's a linear model, thus models the relationship between predictors and observations as linear functions
- Unbiased – the estimation of breeding values and fixed effects are unbiased as each marker is assumed to contribute equally to the genetic variation.
- Prediction – predicts a breeding value based on the estimations of effects.

The following definitions are based on the theory explained in Mrode (2014).

The framework of the LMM model is

$$y = X\beta + Zu + \varepsilon$$

where $\boldsymbol{y}$ is a one column matrix of length n of phenotypic values where n is the number of records in the model, in this case populations. The LMM's system of equations is then solved using Henderson's mixed model equation as

$$\begin{vmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}g \end{vmatrix} \begin{vmatrix} \beta \\ u \end{vmatrix} = \begin{vmatrix} X'y \\ Z'y \end{vmatrix}$$

where **X** and **Z** are the index matrices above, $= \frac{\sigma_e^2}{\sigma_g^2}$, which is estimated using REML (restricted maximum likelihood), **G** is the matrix that relates records in the model. If it's a genetic relationship matrix, it's called a GBLUP and if it's a marker covariance matrix, it's called a SNP-BLUP. The estimate of the fixed effect is called Best Unbiased Linear Estimate (BLUE), and the estimate of the random effect is BLUP, which are given as

$$\beta = \left(X'^{V^{-1}}X\right) - X'V^{-1}y$$

and

$$u = \sigma_g^2 GZ'V^{-1}(y - X\beta)$$

where

$$V = \sigma_g^2 ZGZ' + \sigma_e^2 I$$

If the model has multiple levels of random or fixed effects, the values can either be estimated as an overall mean of the population across all levels of the effects. This is of interest in multivariate models that could include genotype-by-environment interactions (GxE) or multiple traits, where the interest could be an overall GEBV or the value at a specific environment or specific trait.

The genetic component in BLUP models is often defined by the genomic relationship matrix developed by VanRaden (2008). The matrix is a correlation matrix scaled by the allele frequency of the base population where the matrix diagonal is the genetic variance of an individual. If the variance is above 1 the individual can be considered to be inbred (VanRaden *et al.*, 2011). The off-diagonal elements are measures of covariance between individuals, which demonstrate their genetic similarities. However, in his paper, VanRaden stated that while the use of allele frequencies from the base population is theoretically appealing, the results from simulations showed that the counts of shared alleles were the most appropriate measurement of genomic relationship and inbreeding. Hence, the role of the genomic relationship matrix is to describe the similarities between individuals and populations, and within individuals. The math behind the G matrix depends on whether a locus is homozygous or heterozygous. This is not something that can be determined by estimations of allele frequency of a population that does not conform to HWE. However, the covariance between populations can be calculated. The G matrix have been adjusted by Ashraf et al. (2014)

to work on populations from multiple parents by observing the population like a polyploidy. Paper III and IV tired a different approach by using a Gaussian kernel. Then, more complex patterns between populations can be found. The Gaussian kernel is formulated as

$$k(Z_i, Z_j) = e^{\frac{-\|Z_i - Z_j\|^2}{\rho}}$$

where

$$\rho < 0$$

As Z_i and Z_j are allele frequencies for two populations, the Gaussian kernel calculates their relationship using a scaling parameter ρ. Thus, instead of using the G matrix, which is a linear kernel scaled by the allele frequency, we employed a more black box approach by estimating ρ so that the relationship between populations can be described. Although the same depth of information as that of the G matrix could not be obtained using this approach, we could adequately describe the relationship between populations.

When performing a GWAS, the allele dosage of a SNP is set as the fixed effect; however, there is no mathematical requirement for allele dosages. Allele dosage is the number of minor alleles present, say if the rare allele is coded as a and common allele codes as A, the AA genotype would have a value of 0, consequently, Aa = 1 and aa = 2, for a diploid species. These values are used for the interpretation of the results as the increase in Estimated Breeding Value (EBV) is viewed as a linear function of the allele dosage. This will capture the additive effect of a quantitative trait loci (QTL) where any deviation from the linear function can be viewed as the dominance effect (Varona *et al.*, 2018). The difference from using allele frequencies will be that the predictor will be continuous rather than discrete. As for interpreting dominance effects, the effects could either be viewed as deviations of the mean or a non-linear effect of allele frequency on EBV. This was never fully studied in this thesis, but the results and conclusions lay the foundation for further studies.

Thus, when using additive models the path to genomic selection in red clover is straight forward as it models linear relationships between allele frequencies and genotypic values.

6. Genomic prediction and GWAS targeting forage yield and quality

There are multiple approaches to develop models for both GP (paper III) and GWAS (paper IV). The objective is to use spatial data to correct for within-trial variation, and location parameters to adjust for environment and environment-by-genotype interactions (G×E). The genotype effect is assessed by using the genetic relationship matrix as a variance-covariance matrix to estimate the similarity between observations. This can be done in one or two steps. The two-step approach first involves correcting environmental effects using a model in which genotypes are treated as fixed effects. In the second step, the BLUEs are used as a response with the genotypes as random effects, and if an MT model is used, the covariance between measurements is modelled at this step. As genotyping of some accessions was unsuccessful, the two-step approach was used in both papers III and IV. The model design was chosen based on the corresponding Bayesian information criterion (BIC) and Akaike information criterion (AIC) values.

The prediction models were trained on red clover accessions from both Lantmännen and NordGen. In total, 532 accessions were used for phenotyping and genotyping. The majority (488) were Lantmännen's F_2-based cultivars and synthetic populations while the remaining 44 accessions were cultivars, landraces and wild populations from NordGen. Of the 532 accessions, 178 were tetraploids and 354 were diploids. The diploids were further divided into two groups based on maturity, either late or middle-late. The genotyping was performed by genotype-by-sequencing (GBS), which resulted in 8 107 SNPs for the diploids and 13 544 for the tetraploids.

6.1 The target traits

The traits used as response to evaluate the models were dry matter yield and three forage quality traits. The quality of forage is measured by the crude protein (will be called protein), fiber content in form of neutral detergent fiber (NDF) and the net yield of lactation (NEL). All these traits were measured on the dried harvests from the field traits using VDLUFA (Association of German Agricultural Analytic and Research Institutes, https://www.vdlufa.eu/). Protein content, NDF and NEL were selected based on their importance in red clover breeding. As dry matter yield is the first yield of forage, milk yield can arguably be the second yield. Milk yield is dependent on multiple parameters, where protein is important for animal health, NDF is a measure of digestibility and NEL is the amount of absorbed energy used for milk production.

6.2 Linkage disequilibrium analysis

Quantitative trait loci (QTL) are genomic regions that have a genetic effect on quantitative traits. Identification of their precise locations in a genome is complex and labor demanding. Molecular markers, such as SNPs and their association with a trait are used to identify and measure the effect of a QTL without having to map its precise location. Mapping populations phenotyped for various traits can be genotyped with a large number of SNP markers, of which only a few could be associated with a particular trait variation. The key to marker-trait association is the linkage disequilibrium (LD) between a marker and QTL and its stability across generations. The LD is measured as the correlation between the frequencies of two loci across all populations. It is an effect of several factors such as rate of genetic recombination, selection, genetic drift and mutation. LD often decays more rapidly in outcrossing species than in selfing species due to the effect of crossing between two distinct genomes. Hence, the genetic distance between markers and causal QTL needs to be shorter in an outcrossing crop (Charlesworth and Charlesworth, 1979). When De Vega et al. (2015) published the red clover reference genome, they reported a mean LD that ranged between 0.15 to 0.25 across the seven chromosomes. In paper III, the mean LD at 100 kbp were 0.012 in late diploids, 0.011 in middle-late diploids, and 0.064 in tetraploids. The values of paper IV are more in line with what Zanotto et al., (2023) reported. Thus, measurements of LD on multiple populations show more rapid decay.

6.3 The idea of multi-trait models for genomic prediction

The success of a prediction model is measured by its ability to describe all useful genetic variation that affect the biology of a trait of interest. One way to increase genetic variation described by a model is to include multiple responses, e.g. multiple traits. The key idea is that the model can use the correlation between traits to make better predictions about new observations (Henderson, 1984). Selecting multivariate models needs to be done based on the trait biology to capture either genetic correlation patterns or non-linear trends between traits. The target traits were genetically correlated with one another as well as across cuts (Figure 8). In this thesis, multiple approaches were tested and the outcomes showed that a factor analytical model and a heterogeneous variance model were the most effective approaches.

A factor analytical model (FA) is often used when modeling GxE due to its ability to summarize complex variance structures (Meyer, 2009). The FA model summarizes the data points to a reduced number of dimensions based on the explained variance (Gollob, 1968). In this study, this approach was applied for the longitudinal models where the variance between multiple cuts was reduced to two dimensions. A heterogeneous variance model utilizes an $n \times n$ matrix where n is the number of traits observed, the diagonal of the

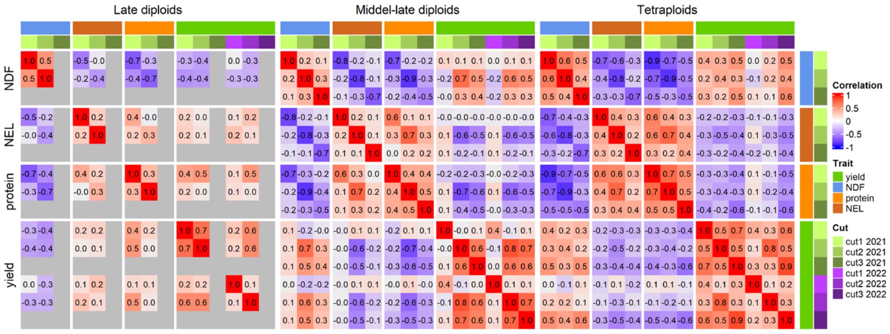


Figure 8 Heatmap showing the correlations within and between traits for each cut for the three subsets sorted by trait and cut. The columns and rows are annotated based on which trait and which cut the correlation belongs to.

matrix is variance of the n^{th} trait, and the off-diagonal is covariance between trait n and $n+1$. The advantage of a heterogeneous variance model over a homogeneous variance model is that each trait has its own variance on the diagonal. Thus, the estimate of each trait is better when using a heterogeneous variance model although it requires more data to fit.

6.4 The dependence of multi-trait model performance on genetic correlation

The success of the multivariate models, when applied to two-traits (MT_1), four-traits (MT_2), and multiple cuts (ST_L), was based on the correlation between the traits or cuts. Per example, the accuracy of predicting yield in middle-late diploids where the correlation of the first cut to the others is

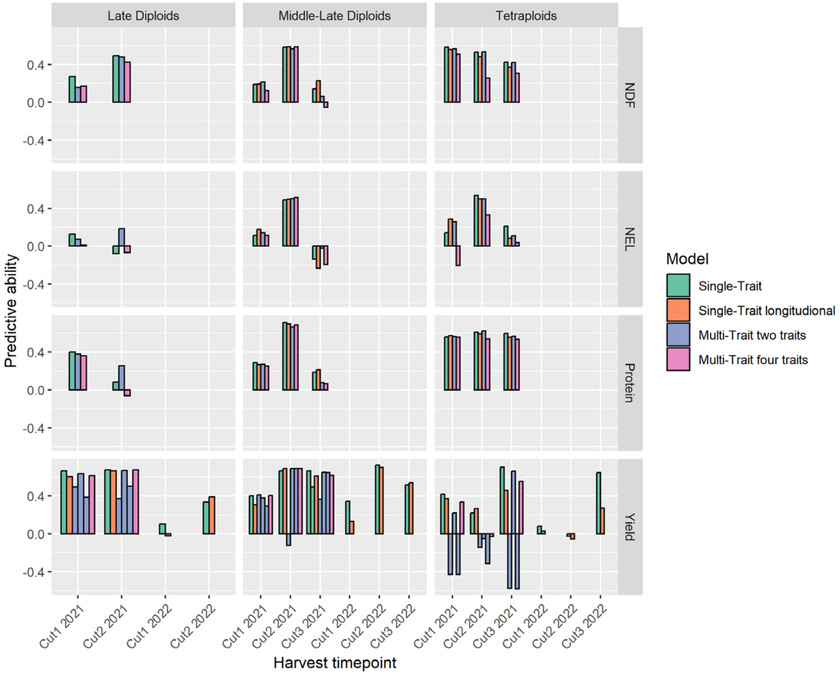


Figure 9 A bar plot of the predictive ability between the BLUEs and the mean BLUP of the 100 iterations of the 10-fold cross validation shown for each trait (row) within each subset (column) for each cut. The bars are colored based on the model used to calculate the BLUPs. Three bars are present for the predictive ability for yield by the multi-trait two trait model, they are yield estimated with net energy lactation (NEL), protein or neutral detergent fiber (NDF).

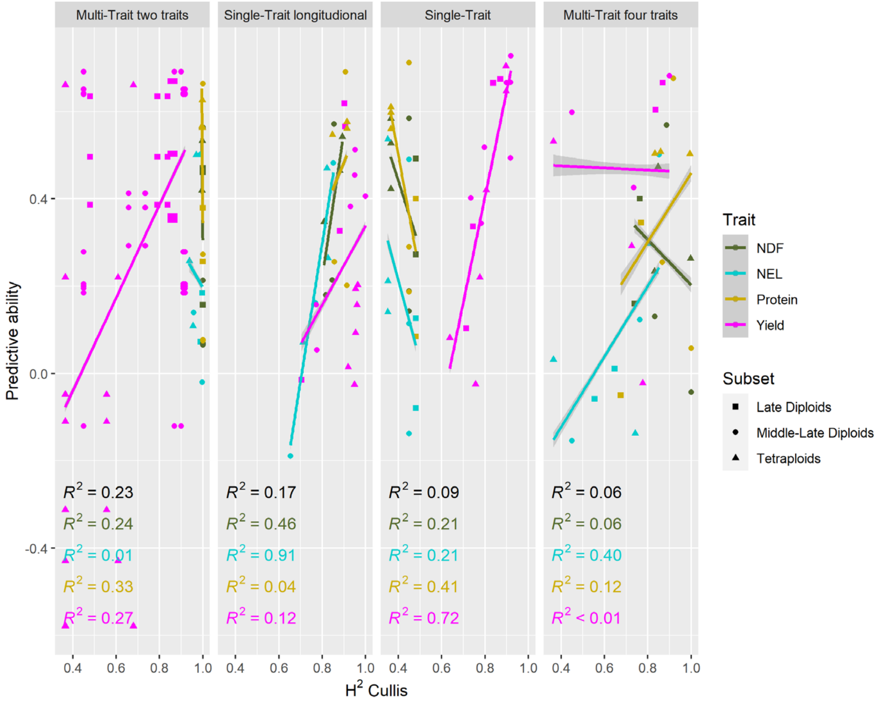


Figure 10 The predictive ability as a function of H^2 separated by model approach. Each point is the result of a model on a specific trait in a specific subset. Each point has a shape according to which subset it belongs and colored by trait. The R^2 value is calculated from a linear regression of Predictive ability on H^2.

between -0.1 and 0.4 and the single-trait (ST) model outperformed the ST_L model (Figure 9). However, the ST_L models outperformed the ST models when predicting second and third cuts (Figure 9).

The negative correlation of multiple models indicates that accessions have a predicted value higher than the observed value (Figure 9). This can be a result of the BLUPs shrinkage on accessions performing below the population mean increasing the estimated value towards the mean. The effect of how the predictive ability of the model is calculated across the 10-fold cross-validation can bias the predictive ability towards negative values as heritability approaches zero (Zhou et al., 2017). The assumption would be that the predictive ability has a positive linear relation to H^2. Hence, the better the genetic variance is described the better the performance of the model.

However, for some traits this study showed that the relation of H^2 on predictive ability was negative (Figure 10). Further examination show that the failure of the positive linear relation was dependent on the subset, where the tetraploids had overestimated accuracy based on H^2. One possible explanation is overestimation of additive variance due to non-additive effects. Thus, the prediction accuracy could be biased towards the non-additive genetic variance of the specific data. By introducing multiple traits or a longitudinal approach the accuracy was in line with expectations due to the genetic correlation between traits. This genetic correlation could decipher the probable non-additive effects. However, more analysis is needed confirm these hypothesis.

When implementing GP in a breeding pipeline, the model of choice need to perform well and be stable across traits, environments and time. This study found that while the predictive ability of the ST models was highest there could be non-additive genetic effects that make the model unreliable. This added reliability would be the biggest benefit of using MT models. As previously reported MT models for other crops rarely report large increases in predictive ability, the largest gain in MT models would be reliability over predictability. Semagn et al. (2022) who tested single-trait and multi-trait models for seven traits in spring wheat found that there was no significant improvement in predictive ability of MT models compared to ST models. However, they stated that by only observing one trait, an MT model could be applied to predict an unobserved trait. Additionally, Cuevas et al. (2023) indicated that the true advantage of MT models was for predicting unobserved traits based on the observation of correlated traits when predicting tuber weight, starch content, and sugar content in potatoes. Hence, the models developed in this thesis should be tested for predicting unobserved quality traits based on yields, as measuring quality traits is both expensive and labor-demanding.

6.5 Genome-wide association analysis across time

The GWAS study aimed to compare a univariate approach to a multivariate. The univariate approach (single-trait, ST) was to identify marker-trait associations separately for each trait at each cut. The multivariate approach (multi-trait, MT) was to expand the model at each trait to include all cuts, i.e.

a longitudinal model. The ST_L model applied in the GP study was used as a base for the MT GWAS. The major obstacle for implementing GWAS over time could be the same reason for the relation between H^2 and predictive ability.

The introduction of multiple cuts overestimated the SNP effect and especially the interaction effect between SNP and cut. Thus, the risk of false positives is great and need a more stringent threshold of acceptance. Non-additive effects that the model failed to capture could be the reason for overestimated SNP effects. This could be seen when plotting how the changes in allele frequency affected the genotypic values (figure 10). By modelling non-linear interaction effects between SNP and cut these non-additive effects can be captured and this would decrease the risk of detecting false positives. However, more data is needed as different polynomial patterns were observed across multiple SNPs (Figure 11).

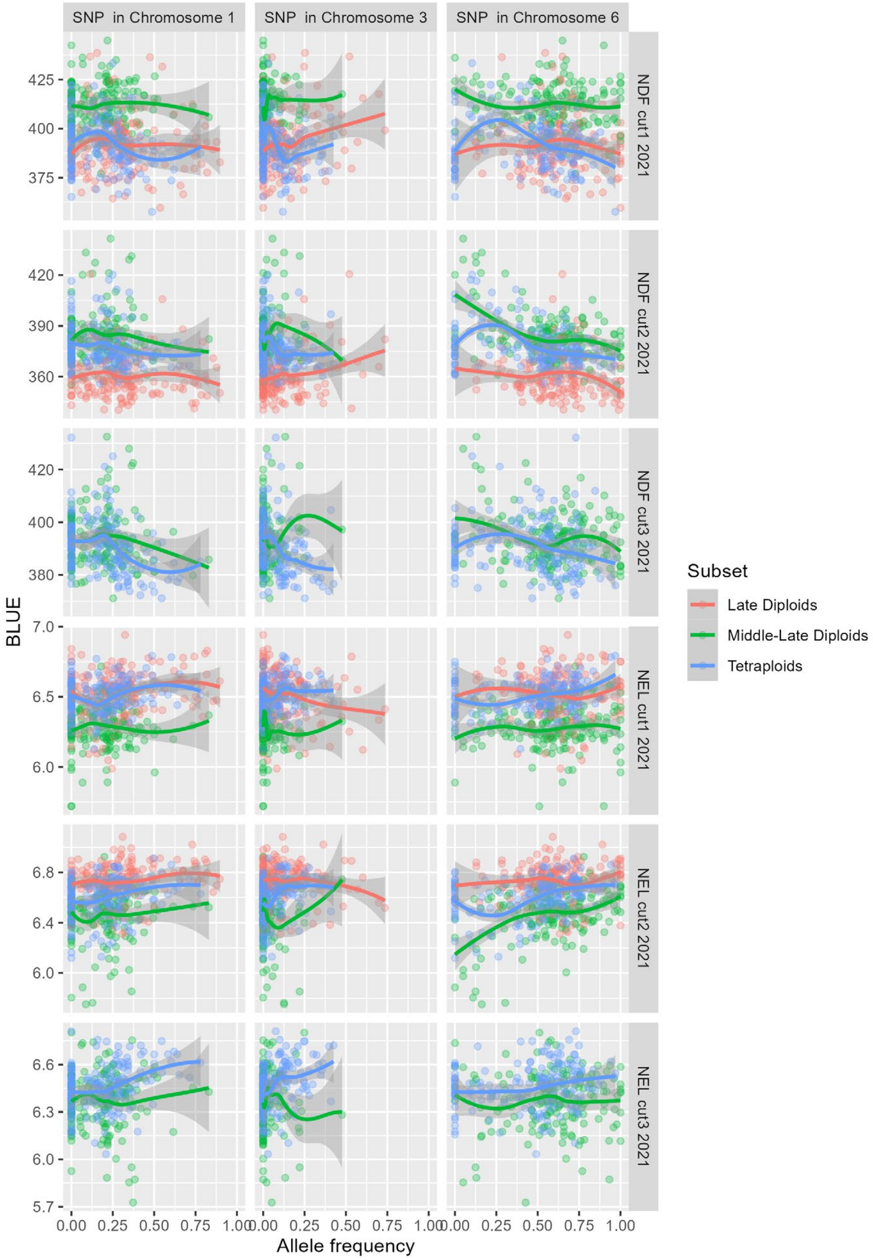


Figure 11 The BLUE as a function of the allele frequency for net energy lactation (NEL) and neutral detergent fiber (NDF) at three SNPs. The points are grouped and colored based on subset and for each group a Loess curve is fitted together with the standard deviation in grey.

The number of significant SNPs across all three subsets across all models summed up to 2 123. Among these SNPs, 165 were found in at least two of the three subsets, disregarding trait. The 165 SNPs would then identify possible QTL regions in the genome that are similar across maturity and ploidy. The results of paper IV showed that many genes have significant effects on all the tested traits and the SNPs were often significant for more than one trait (Figure 12). Thus, the target traits are highly polygenic and the genetic effects are confounded with multiple traits.

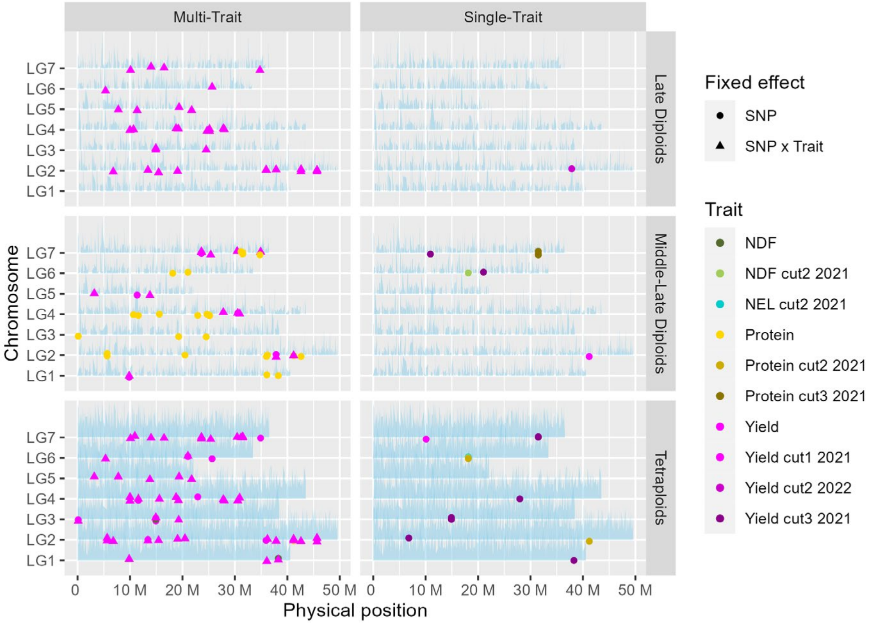


Figure 12 The position of 165 SNPS across the genome of late diploids, middle late diploids and tetraploids from either at single trait or multi trait model. The colors indicate the trait that was used in response with greens for neutral detergent fiber (NDF), blues for net energy lactation (NEL), yellows for protein and pinks for yield. The shape of the point depends on whether the SNP was identified as a single effect or a SNP×Cut interaction. The blue density plot is the mean LD at each position across the chromosome.

7. Conclusion and future prospects

Since the purpose of any breeding program is to maximize genetic gain on target traits, the significance of the results of this thesis can be summed up in terms of the breeder's equation.

$$\Delta G = \frac{ir\sigma_A}{t}$$

- **The genetic variance ($\boldsymbol{\sigma_A}$)** of the germplasm used for red clover breeding covers the wild and landrace gene pool of the southern parts of Northern Europe. However, additional genetic gains, especially to achieve breeding goals for the northern parts, could be achieved by introgression of alleles from landraces into the gene pool for breeding. Additionally, the work within this thesis found that although breeding pipelines implement approaches that minimize inbreeding depression, close genetic relationships exist among potential parent populations in the breeding pool. This can result in inbreeding depression due to close relatedness of parents. Thus, introgression of novel alleles through crossbreeding with germplasm from other sources is highly recommended. However, preferably not from breeding companies breeding for the same target environments due to potentially significant overlaps in the source gene pool. In this thesis probable molecular functions and biological processes underlying red clover's adaptation to temperature and precipitation, was also identified. The addition of increased genetic variance that will positively affect the target traits will increase response and consequently genetic gain.

- The range of a **prediction accuracy (*r*)** is between -1 and 1, hence the most important aspect of a prediction model is that its accuracy is positive. Some models had negative predictive abilities, thus the variance components was not adequately described. The success of the models were dependent on the genetic correlation between traits, which fluctuated across time points. By leveraging the models developed in this study, future research can predict the quality traits of interest based on observed phenotypes of other traits, thereby optimizing selection strategies and accelerating genetic progress in breeding programs. This thesis showed that accuracy differed across cuts, hence, the possible increase in gain is depended on how the breeders weigh their selection across cuts.
- **Time (*t*)** is a major factor on the breeder's equation and is the target of many GS protocols. GS makes it possible to decrease the breeding cycle time by predicting performance, thus, avoiding time consuming evaluation. Further decrease in time can be achieved by implementing speed breeding protocols for red clover.

One major challenge in the genomics-driven red clover breeding is its strict outcrossing nature that hinders the development of inbred lines, hence advanced cultivars are bred as populations. Given that the goal of genomic selection is to increase genetic gain at similar or lower costs than phenotype-based breeding, genotyping adequate individuals from each population is not feasible. The research conducted within this thesis showed that it is possible to perform population genetic studies as well as GP and GWAS using allele frequencies generated through pool-seq genotyping methods. However, there are still a need to describe non-additive effects for model stability across traits and cuts.

References

Abd El Moneim, A.M., Khair, M.A. and Rihawi, S. (1990) 'Effect of Genotypes and Plant Maturity on Forage Quality of Certain Forage Legume Species Under Rainfed Conditions', *Journal of Agronomy and Crop Science*, 164(2), pp. 85–92. Available at: https://doi.org/10.1111/j.1439-037X.1990.tb00790.x.

Amdahl, H. *et al.* (2016) 'Seed Yield of Norwegian and Swedish Tetraploid Red Clover (Trifolium pratense L.) Populations', *Crop Science*, 56(2), pp. 603–612. Available at: https://doi.org/10.2135/cropsci2015.07.0441.

Ashraf, B.H. *et al.* (2014) 'Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing', *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 127(6), pp. 1331–1341. Available at: https://doi.org/10.1007/s00122-014-2300-4.

Busbice, T.H. (1969) 'Inbreeding in Synthetic Varieties1', *Crop Science*, 9(5), p. cropsci1969.0011183X000900050026x. Available at: https://doi.org/10.2135/cropsci1969.0011183X000900050026x.

Campos-de Quiroz, H. and Ortega-Klose, F. (2001) 'Genetic variability among elite red clover (Trifolium pratense L.) parents used in Chile as revealed by RAPD markers', *Euphytica*, 122(1), pp. 61–67. Available at: https://doi.org/10.1023/A:1012617504493.

Charlesworth, B. and Charlesworth, D. (1999) 'The genetic basis of inbreeding depression', *Genetical Research*, 74(3), pp. 329–340. Available at: https://doi.org/10.1017/s0016672399004152.

Charlesworth, D. and Charlesworth, B. (1979) 'The Evolutionary Genetics of Sexual Systems in Flowering Plants', *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 205(1161), pp. 513–530.

Clouard, C. and Nettelblad, C. (2024) 'Genotyping of SNPs in bread wheat at reduced cost from pooled experiments and imputation', *Theoretical and Applied Genetics*, 137(1), p. 26. Available at: https://doi.org/10.1007/s00122-023-04533-5.

Cuevas, J. *et al.* (2023) 'Modeling genotype × environment interaction for single and multitrait genomic prediction in potato (Solanum tuberosum L.)', *G3 Genes|Genomes|Genetics*, 13(2), p. jkac322. Available at: https://doi.org/10.1093/g3journal/jkac322.

Dawson, J.R.O. and Street, H.E. (1959) 'Behavior in Culture of Excised Root Clones of Red Clover', *Botanical Gazette*, 120(4), pp. 217–227.

De Vega, J.J. *et al.* (2015) 'Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement', *Scientific Reports*, 5(1), p. 17394. Available at: https://doi.org/10.1038/srep17394.

Ellison, N.W. *et al.* (2006) 'Molecular phylogenetics of the clover genus (Trifolium—Leguminosae)', *Molecular Phylogenetics and Evolution*, 39(3), pp. 688–705. Available at: https://doi.org/10.1016/j.ympev.2006.01.004.

Futschik, A. and Schlötterer, C. (2010) 'The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples', *Genetics*, 186(1), pp. 207–218. Available at: https://doi.org/10.1534/genetics.110.114397.

Gillett, J.M. *et al.* (2001) *World of clovers*. Iowa State University Press.

Glover, J.D. *et al.* (2010) 'Harvested perennial grasslands provide ecological benchmarks for agricultural sustainability', *Agriculture, Ecosystems & Environment*, 137(1–2), pp. 3–12. Available at: https://doi.org/10.1016/J.AGEE.2009.11.001.

Gollob, H.F. (1968) 'A statistical model which combines features of factor analytic and analysis of variance techniques', *Psychometrika*, 33(1), pp. 73–115. Available at: https://doi.org/10.1007/BF02289676.

Greene, S.L., Gritsenko, M. and Vandemark, G. (2004) 'Relating Morphologic and RAPD Marker Varlation to Collection Site Environment in wild Populations of Red Clover (Trifolium Pratense L.)', *Genetic Resources and Crop Evolution*, 51(6), pp. 643–653. Available at: https://doi.org/10.1023/B:GRES.0000024655.48989.ab.

Henderson, C.R. (1975) 'Best Linear Unbiased Estimation and Prediction under a Selection Model', *Biometrics*, 31(2), pp. 423–447. Available at: https://doi.org/10.2307/2529430.

Henderson, C.R. (1984) 'Estimation of Variances and Covariances under Multiple Trait Models', *Journal of Dairy Science*, 67(7), pp. 1581–1589. Available at: https://doi.org/10.3168/jds.S0022-0302(84)81480-0.

Herrmann, D. *et al.* (2005) 'Optimization of bulked AFLP analysis and its application for exploring diversity of natural and cultivated populations of red clover', *Genome*, 48(3), pp. 474–486. Available at: https://doi.org/10.1139/g05-011.

Jones, C. *et al.* (2020) 'Population structure and genetic diversity in red clover (Trifolium pratense L.) germplasm', *Scientific Reports*, 10(1), p. 8364. Available at: https://doi.org/10.1038/s41598-020-64989-z.

Jordbruksaktuellt (2020) *23 år innan ny sort når marknaden*, *Jordbruksaktuellt*. Available at: https://www.ja.se/artikel/2226512/23-r-innan-ny-sort-nr-marknaden.html (Accessed: 2 November 2020).

Kjærgaard, T. (2003) ‘A Plant that Changed the World: The rise and fall of clover 1000-2000’, *Landscape Research*, 28(1), pp. 41–49. Available at: https://doi.org/10.1080/01426390306531.

Kofler, R., Betancourt, A.J. and Schlötterer, C. (2012) ‘Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila melanogaster’, *PLOS Genetics*, 8(1), p. e1002487. Available at: https://doi.org/10.1371/journal.pgen.1002487.

Li, C. *et al.* (2022) ‘Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding’, *Nature Plants*, 8(7), pp. 750–763. Available at: https://doi.org/10.1038/s41477-022-01190-2.

Merkenshlager, F. (1934) ‘Migration and distribution of red clover in Europe’, *Herb. Rev*, 1934(2), pp. 88–92.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001) ‘Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps’, *Genetics*, 157(4), pp. 1819–1829. Available at: https://doi.org/10.1093/genetics/157.4.1819.

Meyer, K. (2009) ‘Factor-analytic models for genotype × environment type problems and structured covariance matrices’, *Genetics Selection Evolution*, 41(1), p. 21. Available at: https://doi.org/10.1186/1297-9686-41-21.

Mrode, R.A. (2014) *Linear models for the prediction of animal breeding values*. Third edition. Boston, MA : CABI, [2014] ©2014. Available at: https://search.library.wisc.edu/catalog/9910197198702121.

Nay, M.M. *et al.* (2023) ‘Multi-location trials and population-based genotyping reveal high diversity and adaptation to breeding environments in a large collection of red clover’, *Frontiers in Plant Science*, 14. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1128823 (Accessed: 3 March 2023).

Öhberg, H. (2008) *Studies of the persistence of red clover cultivars in Sweden*. Umeå. Available at: https://pub.epsilon.slu.se/1741/ (Accessed: 7 July 2021).

Praat, M., De Smet, I. and van Zanten, M. (2021) ‘Protein kinase and phosphatase control of plant temperature responses’, *Journal of Experimental Botany*, 72(21), pp. 7459–7473. Available at: https://doi.org/10.1093/jxb/erab345.

Riday, H. (2010) ‘Progress made in improving red clover (Trifolium pratense L.) through breeding’, *International Journal of Plant Breeding*, 4(1), pp. 22–29.

Sato, S. *et al.* (2005) ‘Comprehensive Structural Analysis of the Genome of Red Clover (Trifolium pratense L.)’, *DNA Research*, 12(5), pp. 301–364. Available at: https://doi.org/10.1093/dnares/dsi018.

Schneeberger, K. *et al.* (2009) ‘SHOREmap: simultaneous mapping and mutation identification by deep sequencing’, *Nature Methods*, 6(8), pp. 550–551. Available at: https://doi.org/10.1038/nmeth0809-550.

Semagn, K. *et al.* (2022) ‘Comparison of single-trait and multi-trait genomic predictions on agronomic and disease resistance traits in spring wheat’, *Theoretical and Applied Genetics*, 135(8), pp. 2747–2767. Available at: https://doi.org/10.1007/s00122-022-04147-3.

Simeão Resende, R.M., Casler, M.D. and de Resende, M.D.V. (2014) ‘Genomic Selection in Forage Breeding: Accuracy and Methods’, *Crop Science*, 54(1), pp. 143–156. Available at: https://doi.org/10.2135/cropsci2013.05.0353.

Sjödin, J. and Ellerström, S. (1986) ‘Autopolyploid forage crops’.

Steinshamn, H. and Thuen, E. (2008) ‘White or red clover-grass silage in organic dairy milk production: Grassland productivity and milk production responses with different levels of concentrate’, *Livestock Science*, 119(1), pp. 202–215. Available at: https://doi.org/10.1016/j.livsci.2008.04.004.

Sturz, A.V. *et al.* (1997) ‘Biodiversity of endophytic bacteria which colonize red clover nodules, roots, stems and foliage and their influence on host growth’, *Biology and Fertility of Soils*, 25(1), pp. 13–19. Available at: https://doi.org/10.1007/s003740050273.

Tajima, F. (1989) ‘Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism’, *Genetics*, 123(3), pp. 585–595.

Taylor, N.L. (1982) ‘Stability of S Alleles in a Doublecross Hybrid of Red Clover1’, *Crop Science*, 22(6), p. cropsci1982.0011183X002200060032x. Available at: https://doi.org/10.2135/cropsci1982.0011183X002200060032x.

Taylor, N.L. and Quesenberry, K.H. (1996) ‘Reproductive Biology, Genetics and Evolution’, in N.L. Taylor and K.H. Quesenberry (eds) *Red Clover Science*. Dordrecht: Springer Netherlands (Current Plant Science and Biotechnology in Agriculture), pp. 25–43. Available at: https://doi.org/10.1007/978-94-015-8692-4_3.

Thilakarathna, M.S. *et al.* (2017) ‘Red clover varieties show nitrogen fixing advantage during the early stages of seedling development’, *Canadian Journal of Plant Science* [Preprint]. Available at: https://doi.org/10.1139/cjps-2017-0071.

Turner, T.L. *et al.* (2010) ‘Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils’, *Nature Genetics*, 42(3), pp. 260–263. Available at: https://doi.org/10.1038/ng.515.

Ulloa, O., Ortega, F. and Campos, H. (2003) ‘Analysis of genetic diversity in red clover (Trifolium pratense L.) breeding populations as revealed by RAPD genetic markers’, *Genome*, 46(4), pp. 529–535. Available at: https://doi.org/10.1139/g03-030.

VanRaden, P.M. (2008) 'Efficient Methods to Compute Genomic Predictions', *Journal of Dairy Science*, 91(11), pp. 4414–4423. Available at: https://doi.org/10.3168/jds.2007-0980.

VanRaden, P.M. *et al.* (2011) 'Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss', *Journal of Dairy Science*, 94(11), pp. 5673–5682. Available at: https://doi.org/10.3168/jds.2011-4500.

Varona, L. *et al.* (2018) 'Genomic selection models for directional dominance: an example for litter size in pigs', *Genetics Selection Evolution*, 50(1), p. 1. Available at: https://doi.org/10.1186/s12711-018-0374-1.

Vleugels, T., Roldán-Ruiz, I. and Cnops, G. (2015) 'Influence of flower and flowering characteristics on seed yield in diploid and tetraploid red clover', *Plant Breeding*, 134(1), pp. 56–61. Available at: https://doi.org/10.1111/pbr.12224.

Werner, C.R. *et al.* (2023) 'Genomic selection strategies for clonally propagated crops', *Theoretical and Applied Genetics*, 136(4), p. 74. Available at: https://doi.org/10.1007/s00122-023-04300-6.

Whitlock, M.C. (2002) 'Selection, Load and Inbreeding Depression in a Large Metapopulation', *Genetics*, 160(3), pp. 1191–1202. Available at: https://doi.org/10.1093/genetics/160.3.1191.

Williams, R. and Williams, W. (1947) 'Genetics of red clover (Trifolium pratense L.) compatibility: III. The frequency of incompatibility S alleles in two non-pedigree populations of red clover', *Journal of Genetics*, 48, pp. 69–79.

Williams, R.D. and Silow, R.A. (1933) 'Genetics of red clover (Trifolium pratense L.), compatibility. I', *Journal of Genetics*, 27(2), pp. 341–362. Available at: https://doi.org/10.1007/BF02984421.

Wright, S. (1949) 'The Genetical Structure of Populations', *Annals of Eugenics*, 15(1), pp. 323–354. Available at: https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.

Wright, S. (1965) 'The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating', *Evolution*, 19(3), pp. 395–420. Available at: https://doi.org/10.2307/2406450.

Zanotto, S. *et al.* (2023) 'A genome-wide association study of freezing tolerance in red clover (Trifolium pratense L.) germplasm of European origin', *Frontiers in Plant Science*, 14. Available at: https://doi.org/10.3389/fpls.2023.1189662.

Zhou, Y. *et al.* (2017) 'Systematic bias of correlation coefficient may explain negative accuracy of genomic prediction', *Briefings in Bioinformatics*, 18(5), pp. 744–753. Available at: https://doi.org/10.1093/bib/bbw064.

Popular science summary

Plant breeding is under increasing pressure to improve crops' performance due to factors, such as climate change, a growing population and financial demands. Forages are a widely cultivated crop in Northern Europe and are vital for the ruminant industry and ecological farming. Forage is a mixture of grasses and legumes that grows over multiple years and can be harvested up to three times per season. Red clover is one of the key forage legumes and as a legume red clover had a high protein content attributed to its symbiosis with bacteria that binds atmospheric nitrogen.

Improving red clover yields and forage quality faces multiple challenges, due to factors such as changes in growing conditions. One key approach to meeting the demands of red clover is to increase genetic gain by improving plant breeding efficiency. Genetic gain is the improvement of a trait over generations attributed to the change in the plants genetics. Plant breeding aims to increase genetic gain by influencing three components, genetic variance, accuracy, intensity and time. Genetic variation is the genetic resources available in the breeding population. Accuracy and intensity is how well a breeder select parents based on their genetic attributer rather than environmental effects, and how many parents are chosen. The time refers to the time required for each breeding cycle. This thesis focused on studying genetic variation in the northern European red clover and investigated various statistical models in order to implement selection based on genetics (genomic selection) for increased efficiency of red clover breeding.

The red clover genetic resources studied in this thesis include wild populations, landraces (farmer’s varieties), old cultivars, and red clover used in current breeding programs. These include both diploid and tetraploid red clover. The analysis of genetic variation in red clover in current breeding programs may be insufficient for efficient improvement. However, potentially interesting genetic materials for breeding were identified, and their inclusion in the crop's breeding programs can aid in increasing genetic gain.

Field trials were conducted for diploids of two maturity classes (late and middle-late) and tetraploids at different locations in Sweden, and forage yield and forage quality data were collected. Their corresponding genomic data were obtained by DNA sequencing of specific DNA markers. Phenotypic and genomic data were used for genomic prediction and marker-trait association analysis. Genomic prediction models that target a single trait at a time were compared to models targeting single traits across multiple time points or multiple traits at one time point. The aim was to use the information across multiple traits or time points to better describe the underlying genetics of the target traits.

The results showed that even though single-trait models had higher predictive ability; i.e. higher accuracy of estimates, than multi-trait models, they showed potential bias. This bias could over or underestimate the performance of red clover populations and could be due to the model's failure to capture important genetic effects. The bias could be due to genetic effects where the effect of a gene is non-linear to the response on the trait. These effects are often due to the presence of dominant alleles or epistasis effects between multiple genes. However, by introducing multiple traits and using the correlation between them these effects could be captured.

Thus, this thesis highlights key aspects of red clover in terms of exploiting its available genetic resources for increased genetic gain through genomics-driven breeding. Additionally, pitfalls in utilizing various prediction models for increased accuracy and speed were highlighted. This research contributes to red clover breeding efficiency so that new varieties can be developed faster and meet current and future demands.

Populärvetenskaplig sammanfattning

Den stigande efterfrågan på ökad avkastning hos grödor på grund av klimatförändringar samt ekonomiska faktorer sätter en större press på utvecklingen av nya sorter genom växtförädling. Sett till odlingsareal är vall den mest odlande gröda i Norden och är viktig för mjölk- och köttindustrin samt för det ekologiska jordbruket. Vall är en blandning av gräs och baljväxter som ligger över flera år och skördas flera gånger per säsong. Rödklöver är en av de mest odlade vallbaljväxterna och har en hög proteinhalt. Proteinhalten är beroende av mängden tillgängligt kväve och rödklövern har ett samarbete med bakterier i jorden vilka binder atmosfäriskt kväve i jorden så det kan tas upp av växter. Rödklöver står inför många utmaningar när efterfrågan på högre avkastning samt högre foderkvalitet begränsas av ökat tryck från nya sjukdomar och stressfaktorer på grund av klimatförändringarna.

För att möta den efterfrågade avkastningen behöver rödklöverförädlingen effektiviseras genom att öka den genetiska vinsten. Genetisk vinst är den förbättring av förädlingsmaterial över generationer som kan anknytas till genetik. Växtförädlingens mål är att öka den genetiska vinsten genom att optimera fyra faktorer, genetisk variation, precision, intensitet, samt tid. Genetisk variation beskriver de tillgängliga genetiska resurserna hos de populationerna som används i förädlingen. Precision och intensitet syftar på hur väl förädlaren väljer sina föräldrar baserat på genetik framför miljöeffekter och hur många föräldrar som väljs. Tid syftar på hur lång tid en förädlingscykel tar, alltså tiden från korsningen av föräldrar till en ny sort eller nya föräldrar kandidater. Syftet i denna avhandling var att studera den genetiska variationen hos rödklöver från Norden och utformat statistiska modeller för att förutse prestationsförmågan av rödklöverpopulationer baserat på deras genetik. Implementeringen av prediktionsmodeller baserat på genetik kallas genomisk selektion och kan öka precisionen samt förkorta tiden för en förädlingscykel.

Studierna av den genetiska variationen hos rödklöver från Norden inkluderade vild klöver, lantraser, gamla sorter och de sorter som används för förädling idag. Resultaten visade att förädling idag inte utnyttjar all tillgänglig variation. Populationer av rödklöver som skulle kunna användas i rödklöver förädling för att snabbare nå avkastningsmålen identifierades. Dessa populationer var samlingar av vilt material och lantraser.

Rödklövern delas upp i olika förädlingsprogram beroende på när den blommar, Anledningen till att dela upp rödklöver efter mognad (sen, medel-sen) är på grund av hur reaktiv rödklövern är på miljön. För att ha lyckade fältförsök måste rödklövern odlas i miljöer som passar dess mognadsgrad. Tetraploid rödklöver är inte lika känslig och kan odlas på fler platser.

Fältförsök utfördes på två platser med sen rödklöver, två platser med medel-sen rödklöver och tre platser med tetraploid rödklöver, från vilket värden på torrsubstans och foderkvalitet analyserades.

Genomisk data samlades från alla populationer med utvalda DNA-markörer. Informationen om genotyp (hur de ser ut genetiskt) och fenotyp (hur de uttrycker egenskaper) användes för att skapa och evaluera genomiska prediktionsmodeller. De genomiska prediktionsmodellerna undersöktes med både enskilda egenskaper samt flera egenskaper samtidigt för att se hur genetiken hos rödklöver bäst kan matematiskt beskrivas. Resultaten visade att trots att precisionen var högst för modeller som använde sig av enstaka egenskaper så visade de partiskhet. Partiskhet kan bero på att modellen misslyckas med att beskriva viktiga genetiska effekter som i sin tur under- eller överskattar en rödklöveraccessions prestation. I modellen antas relationen mellan genotyp och fenotyp vara linjär, alltså att de två generna från mor och far påverkar fenotypen till samma grad. Om en gen har starkare effekt på fenotypen blir sambandet icke-linjärt och prediktionen av GP modellen kan då över- eller underskatta utkomsten. Genom att introducera fler egenskaper vilka är genetisk korrelerade kan dessa icke-linjära samband bättre förklaras av modellen.

Såldes, visar denna avhandling viktiga aspekter för ökad effektivisering i växtförädling så som nya källor till genetisk variation samt pekar ut potentiella fallgropar för genomisk prediktion. Appliceringen av dessa resultat kan effektivisera rödklöverförädlingen därmed möta de nuvarande och framtida utmaningarna vad gäller avkastning och foderkvalitet snabbare.

Acknowledgements

I want to thank my supervisors for supporting me on this journey.

I want to thank my main supervisor Mulatu Geleta Dida, you have not only taught me key concepts of genetics but also how to be a good scientist. At times when I was overwhelmed with my own deadlines you sat me down and helped me prioritise my work. You have always welcomed my new ideas but made sure that I never made the wrong conclusions or completely diverted from the objectives of my study. All while never making me feel critiqued, only guided. You have always made time for me whenever I've been in need of a meeting and you've worked hard to put quality work into commenting and revising my manuscripts.

I want to thank Cecilia Hammenhag, I feel like you've been my biggest cheerleader through this! Your compliments after I presented my work have always been a huge comfort. I've always been inspired by your efficiency and kindness. Thank you for helping me navigate complex academic processes. Most importantly, you have taught me how to present my research to people outside of the university. Without that skill a scientist won't come far.

I want to thank Rodomiro Ortiz, you are truly a database of knowledge. I've always appreciated our talks and your stories of the professors and breeders you've worked with. It has helped me put my career path in perspective. I've also appreciated the mini-defences at every supervisory meeting; they've prepared me for what's to come.

I want to thank Lucia Gutiérrez, you swooped in as my co-supervisor in the last year of my study and helped me disentangle the maths behind genomic prediction. Your pedagogical way of describing these complex concepts has helped me be confident in my data analysis. How you taught me these concepts will guide me through my future career; it's almost like a cheat code.

I want to thank my project group and co-authors, Linda Öhlund, late Elisabet Nadeau and David Parsons. You've always been positive, curious and helpful; which have guided me a lot. Thank you for sharing your expertise.

I want to thank SLU Grogrund for the funding that made this work possible. What you do for Swedish plant breeding is invaluable for securing a better world for my generation and future generations.

I want to thank my colleagues at SLU who have made this time memorable; from making long and confusing Zoom lectures bearable to fun and interesting topics at lunch. I want to thank my fellow PhD students, past and present for sharing your ups and downs and making this unique journey oh so relatable. I wish you the best of luck in your future careers and hope we cross paths many times more. I want to thank the researchers who motivate and inspire, who share their experiences in research and how to navigate the academic labyrinth. Your assurance of a non-hierarchical workplace for students permits us to reach our full potential.

I want to thank my collaborators and now colleagues at Lantmännen who have been a great support and given me invaluable insight into forage breeding. I want to thank my wonderful forage team and amazing PI group. I'm thankful for the help in detangling everything from field data to complex prediction models. I look forward to working with you and making sure we can implement the results of this research and do more amazing work.

I want to thank Lorena who supported me and put me in contact with Gregor Gorjanc and his lab at the Roslin Institute. Even though I was there for only two months, I learned so many important things about breeding, which I refer back to all the time. It was an invaluable experience!

I want to thank my dad, I remember when I was little and you said that if I kept working hard in school I could get a PhD and that it was the highest achievement in learning. I believe that knowing from a young age that I could reach the highest achievement of anything if I only put my mind to it, has shaped how I view any challenge today.

I want to thank my mom, you've always motivated me to take on any challenge or risk. This is mostly because I know you have a safe space for me to come back to if it doesn't work out.

I want to thank my aunt Camilla, you taught me that if you want to do fun things, you have to do all the tough and boring things first. I owe a lot of my working discipline to you.

I want to thank Alice, you are a great supporter and listener who reminds me to enjoy the simple things in life.

I want to thank Jacob, you have been an amazing support and aid through all the tough times and helped me celebrate all the good times. Your interest in my research has been a constant motivation. I only wish I could support you during your upcoming PhD journey as well as you have supported me.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 25 October 2021
doi: 10.3389/fpls.2021.748750

Insights Into the Genetic Diversity of Nordic Red Clover (*Trifolium pratense*) Revealed by SeqSNP-Based Genic Markers

Johanna Osterman, Cecilia Hammenhag, Rodomiro Ortiz and Mulatu Geleta*

Department of Plant Breeding, Swedish University of Agricultural Sciences, Lomma, Sweden

Red clover (*Trifolium pratense*) is one of the most important fodder crops worldwide. The knowledge of genetic diversity among red clover populations, however, is under development. This study provides insights into its genetic diversity, using single nucleotide polymorphism (SNP) markers to define population structure in wild and cultivated red clover. Twenty-nine accessions representing the genetic resources available at NordGen (the Nordic gene bank) and Lantmännen (a Swedish agricultural company with a red clover breeding program) were used for this study. Genotyping was performed via SeqSNP, a targeted genotype by sequencing method that offers the capability to target specific SNP loci and enables *de novo* discovery of new SNPs. The SNPs were identified through a SNP mining approach based on coding sequences of red clover genes known for their involvement in development and stress responses. After filtering the genotypic data using various criteria, 623 bi-allelic SNPs, including 327 originally targeted and 296 *de novo* discovered SNPs were used for population genetics analyses. Seventy-one of the SNP loci were under selection considering both Hardy-Weinberg equilibrium and pairwise F_{ST} distributions. The average observed heterozygosity (H_O), within population diversity (H_S) and overall diversity (H_T) were 0.22, 0.21 and 0.22, respectively. The tetraploids had higher average H_O (0.35) than diploids (0.21). The analysis of molecular variance (AMOVA) showed low but significant variation among accessions (5.4%; $P < 0.001$), and among diploids and tetraploids (1.08%; $P = 0.02$). This study revealed a low mean inbreeding coefficient ($F_{IS} = -0.04$) exhibiting the strict outcrossing nature of red clover. As per cluster, principal coordinate and discriminant analyses, most wild populations were grouped together and were clearly differentiated from the cultivated types. The cultivated types of red clover had a similar level of genetic diversity, suggesting that modern red clover breeding programs did not negatively affect genetic diversity or population structure. Hence, the breeding material used by Lantmännen represents the major genetic resources in Scandinavia. This knowledge of how different types of red clover accessions relate to each other and the level of outcrossing and heterozygosity will be useful for future red clover breeding.

Keywords: DAPC, gene targeting markers, heterozygosity, loci under selection, population structure, red clover, SeqSNP

OPEN ACCESS

Edited by:
Marcelino Perez De La Vega,
Universidad de León, Spain

Reviewed by:
Leif Skot,
Aberystwyth University,
United Kingdom
Wenli Li,
U. S. Dairy Forage Research Center,
Agricultural Research Service,
United States Department
of Agriculture (USDA), United States

****Correspondence:***
Johanna Osterman
johanna.osterman@slu.se

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *28 July 2021*
Accepted: *01 October 2021*
Published: *25 October 2021*

Citation:
Osterman J, Hammenhag C,
Ortiz R and Geleta M (2021) Insights
Into the Genetic Diversity of Nordic
Red Clover (Trifolium pratense)
Revealed by SeqSNP-Based Genic
Markers. Front. Plant Sci. 12:748750.
doi: 10.3389/fpls.2021.748750

INTRODUCTION

Red clover (*Trifolium pratense*) is an important crop to secure sustainable cattle farming and thus meat and dairy production, as it is one of the most important forage legumes worldwide (Smith et al., 1985; Taylor and Quesenberry, 1996a). It is a perennial forage legume, which can be harvested multiple times within a year. Red clover is favored due to its nutritional value and positive effect on soil quality. It grows in symbiosis with nitrogen fixing bacteria in the rhizosphere, thus increasing soil fertility (Sturz et al., 1997; Thilakarathna et al., 2017). Until the 1930s, when industrial fertilizers started to be widely used, red clover was an important crop for the European agriculture mainly due to its symbiosis with rhizobacteria, thereby securing high yielding harvests (Kjærgaard, 2003).

According to Merkenshlager (1934), red clover was first cultivated in Spain during the 13*th* century and then spread throughout Europe. Records show that it has been cultivated in the southern parts since the 1560s and in the northern parts since the 1770s. It continued to spread to the temperate regions of North and South America, to New Zealand and Australia and, to China and Japan (Taylor and Quesenberry, 1996a). Red clover is naturally diploid (2n = 2x = 14; Evans, 1954), which includes wild populations as well as landraces and traditional cultivars. However, modern autotetraploid red clover cultivars (2n = 4x = 28) have been developed from diploid genotypes through chromosome doubling (Sjödin and Ellerström, 1986; Taylor and Quesenberry, 1996b).

The financial sustainability of small-scale farmers, in competition with bigger foreign markets, is an important aspect to be considered with regard to the production of local and sustainable beef and dairy products, as already noted in Sweden (Fischer and Röös, 2018). Even though the farmers in Sweden follow quality-grazing requirements, the cattle still require additional fodder to ensure their high quality and reliable milk production as well as the health and development of their calves (Åkerlund, 2008; Kilsgård, 2015). The requirement of adding, for example, soybean meal to the fodder to increase crude protein levels is not sustainable for cattle farming in northern Europe or elsewhere where there is no soybean production. The development of local red clover cultivars are therefore important for the global beef and dairy industry.

To assist farmers in providing good quality beef and dairy to consumers, plant breeders need to develop red clover cultivars that have great nutritional value, along with superior yield. Additionally, the crop should show high persistence under local climates and biotic stresses. For example, the root-rot disease caused by fungal pathogens can wipe out an entire field of red clover in the absence of pesticide use (Bengtsson, 1961). The fungicide used to prevent root-rot in the late 1900s is not approved by environmental agencies (United States Environmental Protection Agency, 2006). Hence, the best way to prevent the disease is to develop resistant cultivars against the pathogen. Persistence, under biotic and abiotic stresses, is a difficult trait to breed for and so far, only one quantitative trait locus (QTL) linked to the persistence trait has been detected in *T. pratense* (Herrmann et al., 2008). Unique climate zones, such as the lands closer to the polar circle in the northern hemisphere can be highly variable in temperature, precipitation and day length due to seasonal changes. Thus, there are additional requirements to realize persistent red clover cultivars suitable for mid and high-latitude regions.

DNA markers have been widely used to conduct population genetics research in crops and their wild and weedy relatives, both for understanding their genetic makeup as well as to establish their genetic and phylogenetic relationships. SeqSNP is a targeted genotype by sequencing method that can be designed for genotyping of known highly polymorphic SNPs. If desirable SNPs are unavailable, novel SNP candidates can be identified through allele mining of genomic resources of a target crop from the GenBank. The significance of these novel SNPs can be evaluated by using them for population genetic studies. Based on the results, a core set of SNPs can be developed for various applications including marker-aided breeding. *T. pratense* is among the crops with publicly available whole genome sequences (GenBank accessions: GCA_000583005.2, GCA_900079335.1 and GCA_900292005.1) (Istvánek et al., 2014; De Vega et al., 2015). The latest one, GCA_900292005.1, is a chromosome level assembly with a total sequence length of 351.6 Mbp.[1] This assembly shows high synteny between *T. pratense* and *Medicago truncatula*, a model species for the legume family (Cook, 1999; Frugoli and Harris, 2001). A coding sequence assembly based on a cultivated red clover population is also available at the Legume Information System (LIS[2]). The annotation of the assemblies can be used to locate gene sequences and thus identify SNPs targeting desired genes.

As explained by Poczai et al. (2013), gene targeting markers (GTMs) is a category of markers that target specific genes. When located within the coding regions of a specific gene, GTMs differ from non-genic markers that are more reliant on association with target genes or loci. By developing SNPs within coding regions of a gene, the likelihood of accurately describing potential genotypic variation with regard to a specific trait increases. This enables the description of genetic diversity within and among populations with respect to important traits (van Tienderen et al., 2002). The aim of the present study was to use SNPs as GTMs to determine the genetic diversity and population structure of red clover represented by its genetic resources available in Northern Europe, where it is a major forage legume.

MATERIALS AND METHODS

Plant Material and Sampling

Twenty-nine red clover accessions were used for this study (**Table 1**). Twenty-one of these accessions were obtained from NordGen, a genebank for genetic resources in the Nordic countries, by ordering via Nordic Baltic Genebanks Information System (GeNBIS),[3] whereas the remaining eight were obtained

[1] https://www.ncbi.nlm.nih.gov/assembly/GCA_900292005.1/

[2] https://legumeinfo.org/

[3] https://www.nordic-baltic-genebanks.org/gringlobal/search.aspx

TABLE 1 | Summary of the accessions with origin, type (B = breeding population, C_L = cultivar from Lantmännen, C_N cultivar from NordGen, L = landrace population, S = synthetic population, W = wild population), percent polymorphic loci (%PL), within population diversity (H_S), observed heterozygosity (H_O), inbreeding coefficient (F_{IS}), discriminant analysis of principal components (DAPC) cluster designation and Nei's standard genetic distance (NSGD).

Accession	Origin	Type	%PL	H_S	H_O	F_{IS}	DAPC clusters	NSGD	
								Mean	SD
NGB1730[c]	Denmark	C_N	65.0	0.21	0.20	0.01	1 and 2; 70%, 30%	0.026	0.005
NGB1743[c]	Denmark	C_N	67.1	0.21	0.20	0.02	1 and 2; 60%, 40%	0.026	0.006
NGB4126[c]	Denmark	C_N	63.9	0.19	0.19	0.00	1 and 2; 90%, 10%	0.025	0.005
NGB8393[c]	Denmark	C_N	64.5	0.21	0.20	0.01	1 and 2; 80%, 20%	0.024	0.005
NGB9228[c]	Denmark	C_N	65.7	0.21	0.21	0.01	1 and 2; 60%, 40%	0.027	0.005
NGB11586[c]	Denmark	C_N	61.0	0.21	0.19	0.00	1 and 2; 75%, 25%	0.029	0.006
NGB11605[c]	Denmark	C_N	58.1	0.20	0.22	−0.08	1 and 2; 50% each	0.025	0.005
NGB11608[c]	Denmark	B	58.4	0.19	0.19	0.02	1 and 2; 20%, 80%	0.025	0.005
NGB1142[c]	Finland	L	70.3	0.22	0.21	0.03	1 and 2; 50% each	0.025	0.005
NGB14327[c]	Finland	L	68.2	0.21	0.21	−0.01	1 and 2; 70%, 30%	0.026	0.005
NGB14444[c]	Finland	C_N	63.6	0.19	0.20	−0.03	1 and 2; 80%, 20%	0.026	0.005
NGB14448[c]	Finland	W	61.3	0.21	0.21	−0.03	1 and 2; 50% each	0.024	0.005
NGB2194[c]	Norway	L	64.7	0.20	0.20	−0.02	1 and 2; 90%, 10%	0.026	0.005
NGB2202[c]	Norway	L	63.7	0.20	0.19	0.05	1 and 2; 60%, 40%	0.027	0.006
NGB15558[c]	Norway	W	53.1	0.18	0.19	−0.06	3	0.025	0.005
NGB15623[c]	Norway	W	59.4	0.19	0.19	−0.02	1 and 2; 50% each	0.026	0.005
NGB1009[c]	Sweden	W	48.2	0.16	0.16	0.01	3	0.026	0.005
NGB1420[c]	Sweden	W	55.7	0.18	0.18	0.01	3	0.024	0.005
NGB6184[c]	Sweden	L	70.3	0.21	0.20	0.05	1 and 2; 60%, 40%	0.026	0.005
NGB9966[c]	Sweden	L	65.8	0.21	0.19	0.06	1 and 2; 50% each	0.027	0.006
NGB24176[c]	Russia	W	54.4	0.16	0.16	0.00	3	0.024	0.005
LÖRK0390[a]	Lantmännen	S	68.7	0.22	0.22	−0.02	1 and 2; 50% each	0.026	0.006
SW ARES[a]	Lantmännen	C_L	64.5	0.21	0.21	−0.05	1 and 2; 70%, 30%	0.024	0.005
SW YNGVE[a]	Lantmännen	C_L	65.0	0.21	0.19	0.03	1 and 2; 30%, 70%	0.025	0.005
SWÅ RK07001[a]	Lantmännen	S	65.7	0.21	0.21	0.01	1 and 2; 50% each	0.026	0.005
SWA 1675209[b]	Lantmännen	S	71.3	0.25	0.34	−0.31	1 and 2; 10%, 90%	0.024	0.005
SW RK1158[b]	Lantmännen	S	73.8	0.25	0.36	−0.33	2	0.024	0.005
SW RK1166[b]	Lantmännen	S	75.9	0.26	0.35	−0.29	1 and 2; 40%, 60%	0.026	0.006
Vicky[b]	Lantmännen	C_L	73.8	0.25	0.34	−0.29	1 and 2; 40%, 60%	0.025	0.005
Mean			64.2	0.21	0.22	−0.04		0.025	0.005

[a] = known diploid; [b] = known tetraploid; [c] = ploidy not determined.

from Lantmännen Seed,[4] a plant breeding and agricultural seed company based in Sweden. The NordGen accessions were selected based on their passport data and represent cultivars, landraces and wild forms from the Nordic countries, as well as one Russian accession collected near the Finnish border. The accessions from Lantmännen represent cultivars and synthetic populations and include both diploids and tetraploids. Hereafter, cultivars from NordGen and Lantmännen will be referred for distinction as cultivar$_N$ and cultivar$_L$, respectively.

A minimum of 20 seeds were planted for each accession in a greenhouse at the Swedish University of Agricultural Sciences (SLU, Alnarp, Sweden) in May 2020. For each accession, two 2 L plastic pots filled with soil were used for planting. Poorly germinating seeds were treated by manual scarification according to Asci et al. (2011) and placed on Petri dishes until germination was achieved, and then transferred to pots. Two weeks after planting, extra seedlings were removed and five seedlings per pot were maintained. Later, leaf tissue from ten individual seedlings per accession were separately sampled, except for accession NGB11586, which was represented by eight individual seedlings. Hence, 288 individuals originating from different seeds were separately sampled in total. For sampling the leaf tissue, BioArk Leaf collection kit provided by LGC, Biosearch Technologies[5] was used. From each plant, ten 6 mm leaf discs were sampled using a punch and were put in a sampling plate. The collected leaf tissue was then sent to LGC, Biosearch Technologies (Berlin, Germany) for DNA extraction and subsequent genotyping. Using the Sbeadex plant kit,[6] high quality genomic DNA was extracted for SeqSNP genotyping.

[4] https://www.lantmannenseed.com/

[5] https://biosearchassets.blob.core.windows.net/assetsv6/guide_bioark-leaf-collection-kit.pdf

[6] https://biosearch-cdn.azureedge.net/assetsv6/sbeadex-plant-data-sheet.pdf

Single Nucleotide Polymorphism Selection, SeqSNP Assay Design and Sequencing

A literature review was performed on red clover targeting pathways of growth and development as well as stress and disease responses. Based on this foundation, genes related to disease resistance, stress response (such as hormone regulation or interaction) or growth and development were targeted for single nucleotide polymorphism (SNP) mining. The SNP mining was performed using coding sequences (CDS) of red clover genes downloaded from the legume information system (LIS) database[2] where the CDS of genes of interest were aligned against the red clover reference genome [*Trifolium pratense* genome v3, GenBank assembly (GCA_900292005.1)] using the Basic Local Alignment Search Tool (BLAST) of the National Center for Biotechnology Information (NCBI). As only CDS sequences with a single hit in the red clover genome were used for the SNP identification, all SNPs used in this study were considered to be from single-copy genes. If a CDS could not be clearly associated with the list by its associated protein name, the UniProt database[7] was used to identify the related pathway. If the related pathway was in accordance with the selection criteria, the sequence was chosen for the SNP mining. Alignments were accepted at a threshold of 95% sequence identity between the query and subject sequences. Whenever more than one SNP was selected per target sequence, the SNPs were at least 55 bp apart. In total, 641 target SNPs in 324 CDS sequences were used for SeqSNP assay design. Of the 641 target SNPs, 571 SNPs were fully covered (two oligo probes per target) and passed high specificity assay design (no off-target hits allowed), whereas the remaining 70 SNPs failed. Among the 571 SNPs that passed the high specificity assay design, 400 target SNPs in 247 sequences were then selected for SeqSNP analysis by taking into account the distribution of the SNPs across the genome, the function of the genes as well as the distance between SNPs within a gene (**Supplementary Table 1**). In this final set, the SNPs within a given sequence were at least 70 bp apart. Then, SeqSNP kit containing 800 high-specificity oligo probes for the 400 SNPs were produced, a sequencing library was constructed and the target SNPs were sequenced. The Illumina NextSeq 500/550 v2 platform was used for sequencing in 150 base-pair (bp) single read mode. On average, ca 217,000 reads per sample was conducted, and the average effective target SNP coverage was 501x.

Variant Discovery, Genotype Assignment, and Data Filtering

After sequencing, the raw reads were adapter-clipped and quality-trimmed (reads containing Ns removed, reads trimmed to obtain a minimum average Phred quality score of 30 over a window of ten bases, and reads with final length < 130 bp discarded). For genotype calling, the SNP genotyping pipeline was set to diploid genotyping with a minimum coverage of eight reads per sample per locus, as described below. The alignment of quality trimmed reads against the reference genome using Bowtie2 v2.2.3 (Langmead and Salzberg, 2012) and variant discovery and genotyping of the samples using Freebayes v1.0.2-16 (Garrison and Marth, 2012) showed that out of the 400 target SNP loci, 15 were monomorphic across the 288 individuals. Furthermore, among the 385 polymorphic loci, 352, 30 and 3 were bi-, tri- and tetra-allelic, respectively, within the genotyped samples. Mono-, tri- and tetra- allelic SNP loci were excluded from further analysis. For genotype determination of the bi-allelic loci, alleles with an allele-count of less than eight were set to missing, according to the threshold set for the genotype-calling pipeline, as a procedure to exclude alleles called due to a potential sequencing error. Although the ploidy level of the NordGen accessions were not indicated in their passport data, the wild populations, landraces and traditional cultivars are diploids, as described in the "Introduction." The accessions labeled as cultivar$_N$ were also considered diploids based on observed phenotypic characteristics. On the other hand, the four accessions from Lantmännen are known tetraploids. A preliminary genetic diversity analyses of these tetraploid accessions conducted using their allele frequency data, calculated from their allele counts, provided similar results with the results obtained by treating them as diploids. Hence, the genotype of each sample at each locus was determined based on the allele-count, as diploid for all accessions in the final data analyses. Hence, individuals having only one allele with an allele-count of above eight were recorded as homozygotes whereas those with an allele-count of above eight for two alleles were scored as heterozygotes. The bi-allelic SNP data were then filtered to retain only loci with missing data of less than 5%. Among the 352 bi-allelic SNPs, 21 had a missing data of over 5%, thus only 331 were retained for further analysis. In addition to the target SNPs, 292 bi-allelic SNPs that fulfilled all aforementioned filtering criteria were identified *de novo* within 75 bp range on both sides of the target SNPs.

Analysis of Molecular Variance, Population Structure and Cluster Analyses

R (R Core Team, 2013) was used to assess whether the SNPs were in agreement with the assumptions of an outcrossing crop by computing expected and observed heterozygosity and F-statistics. The analysis was performed using the method proposed by Kamvar et al. (2017) using the R package adegenet version 2.0.0 (Jombart, 2008) and hierfstat version 0.5–7 (Goudet, 2005). The analysis of molecular variance (AMOVA) was performed using Arlequin suite version 3.5.2.2 (Excoffier and Lischer, 2010) without grouping the 29 accessions as well as by grouping them based on different criteria, as presented in the results. To determine the extent of differentiation between the populations, the F$_{ST}$ (fixation index; Wright, 1931) and Nei unbiased genetic distance (Nei, 1987) between subpopulations were calculated using Arelquin ver 3.5.2.2. The principal coordinate analysis (PCoA) was calculated in R with the stats package (R Core Team, 2013) based on the Nei unbiased genetic distance matrix. The neighbor-joining (NJ) tree at a population level was constructed using MEGA7 (Kumar et al., 2016) based on the Nei unbiased genetic distance matrix. Additionally, a NJ

[7] https://www.uniprot.org/

tree at individual sample level was constructed using MEGA7 based on a pairwise genetic distance matrix calculated using the Tamura and Nei (1993) method. The discriminant analysis of principal components (DAPC) was performed in R using the adegenet package on individual genotypic data and 100 principal components based on the method of Jombart et al. (2010). The principal components were selected using a function from the adegenet package that evaluated the most informative principal components using cross-validation.

Hardy-Weinberg Equilibrium and Neutrality Tests, and Mutation Analysis

Arlequin suite version 3.5.2.2 was used to identify loci that significantly deviated from the Hardy-Weinberg equilibrium (HWE) assumptions, at 0.05 level of significance. For this analysis, 1000,000 steps in Markov-chain and 100,000 dememorization steps were used. Similarly, Arlequin was used to detect loci under selection (at 0.05 level of significance) through the examination of the joint distribution of F_{ST} and heterozygosity under a non-hierarchical finite island model, using 20,000 coalescent simulations (Excoffier et al., 2009). This led to the identification of 51 loci that deviated from HWE (each with a p-value < 0.01), and 88 loci that are presumed to be under selection (each with a p-value < 0.05). An in-house python script using the Biopython package (Cock et al., 2009) was used to determine the longest open reading frame for each marker, which was then translated to amino acid sequences using the reference and alternative allele, respectively. The two amino acid sequences generated based on the two alleles at each locus were then compared for identification of missense and nonsense mutations.

RESULTS

The 400 SNP loci (**Supplementary Table 1**) were identified by aligning one CDS per locus to the reference genome, generating a maximum of two alleles at a bi-allelic locus. However, the SeqSNP based sequencing of the 288 genotypes targeting these SNP loci revealed three alleles at 30 loci and four alleles at three loci. Contrary to this, singletons (only the reference allele) were detected at 15 of the 400 loci across the 288 individuals resulting in 352 bi-allelic SNP loci. Hence, mono-, bi-, tri-, and tetra-allelic loci accounted for 3.8%, 88%, 7.5%, and 0.8% of the 400 loci targeted for genotyping of the 29 populations (288 individuals). Interestingly, no additional sequence variation was detected within 75 bp range up and downstream of the 400 target SNP loci during the original alignment of a single CDS to the reference genome. However, *de novo* SNP calling by mapping the reads from the 288 individuals to the reference genome led to the discovery of an additional 296 bi-allelic SNPs (**Supplementary Table 1**), clearly indicating the advantages of using the SeqSNP genotyping method. Due to the complexity of analyzing tri- and tetra-allelic loci, only the bi-allelic SNPs were included in the population genetic analyses of the 29 accessions. However, the tri- and tetra-allelic loci identified in this study can be further investigated, for example to determine the ploidy of individual red clover genotypes. Among the 352 bi-allelic SNP loci, 25 loci had missing values of over 5% and hence they were excluded. Overall, 623 SNPs that include 327 original SNPs and 296 *de novo* discovered SNPs were used for final data analysis (**Supplementary Table 1**). The translation of the coding sequences containing the bi-allelic SNPs revealed that 362 of the 623 SNPs induced missense mutations (data not shown).

Population Structure, F-Statistics and Analysis of Molecular Variance

This study revealed a low genetic distance between the accessions, where the mean Nei's standard genetic distance of the accessions were quite similar, ranging only from 0.024 to 0.029 (**Table 1**). The inbreeding coefficient (F_{IS}) of known diploids ranged from −0.05 to 0.03 whereas that of known tetraploids ranged from -0.33 to -0.29. Those with unknown ploidy had F_{IS} ranging from −0.08 to 0.06 (**Table 1**) spanning the ranges for the known diploids. The average observed heterozygosity (H_O) across all loci and accessions was 0.22 with individual values ranging from 0.16 (accession NGB1009 and NGB24176) to 0.36 (accession SW RK1158; **Table 1** and **Figure 1A**). The within population diversity (H_S) ranged from 0.16 (accession NGB1009 and NGB24176) to 0.26 (SW RK1166) with an overall average of 0.21 (**Table 1** and **Figure 1A**). The overall diversity (H_T) mean estimate was 0.22 and share similar range of values with that of H_S (**Figure 1A**). The mean diversity estimate among the accessions (F_{ST}) was 0.05, whereas the mean inbreeding coefficient (F_{IS}) was −0.04 (**Figure 1A**). The summary of population genetics parameters by grouping the accessions according to their origin and type is presented in **Figures 1B,C**, respectively. Larger differences were observed between the groups in $H_{O,}$ F_{ST} and F_{IS} as compared to H_S and H_T for both origin and type-based groupings of the accessions (**Figures 1B,C**). There was low variation between the groups in general; however, most of the variation was between accessions from Lantmännen and NordGen (**Figures 1B,C**). This is most prominently shown by the F_{IS} values. Interestingly, the within population variation (H_S) of the single Russian accession (NGB24176) had the lowest mean value but the range of the individuals was corresponding to the other groups. The range of estimated values of population differentiation (F_{ST}) was larger for accessions from Sweden and Norway compared with those from other sources (**Figure 1B**) and for wild types and Lantmännen cultivars (**Figure 1C**).

The AMOVA analysis on the 29 accessions (**Table 2**) prior to grouping revealed a highly significant ($P < 0.001$) differentiation among the accessions accounting for 5.4% of the total variation. AMOVA was then conducted by grouping the 29 accessions or their subsets into different groups according to origin, population type, maturity type and ploidy. In all cases, a highly significant differentiation was obtained among accessions within groups ($P < 0.001$) explaining from 0.5% to 1.8% of the total variation (**Table 2**). A low but significant differentiation ($P = 0.012$) was revealed when the eight Lantmännen accessions were compared with the 21 NordGen

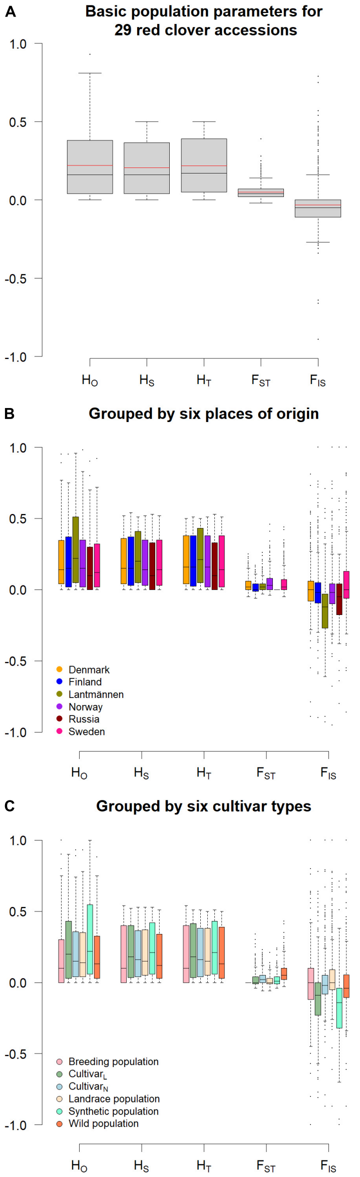


FIGURE 1 | A boxplot visualizing the summary of different population genetics parameters for **(A)** all accessions, together with a red line marking the mean values, **(B)** each group of accessions grouped according to their origin, **(C)** each group of accessions grouped according to their type. The y-axis refers to the range of values for the parameters whereas the different parameters were given along the x-axis. H_O = observed heterozygosity; H_S = within population gene diversity; H_T = overall gene diversity; F_{ST} = population differentiation; F_{IS} = inbreeding coefficient.

accessions, accounting for 0.5% of the total variation. Similarly, significant differentiation ($P = 0.018$) was obtained among groups of NordGen accessions from different countries, accounting for 0.9% of the total variation. The differentiation between different population types of the NordGen accessions was highly significant ($P < 0.001$) with 1.8% of the total variation residing among them. The eight Lantmännen accessions were grouped based on three different criteria (ploidy, maturity type and population type). A low but significant differentiation was obtained among diploids and tetraploids, accounting for 1.1% of the total variation ($P = 0.02$). However, no differentiation was observed between cultivars and synthetic populations as well as between different maturity groups. Even though there was significant structure distinguishing the accessions, the variation explained by each grouping criteria was very low. The highest variation explained by grouping the accessions was between the NordGen population types, accounting for 1.8 % of the total variation.

The F_{ST} based pairwise differentiation matrix between accessions is illustrated in **Figure 2**, where the darker color indicates higher differentiation between the accessions. As shown in this figure, seven accessions had highly significant ($P < 0.001$) pairwise differentiation with average F_{ST} values ranging from 0.07 to 0.11. These are Swedish (NGB1420 and NGB1009), Russian (NGB24176) and Norwegian (NBG15558 and NGB15623) wild accessions, as well as the Danish cultivars (NGB11605 and NGB11608), all of which were obtained from NordGen. The pairwise F_{ST} between two Lantmännen accessions, SW RK1166 (a synthetic population) and 'Vicky' (a cultivar) was 0, thus suggesting that they are closely related and might have been developed based on the same source. Additionally, the differentiation between SW RK1166 and SW RK1158 (synthetic populations from Lantmännen) was also 0, thereby suggesting that they too might have been developed from the same source. Interestingly, 'Vicky', SW RK1166 and SWRK1158 are three of the four known tetraploids included in this study. The fourth tetraploid accession (SWA 1675209) showed slightly higher differentiation from these three tetraploids.

A matrix illustrating the pairwise F_{ST} values between groups of accessions according to their origin (**Figure 3A**) shows a low differentiation between accessions from Lantmännen and those from Denmark or Finland (F_{ST} = 0.009 and 0.006, respectively). The differentiation between Finland and Denmark was also low (F_{ST} = 0.010), which is interestingly half the value of the differentiation between Denmark and Norway or Sweden (F_{ST} = 0.022 and 0.023, respectively). Furthermore, the differentiation between Denmark and Sweden or Norway are almost equal to the differentiation between Lantmännen and Sweden or Norway (F_{ST} = 0.020 for both; **Figure 3A**). The pairwise F_{ST} values between population types (**Figure 3B**) showed similarly low differentiation of NordGen cultivars and landrace populations from Lantmännen accessions, whereas the NordGen wild populations showed a relatively higher differentiation from the Lantmännen accessions (F_{ST} = 0.033). Overall, the wild populations are the most differentiated among these groups with average F_{ST} value of 0.023 (**Figure 3B**).

TABLE 2 | Summary of analysis of molecular variance (AMOVA) based on 1000 permutations for (1) all 29 accessions without grouping, and by grouping (2) the 21 accessions from NordGen, and (3) the eight Lantmännen accessions.

Source of variation	Degrees of freedom	Sum of squares	Variance components	Percentage of variation	Fixation indices	Probability (*P*) value
Among accessions	28	3777.030	0.3563 V_a	5.45	F_{ST} = 0.054	V_a and F_{ST} = 0.0000
Among individuals within Accessions	259	15579.36	−5.169 V_b	−7.48	F_{IS} = −0.079	V_b and F_{IS} = 1.0000
Within individuals	288	20301.0	70.49 V_c	102.03	F_{IT} = −0.02	V_c and F_{IT} = 0.9550
Total	575	39657.4	69.08			
[a]Among groups	1	214.51	0.356 V_a	0.51	F_{CT} = 0.005	V_a and F_{CT} = 0.0117
Among accessions within groups	27	3562.51	3.341 V_b	4.82	F_{SC} = 0.048	V_b and F_{SC} = 0.0000
Within accessions	547	35880.36	65.595 V_c	94.66	F_{ST} = 0.053	V_c and F_{ST} = 0.0000
Total	575	39657.39	69.292			
[b]Among groups	4	726.833	0.63 V_a	0.94	F_{CT} = 0.009	V_a and F_{CT} = 0.0185
Among accessions within groups	16	2119.99	3.55 V_b	5.34	F_{SC} = 0.054	V_b and F_{SC} = 0.0000
Within accessions	395	24601.52	62.28 V_c	93.71	F_{ST} = 0.063	V_c and F_{ST} = 0.0000
Total	415	27448.34	66.46			
[c]Among groups	3	717.65	1.186 V_a	1.78	F_{CT} = 0.018	V_a and F_{CT} = 0.0000
Among accessions within groups	17	2129.17	3.182 V_b	4.77	F_{SC} = 0.049	V_b and F_{SC} = 0.0000
Within accessions	395	2460.52	62.282 V_c	93.45	F_{ST} = 0.065	V_c and F_{ST} = 0.0000
Total	415	27448.34	66.650			
[d]Among groups	1	158.31	0.818 V_a	1.08	F_{CT} = 0.010	V_a and F_{CT} = 0.0244
Among accessions within groups	6	557.61	0.932 V_b	1.23	F_{SC} = 0.012	V_b and F_{SC} = 0.0000
Within accessions	152	11293.45	74.30 V_c	97.70	F_{ST} = 0.023	V_c and F_{ST} = 0.0000
Total	159	12009.40	76.05			
[e]Among groups	1	97.36	−0.076 V_a	−0.10	F_{CT} = −0.001	V_a and F_{CT} = 0.5796
Among accessions within groups	6	618.59	1.440 V_b	1.90	F_{SC} = 0.019	V_b and F_{SC} = 0.0000
Within accessions	152	11293.45	74.299 V_c	98.20	F_{ST} = 0.018	V_c and F_{ST} = 0.0000
Total	159	12009.41	75.662			
[f]Among groups	1	128.12	0.377 V_a	0.50	F_{CT} = 0.005	V_a and F_{CT} = 0.0880
Among accessions within groups	6	587.84	1.184 V_b	1.56	F_{SC} = 0.016	V_b and F_{SC} = 0.0000
Within accessions	152	11293.45	74.299 V_c	97.94	F_{ST} = 0.020	V_c and F_{ST} = 0.0000
Total	159	12009.41	75.859			

[a] *The 29 accessions grouped into two according to source: NordGen and Lantmännen.*
[b] *The 21 NordGen accessions grouped into five groups according to country of origin.*
[c] *The 21 NordGen accessions grouped into four population types: breeding populations, cultivars, landraces and wild.*
[d] *The eight Lantmännen accessions grouped into two ploidy groups: diploids and tetraploids.*
[e] *The eight Lantmännen accessions grouped into two population types: cultivars and synthetic populations.*
[f] *The eight Lantmännen accessions grouped into two maturity groups: early-middle late and late.*

Cluster Analysis and Discriminant Analysis of Principal Components

A neighbor joining (NJ) tree (Saitou and Nei, 1987) constructed based on the Nei's unbiased genetic distance between each pair of the 29 accessions revealed small and large clusters as well as solitary accessions (**Figure 4**). Five of the six wild accessions were clearly separated from the cultivated groups with four of them forming a separate cluster despite representing different countries (Norway, Russia and Sweden). However, a wild accession from Finland (NGB14448) was clustered with a landrace accession from Finland (NGB1142) and two Lantmännen diploid accessions (SW YNGV and SWÅ RK0700). The NJ tree also clearly depicted a close clustering of the Lantmännen tetraploid accessions while the four diploids were spread across three clusters. With regard to country of origin, the Danish accessions (which are all cultivars except one) were closely

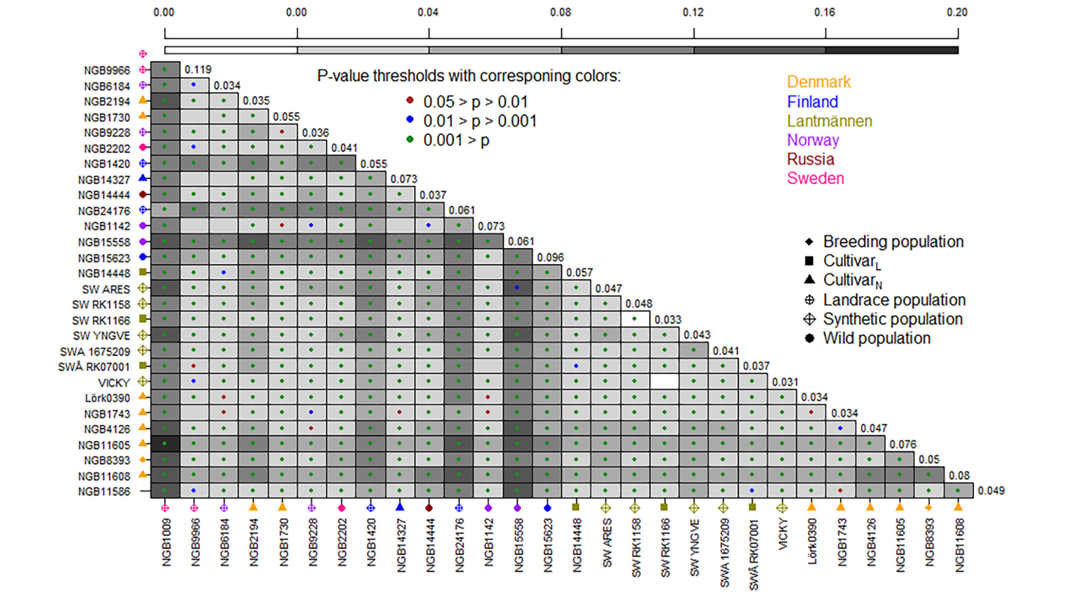


FIGURE 2 | A graphical illustration of pairwise F_{ST} values between populations. The range of background color from light to dark corresponds to low to high F_{ST} values. The color of a dot in each square indicates the level of significance with brown for $0.05 > p > 0.01$, blue for $0.01 > p > 0.001$ and green for $p < 0.001$. The average F_{ST} values are given in the diagonal. The accessions are given on x- and y-axes, with corresponding label shape and color representing population type and country of origin, respectively (see figure keys). Cultivar$_L$ = cultivar from Lantmännen; Cultivar$_N$ = Cultivar from NordGen.

clustered together with the exception that two accessions (a cultivar and a breeding population) were placed in a separate but closely related cluster. Interestingly, the Lantmännen tetraploid accessions showed higher genetic similarity with the Danish accessions than with those from other countries, and were closely clustered with the NGB11608 (a breeding population). The diversity among the landrace accessions was also displayed in the NJ tree, where even those from the same country were placed far apart in different clusters.

The genetic distance between the 288 individual genotypes calculated based on the Tamura-Nei method was used for NJ cluster analysis and resulted in eight clusters (**Figure 5**). The tree exhibited clear differences between the accessions in terms of the within accession genetic diversity. Interestingly, all genotypes of the four known tetraploid accessions (SW RK1158, SW RK1166, SWA 1675209 and 'Vicky') were clustered in cluster-1 whereas the genotypes of the four known diploids (LÖRK0390, SWÅ RK07001, SW ARES and SW YNGVE) were spread across three to five clusters. With the exception of five accessions representing wild populations, all NordGen accessions had members in more than one cluster. All genotypes of the NordGen wild accessions (NGB15558 and NGB15623 from Norway; NGB24176 from Russia; and NGB1009 and NGB1420 from Sweden) were clustered in cluster-8 with the exception of accession NGB14448 (from Finland) that had members in clusters-1, -6 and -8 (**Figure 5**).

A Nei's unbiased genetic distance based two-dimensional PCoA plot depicting the relationship between the 29 accessions (**Figure 6**) explained 53.7% and 19.5% of the total variation in the first and second principal coordinates, respectively. The PCoA showed a close relationship among the majority of the accessions forming a major cluster. However, the wild populations were clearly separated from the rest with the exception of the Norwegian wild population NGB15623 that was placed close to the major cluster. Interestingly, the two Swedish (NGB1009 and NGB1420) and the Russian (NGB24176) wild populations were placed close to each other while the Norwegian wild population (NGB15558) was separated from them along the second principal coordinate. Among the seven Danish accessions, two cultivars (NGB11586 and NGB11605) and a breeding population (NGB11608) were slightly separated from the main cluster along with a Finnish cultivar (NGB14444).

The DAPC analysis (**Figure 7**) accounted for 76.9% of the cumulative variance and separated the 288 individuals into three clusters (the cluster allocation of members of each accession is given in **Table 1**). All individual genotypes of four accessions representing wild populations from Sweden (NGB1009 and NGB1420), Norway (NGB15558) and Russia (NGB24176) clearly separates from the rest and formed Cluster-3. These accessions also had the highest mean F_{ST} values (diagonal values in **Figure 2**) and were clearly separated from the main PCoA cluster

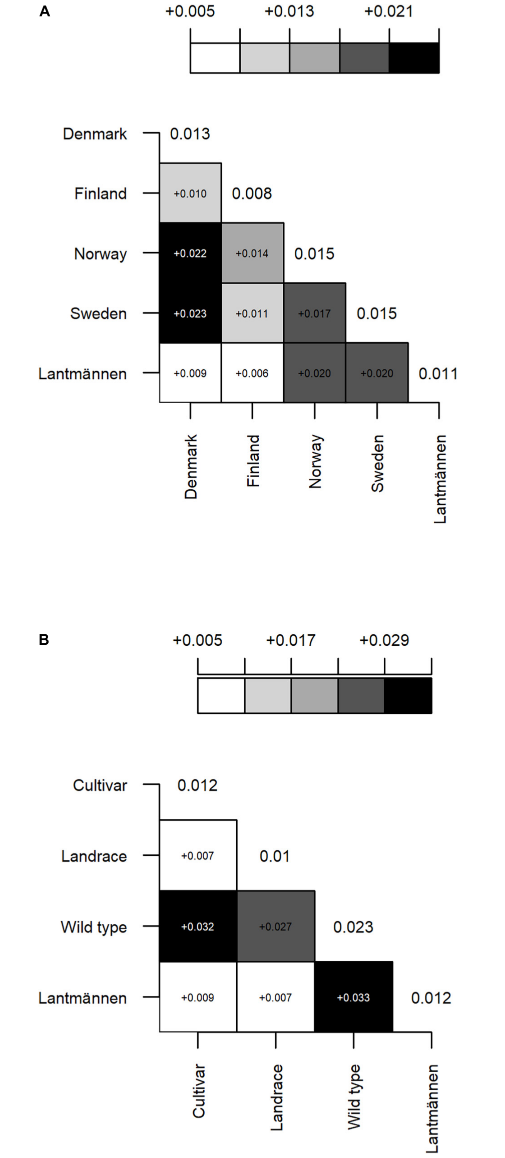


FIGURE 3 | A graphical illustration of pairwise F_{ST} between groups of accessions representing different **(A)** country of origin: Sweden, Norway, Denmark, Finland and Lantmännen, and **(B)** population type: cultivar, landrace and wild populations from NordGen, and Lantmännen accessions. The gradient color intensity corresponds to low to high F_{ST} values (the deeper the color the higher the F_{ST} value). All values are significant at a threshold value of 0.05. F_{ST} value of each pair is shown in the corresponding square. The diagonal values are mean F_{ST} values of each group. Lantmännen is represented as a separate group under country of origin.

(**Figure 5**). The remaining populations were split over cluster-1 and cluster-2.

Loci Deviated From Hardy-Weinberg Equilibrium and Loci Under Selection, and Their Corresponding Mutation Types

The HWE test revealed that 51 out of the 623 SNP loci, showed a highly significant deviation ($P < 0.01$) from HWE. Eighteen of these loci showed excess heterozygosity whereas 33 loci were heterozygote deficient (**Supplementary Table 2**). The translation of the coding sequences containing the 51 SNP loci using their alternative alleles resulted in missense mutations in 28 loci that led to amino acid changes and 1 nonsense mutation that resulted in premature termination of the amino acid sequence (**Supplementary Table 2**). The nonsense mutation is located in the oxysterol binding protein (OSBP)-related protein 1C gene (The UniProt Consortium, 2021), which is involved in the regulation of sterol transportation. The changes in the remaining 22 loci were same-sense mutations.

The examination of the joint distribution of F_{ST} and the heterozygosity at the 623 loci under a non-hierarchical finite island model resulted in 88 loci that had significant F_{ST} P-values (<0.05), and were therefore considered to be under selection. The translation of the coding sequences of genes containing these loci using the alternative SNPs resulted in 42 and two missense and nonsense mutations, respectively (**Supplementary Table 3**). The remaining 44 SNPs did not lead to change in amino acids.

All the alternative allele substitutions that were shown to be under selection and were located within genes coding for PPR proteins resulted in missense mutations. The assessment of the zygosity of the 288 individuals revealed differences in the level of heterozygosity for the PPR proteins between the wild populations and Lantmännen cultivars (**Figure 8**). Among the PPR proteins, the PPR repeat protein 1 marker differs a lot within the 29 accessions but the reference allele seems to be preferred within the Danish breeding population and the Lantmännen synthetic populations. Furthermore, homozygosity of the reference allele was absent for PPR repeat protein marker 2. For PPR repeat protein marker 4, all individuals except two synthetic accessions and five of the Russian wild type individuals are homozygous for the reference allele. The overall level of heterozygosity in the PPR proteins are higher in the four tetraploid accessions comprising three synthetic populations (SW RK1158, SW RK1166, and SWA 1675209) and a cultivar ('Vicky') from Lantmännen.

Within the group of disease resistance proteins, one of the mutations within the HXXD-type acyl-transferase family protein (HXXD-type protein 1 in **Figure 8**) was a missense mutation (**Supplementary Tables 1, 2** and **Figure 8**). For most of the loci within disease resistance genes, one of the two alleles were more dominant, and hence the level of heterozygosity is lower and individual genotypes are mostly homozygous for either the reference or alternative allele. The LRR receptor-like protein 1 marker is the most

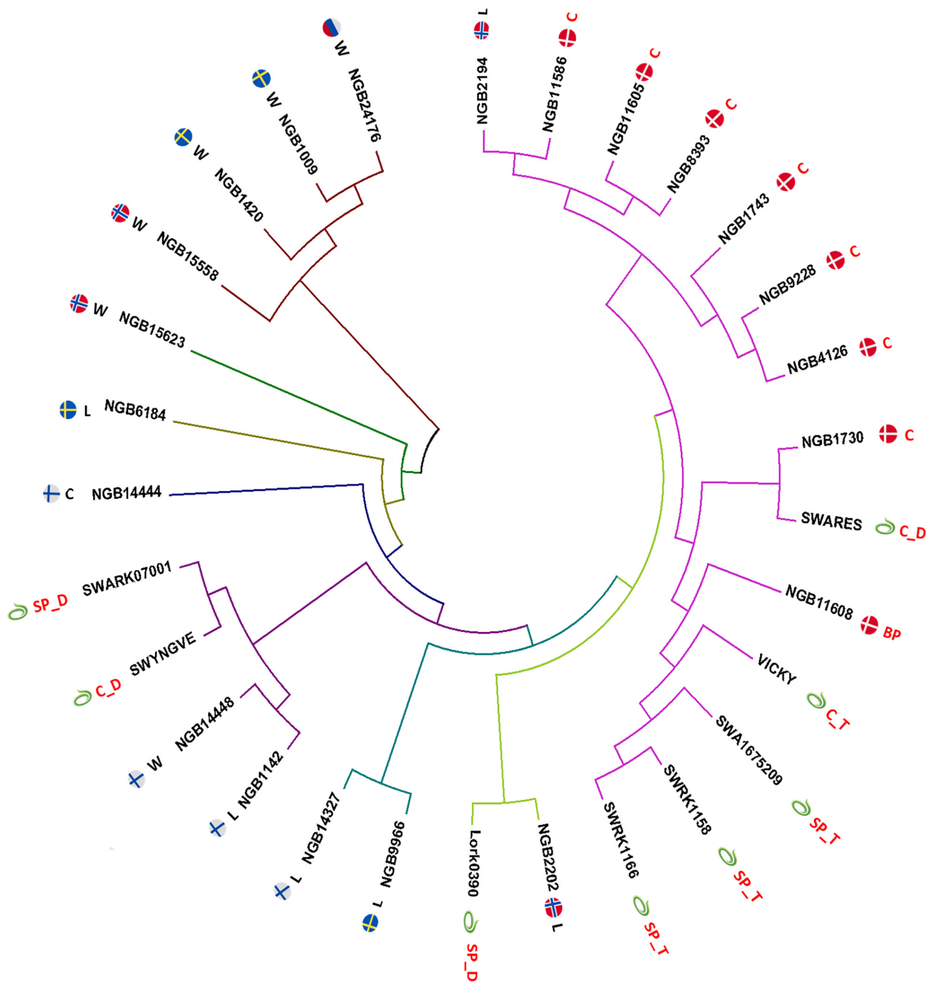


FIGURE 4 | A neighbor joining (NJ) tree constructed based on Nei's unbiased genetic distance values of the 29 accessions with a fan layout and edges scaled to equal length. Each node is labeled with the accession name as well as an icon of the flag of its country of origin or the Lantmännen symbol. The letter abbreviations following the flag corresponds to the population type of the accession with W for wild population, L for landrace, C for cultivar, SP for synthetic population and BP for breeding population. The letter following the underscore (_) signifies the ploidy of the accession (if it is known) with D for diploid and T for tetraploid.

heterozygous loci of the eight. Interestingly, the alternative alleles at the HXXP protein 1 locus resulted in premature termination of the amino acid sequence, and the analysis using the Protein Variation Effect Analyzer (PROVEAN; Choi and Chan, 2015) revealed that this mutation has a deleterious effect on the protein. For this locus, homozygous individuals for alternative allele were preferred (**Figure 8**). Thus, the mutation to alterative alleles in these loci might have noteworthy effects on the genotypes' disease resistance and consequently survival.

DISCUSSION

In the present study, SeqSNP is used as genotyping method for genotyping the target SNP loci. The method also allows *de novo* discovery of new SNPs near the target SNP loci, which resembles genotyping by sequencing (GBS). However, it should be noted that our approach might have excluded the possibility of discovering SNPs in previously uncharacterized genes of potential significance in red clover improvement. The 623 bi-allelic polymorphic markers genotyped across 288 individuals

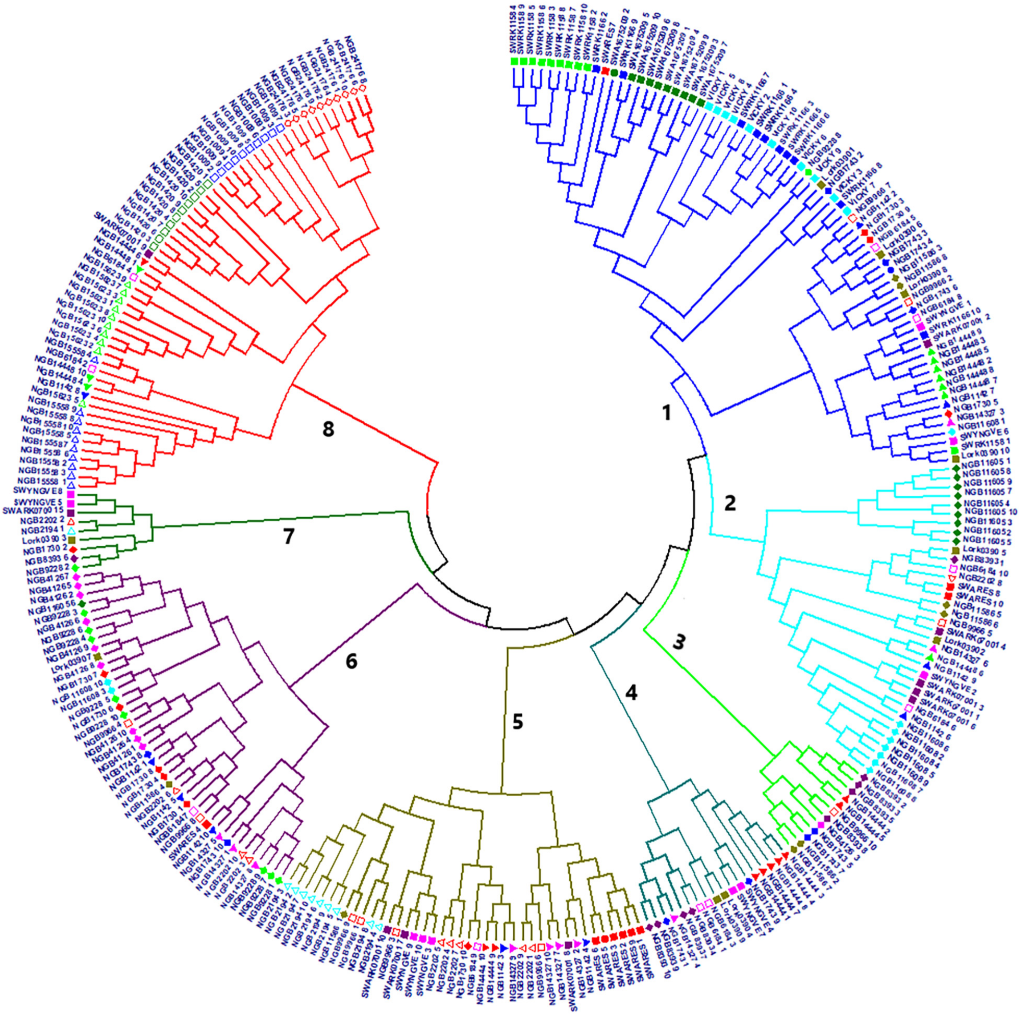


FIGURE 5 | A neighbor joining (NJ) tree of 288 individuals representing 29 accessions based on evolutionary distances computed by the Tamura-Nei method. Individuals sharing the same symbol and color belong to the same accession.

representing the 29 red clover accessions using SeqSNP revealed significant population structure differentiating accessions due to origin and cultivar type. Four wild accessions, two Swedish (NGB1420 and NGB1009), one Russian (NGB24176) and one Norwegian (NGB15558) were consistently grouped together while separated from the other accessions. Additionally, the known tetraploids ('Vicky', SWA 1675209, SW RK1166, and SW RK 1158) grouped together in the NJ trees while the known diploids were dispersed across different clusters together with NordGen accessions. Furthermore, the tetraploids showed similar patterns in zygosity in markers associated with growth and development (**Figure 8**).

F-Statistics, Heterozygosity and Within Population Diversity

The fixation indices measure inbreeding for each locus (Wright, 1949, 1965) and a maximum value of one is attained when all individuals in an accession are homozygous at a locus. Negative values of F_{IS} indicate excess heterozygosity. In the

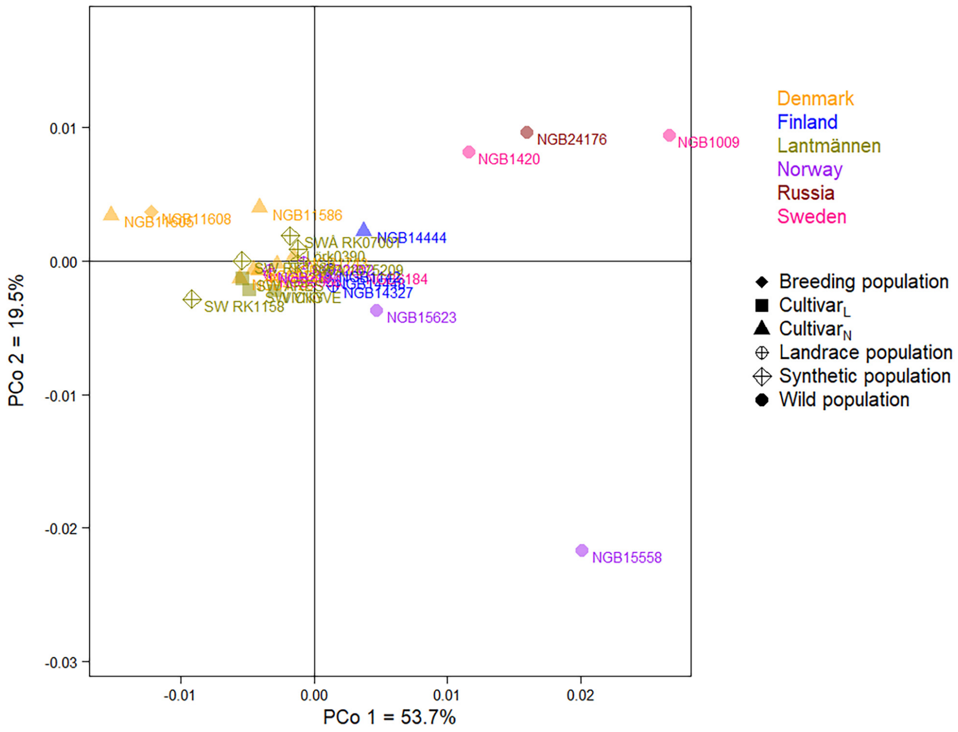


FIGURE 6 | A principal coordinate analysis (PCoA) plot depicting the relationship between the 29 red clover accessions. Accessions with the same font-color belong to the same origin (a country or Lantmännen; see Figure key). Accessions with the same label-shape and color belong to the same population type (see Figure key).

present study, The F_{IS} values ranged from −0.33 to 0.06 with 45% of the accessions having negative values. The mean F_{IS} value (−0.31) of the tetraploids were significantly lower than that of the diploids (−0.01) as expected, because there is higher probability of heterozygosity in tetraploids than in diploids under an outcrossing reproductive system. The F_{IS} values for the known diploids ranged from 0.05 to 0.03 whereas that of the 21 NordGen accessions with unknown ploidy ranged from 0.08 to 0.06 (mean = 0.00). Jones et al. (2020) also reported similar F_{IS} values for red clover ecotypes (diploids) from United Kingdom in their GBS based study. Whereas the ecotypes from central Europe recorded a slightly higher F_{IS}, suggesting the Scandinavian red clover may be more similar to that of United Kingdom than to those of central Europe. In a recent bi-allelic SNP marker-based study by Tsehay et al. (2020) in *Guizotia abyssinica* (a strictly outcrossing diploid species; Geleta and Bryngelsson, 2010), a mean F_{IS} value of 0.13 was reported, which is higher than the values obtained in red clover. This could suggest that red clover is as strictly outcrossing as *G. abyssinica* if not stricter.

The mean values for observed heterozygosity (H_O) for tetraploids (0.35) was higher than that of the diploids (0.21) and accessions with unknown ploidy (0.20). The average observed heterozygosity for European red clover ecotypes in Jones et al. (2020) was 0.25. The F_{IS} and H_O values suggest that all 21 accessions with unknown ploidy obtained from NordGen are most likely diploids. Overall, the heterozygosity and F_{IS} values obtained in this study (**Figure 1A**) are in line with expected values in an outcrossing species (Schoen and Brown, 1991; Huang et al., 2019). Similar to observed heterozygosity, the within accession diversity was higher for the tetraploids (mean H_S = 0.25) when compared to all other accessions (mean H_S = 0.20). According to Öhberg (2008) and Vleugels et al. (2013) tetraploid red clover were shown to have higher persistence than diploid red clover. Furthermore, Ergon et al. (2019) indicated that there were differences in allele frequencies in red clover survivor populations after exposure to certain environmental stresses vis-á-vis the original populations. Generally, they showed that higher genetic diversity within populations resulted in higher chance of survivors following exposure to various selection pressures. Thus, the higher within accession diversity shown by the present study in tetraploids could be a result of persistence and resistance breeding by Lantmännen. Persistence and resistance are two of the major differentiating traits between tetraploid and diploid

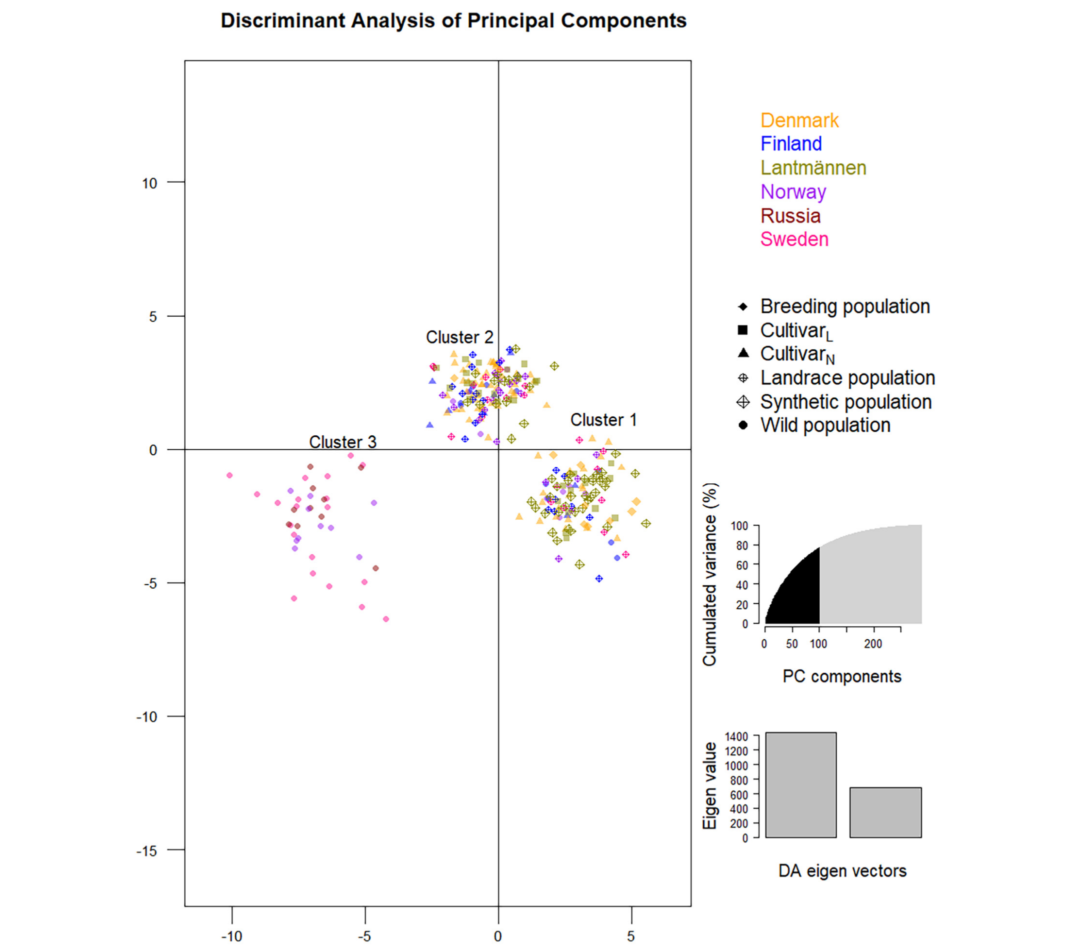


FIGURE 7 | A discriminant analysis of principal components (DAPC) depicting the clustering of the 288 red clover genotypes of the 29 accessions into three clusters. Genotypes sharing the same label-color belong to the same origin (a country or Lantmännen; see Figure key) whereas genotypes sharing the same label-shape belong to the same population type (see Figure key). The graph for the selection of 100 principal components and Eigenvalues of the discriminant analysis are shown on the right.

red clover in an agricultural perspective together with seed yield and dry biomass. The tetraploid red clover produces less seeds due to either a lower number of flower heads or higher levels of embryo abortion compared to diploids (Vleugels et al., 2015). The low seed production is a disadvantage for seed producers but the tetraploid red clover is more competitive for the traits regarding quality fodder such as dry biomass yield, persistence and resistance (Taylor and Quesenberry, 1996b).

It is interesting that the wild accessions had lower average within population diversity than the other groups with four of the six accessions recording the lowest H_S values (**Table 1**). This is likely due to the dominance of wild type alleles over mutant alleles in the wild populations (with mutant alleles having lower frequencies than in the cultivated accessions). This is in line with the generally lower percent polymorphic loci (%PL) obtained in the wild accessions compared to the cultivated ones. The overall mean H_S and H_T (0.21 and 0.22, respectively) in the present study was slightly lower than those reported in Jones et al. (2020) suggesting a slightly lower genetic diversity in Scandinavian red clover as compared to the Central European ones. However, the differences could also be due to differences in the number of markers used and the genomic regions targeted. In the present

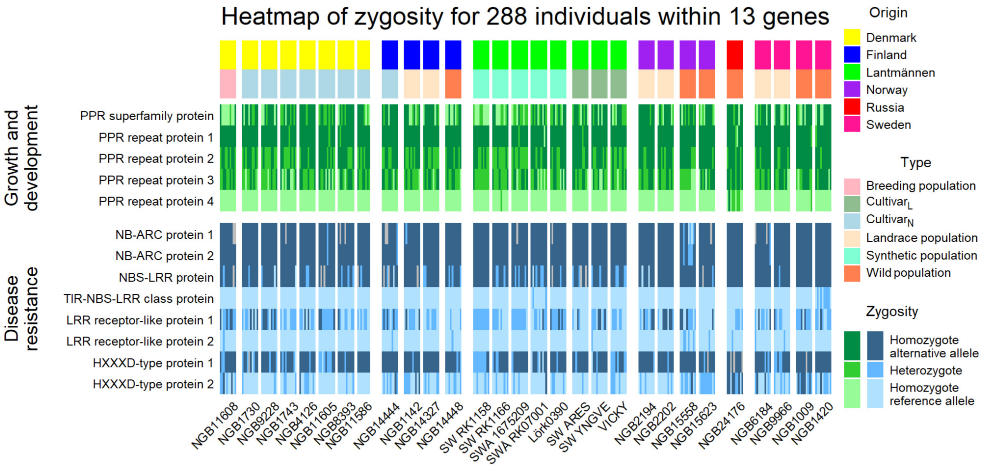


FIGURE 8 | A heatmap illustrating the zygosity level of five PPR proteins, known to play an important role in growth and development, and eight disease resistance proteins in the 288 individuals. The individuals were grouped according to their accession and the top annotation sorted the individuals further into place of origin and type. Gray fields indicate missing genotype for that individual at the particular locus.

study, gene-coding regions were targeted whereas the GBS based study in Jones et al. (2020) was not specifically targeting genes and hence higher genetic variation is likely. The comparison of different population genetics parameters between origin and type-based groups of accessions showed that the difference is very low in general, but it was relatively higher between the NordGen accessions and Lantmännen, which was clearly due to the presence of tetraploids in the latter.

The range of F_{ST} values was larger for accessions from Sweden and Norway (**Figure 1B**) as well as for both wild populations and Lantmännen cultivars (**Figure 1C**). The results agree with the pairwise F_{ST} values for accessions grouped according to origin and cultivar type (**Figures 3A,B**). Sweden and Norway accounted for two-thirds of the wild accessions included in the present study. Presumably, the genetic variance between the wild populations had stronger contribution to the wider range in F_{ST} values obtained for accessions from these two countries, due to restricted gene flow in nature. On the other hand, the wider range in F_{ST} values observed within the Lantmännen group was contributed by the presence of both diploids and tetraploids. There were relatively higher pairwise F_{ST} values between Russia, Sweden and Norway (**Figure 3A**) but low values between the Lantmännen, Finnish and Danish accessions (**Figure 3B**). The accessions representing Denmark are all cultivars except a single breeding population, and most of the Finnish material represents cultivated gene pool, which can explain the lower F_{ST} values.

The Norwegian and Swedish populations, on the other hand, included only wild and landrace accessions explaining the higher F_{ST} values. The effects of wild populations versus cultivars are further noted in **Figure 3C** where the F_{ST} values are lower between the cultivated types and higher between cultivars and wild populations. Furthermore, it is worth noting that the landraces are closer to the cultivars than to the wild populations. Landraces are cultivated types whose genetic diversity is shaped through traditional methods of selection and not through modern breeding programs. It can be hypothesized that different landraces have been developed independently and selection by human has been less stringent (than the case in cultivars), and thus contain a higher level of within population genetic diversity and have higher F_{ST} values compared to the modern cultivated types that include breeding and synthetic populations and cultivars. However, the present study revealed that genetic diversity and population differentiation of all cultivated types are generally similar indicating that modern red clover breeding programs did not lead to loss of genetic diversity and affect population structure. As suggested by Jones et al. (2020) the relatively short period in the cultivation and breeding of red clover could have contributed to similar level of genetic diversity and low differentiation between its wild and cultivated gene pools.

Cluster Analysis, Principal Coordinate Analysis and Discriminant Analysis of Principal Components

As previously discussed, red clover is a strictly outcrossing species, and outcrossing species have a greater gene flow than self-pollinating species due to the role of their pollinators. This increased gene flow between subpopulations reduces the difference between them. From the AMOVA (**Table 2**), 5.4% of the total variation was explained by the variation among accessions, which is slightly higher than that of a strictly outcrossing species *G. abyssinica* (Tsehay et al., 2020). Compared with studies on *Arabidopsis thaliana* (Brennan et al., 2014) and both outcrossing and selfing *Zingiber* (Huang et al., 2019), which generated 81.6%, 65.3% and 91.7% variation among accessions, respectively, the differentiation between red clover accessions are in the lower range. In these examples, the difference between the three outcrossing species is interesting with 65.3% for the outcrossing *Zingiber*, 5.4% for red clover and 4.5% for *G. abyssinica*. *Zingiber* is less strictly outcrossing than red clover or *G. abyssinica*, which can be a reason for the great difference in variation explained among accessions. This research finding suggests that the degree of stringency of the outcrossing reproductive system can be shown through the AMOVA results. However, there are other contributors to the variation between accessions other than the breeding system. One factor that should not be over looked is the effect of sample selection. The low level of population structure presented in this study is in line with the results of Dias et al. (2008), who reported that the vast majority of the variance in red clover were resided within the populations although the level of differentiation was relatively higher in their study. Similarly, Jones et al. (2020) reported a significant but low population differentiation (F_{ST} = 0.076) in their GBS based study that included European and Asian ecotypes and cultivars.

AMOVA analysis was further performed after six separate criteria were applied to group the 29 accessions. Four of the groupings revealed a very low but significant differentiation between the groups. Overall, the variation explained by the groupings was very low with only NordGen accessions grouped by population types and the Lantmännen accessions grouped by ploidy explaining over 1% of the total variance (1.78% and 1.08%, respectively).

The cluster analysis at both accession and individual genotype levels demonstrated the genetic relationship between the accessions used in the present study. The eight Lantmännen accessions displayed interesting pattern at both levels. The four tetraploid accessions are closely related and all accessions and their individual members were placed in the same cluster. The result suggests a narrow genetic base of the tetraploid red clover and hence crossbreeding among them may not lead to a significant improvement of desirable traits. Since tetraploid red clover is a result of successful plant breeding, based on limited tetraploid genotypes developed through chromosome doubling (Sjödin and Ellerström, 1986; Taylor and Quesenberry, 1996b), the obtained results are not unexpected. Hence, further genetic analyses among a larger set of tetraploid accessions need to be conducted to identify genetically distinct groups suitable for crossbreeding. On the other hand, the Lantmännen diploid accessions are diverse and appeared to have close relationship with landraces or cultivars from Finland, Norway, Sweden and Denmark. This indicates that the diploid cultivars and synthetic populations that are currently in use at Lantmännen largely represents the landrace gene pool from the different countries in Scandinavia.

Similar to the cluster analyses, the relatively close relationship between the cultivated accessions was also shown in the PCoA, where they were clustered close to the center of the PCoA bi-plot with a slight separation of the three Danish cultivars (NGB11586, NGB11708, and NGB11605) and one Finnish cultivar (NGB14444). The relatively higher divergence between the wild accessions as compared to the case of cultivated pool is expected due to a limited gene flow among the wild populations. In agreement with the cluster analyses, five of the six wild accessions were separated from the other groups along the principal coordinate 1, which accounted for 54% of the total variation whereas the differentiation between them is clearly observable along the principal coordinate 2 (**Figure 6**). The clear separation of most wild accessions from the cultivated gene pool is mainly the results of human selection, and is in agreement with the results of previous research (Shim and Jørgensen, 2000; Geleta et al., 2007). The case of the Finnish wild accession NGB14448 is interesting. Unlike all other wild accessions, it was closely grouped with cultivated accessions including a Lantmännen cultivar (at accession level), as shown by cluster analysis, PCoA and DAPC. At individual level, its genotypes were placed in different clusters, of which three individuals clustered with the other wild genotypes. This may suggest that the accession is a result of mixture of seeds (accidental or intentional) from cultivated and wild accessions. Hence, it is interesting to investigate this accession further and other similar accessions at NordGen.

Unlike the PCoA bi-plot that was the result of the first two principal coordinates, the DAPC, which described 100 principal components (**Figure 7**), led to three significant clusters (**Table 1**). It shall be noted that the PCoA was analyzed based on Nei's unbiased genetic distance while the principal components of DAPC was based on an Euclidian distance. Hence, the depiction of the PCoA is a representation of the variation between accessions on a more genetic level while Euclidian distance is strictly mathematical explanation of the variation between SNP genotypes. The separation of the first two DAPC clusters is not clear-cut since they are a mixture of different population types and from different origins. The lack of complete differentiation between cluster one and two with regard to population affiliation can be explained by a high level of within population variation (**Figure 1A**). One exception is the known tetraploid SW RK11158 for which all its individuals were located in cluster-2. Interestingly, the other tetraploids also had the majority of their individuals in cluster-2. The third DAPC cluster contained all individuals of four of the six wild populations, NGB1009 (Sweden), NGB1420 (Sweden), NGB15558 (Norway) and NGB24176 (Russia), which were also clustered together in the NJ trees and PCoA bi-plot. The affiliation of populations in the DAPC clusters are further confirmed by the results of the individual NJ tree (**Figure 5**), which showed a similar clustering pattern.

The Differentiation Between Accessions Shown by Markers of Interest

The 623 SNP loci used in the present study are located within the coding sequences of 231 genes, of which 51 loci (8.2%) showed significant deviation from HWE. Because, red clover is an outcrossing species with a gametophytic self-incompatibility system (Taylor, 1982), excess heterozygosity is likely even at neutral loci. However, excess heterozygosity coupled with nonsense mutation, such as the case at locus ASHM01029522C

(**Supplementary Table 2**), within the oxysterol binding protein (OSBP)-related protein 1C gene (The UniProt Consortium, 2021) regulating the sterol transportation, could be due to heterozygote advantage. On the other hand, heterozygote deficiency in plants with a self-incompatibility system, strongly suggests that the loci may have fitness values in the form of homozygote advantage or linked to such loci. Two-thirds of the 51 loci were heterozygote deficient, and further investigation of their potential roles in red clover fitness is therefore recommended. Li et al. (2019) reported that 28% of the SNPs identified using publicly available red clover RNAseq data accounted for missense mutations. As a comparison, the present study found that 58% of the SNPs gave rise to missense mutations. Among the SNPs shown to be under selection pressure, 48% resulted in missense mutations whereas missense mutations accounted for 55% of the loci deviated from HWE. There are multiple reasons for the differences in the proportion of missense mutations between the studies, which include differences in sequencing methods or the stringency of steps to determine missense mutations. The present study used DNA sequences of known coding regions at specific loci across 29 accessions while Li et al. (2019) used a combination of targeted genomic amplicon sequences of 72 genotypes of a breeding population and RNA-seq data from three genotypes of different cultivars. Furthermore, the present study used the longest open reading frame on each protein sequence with no additional penalties or weights applied in scoring. However, Li et al. (2019) used a software developed to score all types of mutations and might have been more stringent in calling missense mutations.

Among the 231 protein coding genes containing the 623 bi-allelic SNPs, those coding for the pentatricopeptide repeat (PPR) family proteins and proteins related to disease resistance are among the most frequent groups. The PPR protein family is one of the largest protein families in land plants and are known to regulate RNA expression and have an effect on respiration, photosynthesis, growth and development, and responses to environmental stresses (Barkan and Small, 2014). The zygosity pattern of markers located within the five PPR proteins and eight disease resistance proteins showed that there was a difference in heterozygosity between the Lantmännen and NordGen accessions. This is in agreement with the levels of H_O (**Figure 1C**) where Lantmännen populations were shown to have a higher level of heterozygosity.

The level of heterozygosity was higher in the PPR protein markers than in the disease resistance proteins and the SNP loci in the disease resistance genes were more often homozygous. Interestingly, the nonsense mutation (HXXXD-type protein 1 in **Figure 8**) showed that the majority of individuals were homozygous for the alternative allele, which was higher than expected based on the allele frequency. Using the prediction tool on protein function, PROVEAN (Choi and Chan, 2015), the premature stop codon induced by the alternative allele had significant effect on the protein function. It would therefore be interesting to further study the effect of the end terminal of the HXXXD-type protein and the response to pathogens. However, the result of selection based on HWE and F_{ST} could be due to association with another loci; an effect referred to as "hitch-hiking" by Smith and Haigh (1974). If there is linkage disequilibrium between one neutral locus and another locus under selection, the neutral allele will be inherited with the allele under selection thus appears to be under selection as per HWE an F_{ST} assumptions. De Vega et al. (2015) found, however, a low level of linkage disequilibrium between red clover markers, thus decreasing the possibility of a hitch-hiking locus.

CONCLUSION

This research revealed the population structure of red clover accessions gathered from the Nordic countries based on SNPs chosen to represent variation in genes that probably regulate traits important in agriculture. The results show a significant differentiation among accessions from different countries as well as population types, even though it represented a small proportion of the total genetic variation. The inbreeding coefficient was significantly lower whereas observed heterozygosity and within population diversity were significantly higher in tetraploids than in diploids. This is in line with general expectation in a species with outcrossing reproductive system. The present study revealed high genetic similarity among the tetraploid accessions, suggesting a narrow genetic basis of tetraploid red clover being used at Lantmännen. A comprehensive genetic study, incorporating all available tetraploid accessions at Lantmännen and elsewhere, would reveal the overall genetic diversity of tetraploid red clover, which facilitates an adequate comparison with the diploids, as well as designing effective breeding programs. The comparative population genetics analysis of the tetraploids and diploids suggests that the NordGen red clover accessions included in the present study are diploids, as expected based on the background information. The comparison of the wild populations (ecotypes) with the cultivated ones showed a relatively lower within population diversity in the wild populations reflecting differences in allele frequency between the two groups, likely due to the dominance of wild type alleles across loci in the ecotypes. Most, but not all, wild red clover populations were distinctly separated from the cultivated gene pool. In relation to this, the case of wild accession NGB14448 that had highly similar genetic profile with that of some cultivated accessions needs further investigation, to rule out any misclassification at NordGen and to shed more light on the overall genetic relationship between the wild and cultivated red clover gene pool. Interestingly, all cultivated diploid accessions had similar levels of genetic diversity. Hence, modern red clover breeding programs did not lead to significant loss of genetic diversity. This study demonstrates how the breeding material used today reflects the red clover genetic diversity in the Nordic countries and provides the breeders with the knowledge needed to develop genomic breeding tools for red clover. Future research is needed using a greater number of both accessions and individuals together with phenotypic data to obtain a definitive conclusion about the genetic diversity in Nordic red clover and beyond.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**. All raw sequences are available at SRA, BioProject PRJNA765476.

AUTHOR CONTRIBUTIONS

MG secured the funding with RO and CH as co-applicants. JO and MG selected and ordered the seed material from NordGen and Lantmännen Seed, analyzed the data, and compiled the results with input from RO and CH. JO was responsible for the planting and taking care of the plant material in the greenhouse, together with MG and CH sampled the plant material, performed the SNP mining with the assistance of MG, and wrote the manuscript. All authors contributed to the design of the study, provided inputs, revised the manuscript, and approved its final version.

FUNDING

This work was funded by SLU Grogrund – Centre for Breeding of Food Crops, Swedish University of Agricultural Sciences.

ACKNOWLEDGMENTS

We would like to thank NordGen for providing red clover accessions and for sharing their knowledge about treating poorly germinating seeds. We also would like to thank Lantmännen Seed for providing red clover accessions. We acknowledge Linda Öhlund (Lantmännen Seed) for providing information useful for selecting suitable germplasm for this study. Thanks to Mohammed Elsafy for his help in planting and taking care of the plants in a greenhouse.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.748750/full#supplementary-material

REFERENCES

Åkerlund, H. (2008). *Studie av Introduktionen av NorFor Plan För Foderstatsberäkning Till Mjölkkor I Sverige.* Ph.D. thesis. Uppsala (SLU): Swedish University of Agricultural sciences.

Asci, O. O., Acar, Z., Ayan, I., BaŞaran, U., and Mut, H. (2011). Effect of pretreatments on seed germination rate of red clover (*Trifolium pratense* L.) populations. *Afr. J. Agric. Res.* 6, 3055–3060.

Barkan, A., and Small, I. (2014). Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant Biol.* 65, 415–442. doi: 10.1146/annurev-arplant-050213-040159

Bengtsson, A. (1961). "Klöverröta och dess motverkande. Resultat från försök i mellansverige," in *Kungliga Lantbrukshögskolan och Statens lantbruksförsök*, (Uppsala: Särtryck och småskrifter), 1–12.

Brennan, A. C., Méndez-Vigo, B., Haddioui, A., Martínez-Zapater, J. M., Picó, F. X., and Alonso-Blanco, C. (2014). The genetic structure of *Arabidopsis thaliana* in the south-western mediterranean range reveals a shared history between North Africa and Southern Europe. *BMC Plant Biol.* 14:17. doi: 10.1186/1471-2229-14-17

Choi, Y., and Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31, 2745–2747. doi: 10.1093/bioinformatics/btv195

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Cook, D. R. (1999). Medicago truncatula - a model in the making!: commentary. *Curr. Opin. Plant Biol.* 2, 301–304. doi: 10.1016/S1369-5266(99)80053-3

De Vega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, Å, et al. (2015). Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Sci. Rep.* 5:17394. doi: 10.1038/srep17394

Dias, P. M. B., Julier, B., Sampoux, J.-P., Barre, P., and Dall'Agnol, M. (2008). Genetic diversity in red clover (*Trifolium pratense* L.) revealed by morphological and microsatellite (SSR) markers. *Euphytica* 160, 189–205. doi: 10.1007/s10681-007-9534-z

Ergon, Å., Skøt, L., Sæther, V. E., and Rognli, O. A. (2019). Allele frequency changes provide evidence for selection and identification of candidate loci for survival in red clover (*Trifolium pratense* L.). *Front. Plant Sci.* 10:718. doi: 10.3389/fpls.2019.00718

Evans, A. M. (1954). The production and identification of polyploids in red clover, white clover and Lucerne. *New Phyt.* 54, 149–162. doi: 10.1111/j.1469-8137.1955.tb06169.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Excoffier, L., Hofer, T., and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* 103, 285–298. doi: 10.1038/hdy.2009.74

Fischer, K., and Röös, E. (2018). Controlling sustainability in Swedish beef production: outcomes for farmers and the environment. *Food Ethics* 2, 39–55. doi: 10.1007/s41055-018-0027-7

Frugoli, J., and Harris, J. (2001). Medicago truncatula on the move! *Plant Cell* 13, 458–465. doi: 10.2307/3871399

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv* [Preprint] arXiv:1207.3907,

Geleta, M., and Bryngelsson, T. (2010). Population genetics of self-incompatibility and developing self-compatible genotypes in *Guizotia abyssinica* (L. f.) Cass. (*Asteraceae*). *Euphytica* 176, 417–430. doi: 10.1007/s10681-010-0184-1

Geleta, M., Bryngelsson, T., Bekele, E., and Dagne, K. (2007). Comparative analysis of genetic relationship and diagnostic markers of several taxa of *Guizotia* Cass. (*Asteraceae*) as revealed by AFLPs and RAPDs. *Plant Syst. Evol.* 265, 221–239. doi: 10.1007/s00606-007-0521-6

Goudet, J. (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* 5, 184–186. doi: 10.1111/j.1471-8286.2004.00828.x

Herrmann, D., Boller, B., Studer, B., Widmer, F., and Kölliker, R. (2008). Improving persistence in red clover: insights from QTL analysis and comparative phenotypic evaluation. *Crop Sci.* 48, 269–277. doi: 10.2135/cropsci2007.03.0143

Huang, R., Chu, Q. H., Lu, G. H., and Wang, Y.-Q. (2019). Comparative studies on population genetic structure of two closely related selfing and outcrossing *Zingiber* species in Hainan Island. *Sci. Rep.* 9:17997. doi: 10.1038/s41598-019-54526-y

Istvánek, J., Jaros, M., Krenek, A., and Řepková, J. (2014). Genome assembly and annotation for red clover (*Trifolium pratense*; Fabaceae). *Am J Bot.* 101, 327–337. doi: 10.3732/ajb.1300340

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11:94. doi: 10.1186/1471-2156-11-94

Jones, C., De Vega, J., Lloyd, D., Hegarty, M., Ayling, S., Powell, W., et al. (2020). Population structure and genetic diversity in red clover (*Trifolium pratense* L.) germplasm. *Sci. Rep.* 10:8364. doi: 10.1038/s41598-020-64989-z

Kamvar, Z. N., López-Uribe, M. M., Coughlan, S., Grünwald, N. J., Lapp, H., and Manel, S. (2017). Developing educational resources for population genetics in R: an open and collaborative approach. *Mol. Ecol. Resour.* 17, 120–128. doi: 10.1111/1755-0998.12558

Kilsgård, S. (2015). *Faktorer Som Påverkar Tillväxt Hos Kalvar Under Mjölkutfodringsperioden.* Ph.D. thesis. Uppsala (SLU): Swedish University of Agricultural Sciences.

Kjærgaard, T. (2003). A plant that changed the world: the rise and fall of clover 1000-2000. *Landsc. Res.* 28, 41–49. doi: 10.1080/01426390306531

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, W., Riday, H., Riehle, C., Edwards, A., and Dinkins, R. (2019). Identification of single nucleotide polymorphism in red clover (*Trifolium pratense* L.) using targeted genomic amplicon sequencing and RNA-seq. *Front. Plant Sci.* 10:1257. doi: 10.3389/fpls.2019.01257

Merkenshlager, F. (1934). Migration and distribution of red clover in Europe. *Herb. Rev.* 1934, 88–92.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York, NY: Columbia University Press. doi: 10.7312/nei-92038

Öhberg, H. (2008). *Studies Of The Persistence Of Red Clover Cultivars In Sweden.* Ph.D. Dissertation. Umeå (SLU): Swedish University of Agricultural Sciences.

Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J. P., et al. (2013). Advances in plant gene-targeted and functional markers: a review. *Plant Methods* 9, 6. doi: 10.1186/1746-4811-9-6

R Core Team (2013). *R: A Language And Environment For Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. doi: 10.1093/oxfordjournals.molbev.a040454

Schoen, D. J., and Brown, A. H. (1991). Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proc. Natl. Acad. Sci. U.S.A.* 88, 4494–4497. doi: 10.1073/pnas.88.10.4494

Shim, S. I., and Jørgensen, R. B. (2000). Genetic structure in cultivated and wild carrots (*Daucus carota* L.) revealed by AFLP analysis. *Theor. Appl. Genet.* 101, 227–233. doi: 10.1007/s001220051473

Sjödin, J., and Ellerström, S. (1986). "Autopolyploid forage crops," in *Research and Results in Plant Breeding, Svalöf 1886–1986*, ed. G. Olsson (Stockholm: LTs Förlag), 102–113.

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35. doi: 10.1017/S0016672300014634

Smith, R. R., Taylor, N. L., and Bowley, S. R. (1985). "Red clover," in *Clover Science and Technology*, ed N. Taylor (Hoboken, NJ: John Wiley & Sons, Ltd), 457–470. doi: 10.2134/agronmonogr25.c19

Sturz, A. V., Christie, B. R., Matheson, B. G., and Nowak, J. (1997). Biodiversity of endophytic bacteria which colonize red clover nodules, roots, stems and foliage and their influence on host growth. *Biol. Fertil. Soils* 25, 13–19. doi: 10.1007/s003740050273

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526. doi: 10.1093/oxfordjournals.molbev.a040023

Taylor, N. L. (1982). Stability of S alleles in a double cross hybrid of red clover. *Crop Sci.* 22, 1222–1225. doi: 10.2135/cropsci1982.0011183X002200060032x

Taylor, N. L., and Quesenberry, K. H. (1996a). "Historical perspectives," in *Red Clover Science (Current Plant Science And Biotechnology In Agriculture)*, eds N. L. Taylor and K. H. Quesenberry (Dordrecht: Springer Netherlands), 1–10. doi: 10.1007/978-94-015-8692-4_1

Taylor, N. L., and Quesenberry, K. H. (1996b). "Tetraploid red clover," in *Red Clover Science (Current Plant Science And Biotechnology In Agriculture)*, eds N. L. Taylor and K. H. Quesenberry (Dordrecht: Springer Netherlands), 161–169. doi: 10.1007/978-94-015-8692-4_13

The UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi: 10.1093/nar/gkaa1100

Thilakarathna, M. S., Papadopoulos, Y. A., Grimmett, M., Fillmore, S. A. E., Crouse, M., and Prithiviraj, B. (2017). Red clover varieties show nitrogen fixing advantage during the early stages of seedling development. *Can. J. Plant Sci.* 98: 3. doi: 10.1139/cjps-2017-0071

Tsehay, S., Ortiz, R., Johansson, E., Bekele, E., Tesfaye, K., Hammenhag, C., et al. (2020). New transcriptome-based SNP markers for noug (*Guizotia abyssinica*) and their conversion to KASP markers for population genetics analyses. *Genes* 11:1373. doi: 10.3390/genes11111373

United States Environmental Protection Agency (2006). *Prevention, Pesticides and Toxic Substances*. Washington DC: EPA.

van Tienderen, P. H., de Haan, A. A., van der Linden, C. G., and Vosman, B. (2002). Biodiversity assessment using markers for ecologically important traits. *Trends Ecol. Evol.* 17, 577–582. doi: 10.1016/S0169-5347(02)02624-1

Vleugels, T., Cnops, G., and van Bockstaele, E. (2013). Screening for resistance to clover rot (*Sclerotinia* spp.) among a diverse collection of red clover populations (*Trifolium pratense* L.). *Euphytica* 194, 371–382. doi: 10.1007/s10681-013-0949-4

Vleugels, T., Roldán-Ruiz, I., and Cnops, G. (2015). Influence of flower and flowering characteristics on seed yield in diploid and tetraploid red clover. *Plant Breed.* 134, 56–61. doi: 10.1111/pbr.12224

Wright, S. (1931). Evolution in mendelian populations. *Genetics* 16, 97–159. doi: 10.1093/genetics/16.2.97

Wright, S. (1949). The genetical structure of populations. *Ann. Eugen.* 15, 323–354. doi: 10.1111/j.1469-1809.1949.tb02451.x

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* 19:395. doi: 10.1111/j.1558-5646.1965.tb01731.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

II

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fpls.2022.997860

Discovering candidate SNPs for resilience breeding of red clover

Johanna Osterman*, Cecilia Hammenhag, Rodomiro Ortiz and Mulatu Geleta

Department of Plant Breeding, Swedish University of Agricultural Sciences, Lomma, Sweden

OPEN ACCESS

EDITED BY
Jinyoung Y. Barnaby,
United States Department of Agriculture (USDA), United States

REVIEWED BY
João Ricardo Bachega Feijó Rosa,
Federal University of Viçosa, Brazil
Tudor Borza,
Dalhousie University, Canada

*CORRESPONDENCE
Johanna Osterman
johanna.osterman@slu.se

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 19 July 2022
ACCEPTED 07 September 2022
PUBLISHED 28 September 2022

CITATION
Osterman J, Hammenhag C, Ortiz R and Geleta M (2022) Discovering candidate SNPs for resilience breeding of red clover. *Front. Plant Sci.* 13:997860. doi: 10.3389/fpls.2022.997860

Red clover is a highly valuable crop for the ruminant industry in the temperate regions worldwide. It also provides multiple environmental services, such as contribution to increased soil fertility and reduced soil erosion. This study used 661 single nucleotide polymorphism (SNP) markers via targeted sequencing using seqSNP, to describe genetic diversity and population structure in 382 red clover accessions. The accessions were selected from NordGen representing red clover germplasm from Norway, Sweden, Finland and Denmark as well as from Lantmännen, a Swedish seed company. Each accession was represented by 10 individuals, which was sequenced as a pool. The mean Nei's standard genetic distance between the accessions and genetic variation within accessions were 0.032 and 0.18, respectively. The majority of the accessions had negative Tajima's D, suggesting that they contain significant proportions of rare alleles. A pairwise F_{ST} revealed high genetic similarity between the different cultivated types, while the wild populations were divergent. Unlike wild populations, which exhibited genetic differentiation, there was no clear differentiation among all cultivated types. A principal coordinate analysis revealed that the first principal coordinate, distinguished most of the wild populations from the cultivated types, in agreement with the results obtained using a discriminant analysis of principal components and cluster analysis. Accessions of wild populations and landraces collected from southern and central Scandinavia showed a higher genetic similarity to Lantmännen accessios. It is therefore possible to link the diversity of the environments where wild populations were collected to the genetic diversity of the cultivated and wild gene pools. Additionally, least absolute shrinkage and selection operator (LASSO) models revealed associations between variation in temperature and precipitation and SNPs within genes controlling stomatal opening. Temperature was also related to kinase proteins, which are known to regulate plant response to temperature stress. Furthermore, the variation between wild populations and cultivars was correlated with SNPs within genes regulating root development. Overall, this study comprehensively investigated Nordic European red clover germplasm, and the results provide forage breeders with valuable information for further selection and development of red clover cultivars.

KEYWORDS

pool-seq, phylogenetic tree, DAPC, LASSO, bioclimatic variables, red clover

Introduction

Red clover is a perennial forage legume that grows in temperate regions worldwide. Due to its high protein content, it is considered an important crop for the ruminant industry (Smith et al., 1985; Taylor and Quesenberry, 1996a). In addition to its great nutritional value, red clover provides several important ecosystem services. Due to its symbiotic relationship with nitrogen-fixing bacteria (Sturz et al., 1997; Thilakarathna et al., 2017), red clover increases soil fertility. Compared to alfalfa and white clover, which have similar symbiotic relationships with the *Rhizobium* bacteria, red clover is more efficient at nitrogen fixation (Dhamala et al., 2017). As a perennial crop, red clover also contributes to soil carbon sequestration, reduces soil erosion during the winter, and suppresses weeds (McKenna et al., 2018). However, persistence is generally low in red clover, which adversely affects its overall performance as a forage crop (Taylor and Quesenberry, 1996b). Red clover is a diploid species ($2n = 2x = 14$); however, tetraploid ($2n = 4x = 28$) cultivars have been developed through chromosome doubling techniques (Taylor and Quesenberry, 1996c). The tetraploid red clover cultivars generally have a higher green biomass yield as well as a higher persistence and resilience than the diploids (Öhberg, 2008). However, their seed yield is generally lower than that of the diploids because of their flower anatomy and a higher rate of embryo abortion (Amdahl et al., 2016).

The development of modern DNA marker-based plant breeding techniques for red clover is lagging behind, despite its economic and ecological significance, although it has been picking up pace in recent years. For instance, the publication of its reference genome (De Vega et al., 2015) has facilitated the discovery of quantitative trait loci (QTL) for various traits, and the “mining” of single nucleotide polymorphism (SNP) markers (Herrmann et al., 2008; Ergon et al., 2019; Li et al., 2019). In two recent papers on population genetics, SNPs were used to assess the population structure in individually genotyped red clover (Jones et al., 2020; Osterman et al., 2021). Jones et al. studied 75 accessions from Europe, Asia, and Iberia where 70 were wild populations and five were commercially available breeding populations. They found that the population structure of red clover is highly correlated with its geographical location and associated climatic conditions. Osterman et al. focused more on the genetic differences between accessions representing different populations and found that, for instance, wild populations were clearly differentiated from cultivated populations. Both studies noted the effect of the outcrossing nature of red clover in the overall higher heterozygosity which decreases the levels of genetic differentiation. Since red clover is a strictly outcrossing species, genetic research should ideally be performed at a population level. Currently, it is quite expensive to sequence an adequate number of individuals within each population for a comprehensive genetic analysis of multiple populations. With a method that is generally referred to as Pool-seq (Schlötterer et al., 2014), individuals can be pooled and sequenced simultaneously using different next-generation sequencing (NGS) methods. SeqSNP is a targeted genotype by sequencing method for genotyping known SNPs, which is also amenable to *de novo* discovery of SNPs located close to the target SNP positions (Osterman et al., 2021). Hence, Pool-seq via SeqSNP is an NGS method in which individuals are sampled, pooled, and sequenced together, targeting known SNP loci. The target SNPs can be selected from the available SNP databases or developed through allele-mining approaches based on existing genomic resources.

SNP markers are codominant single nucleotide markers that have been widely used in various applications, including genomic selection (GS; Heffner et al., 2009) and marker-assisted selection (MAS; Lande and Thompson, 1990). Compared to the phenotype-based selection, these two breeding methods are quicker and can facilitate the development of high-yielding cultivars that are resilient and nutritious within a shorter period of time. Gene-specific SNP markers are preferred over SNPs in other genomic regions since they are more likely to be associated with genes that regulate desirable traits (Poczai et al., 2013). Hence, genes that are highly desirable from the viewpoint of plant breeding can be targeted for genetic diversity analyses. This will enable the determination of suitable genetic resources that could be used in plant breeding programs. Because genetic similarity between populations might reflect phenotypic similarity, gene-specific markers could provide crucial insights into the differentiation of populations in terms of traits, such as growth and development.

Due to its proximity to the North Pole and the effects of the passing Gulf Stream, the Nordic Region of Europe has highly variable weather with large differences in day length over seasons despite its geographically small area. Consequently, the region requires unique crop cultivation conditions, and the key to crop persistence could be found in its wild relatives. In this regard, genetic analyses of both wild and cultivated gene pools could link the breeding material used by Scandinavian breeding enterprises to resilient wild populations.

The purpose of this study was to compare and examine the genetic resources of red clover available in northern Europe by targeting its cultivated and wild gene pools that represent the Nordic countries. Here, SeqSNP was used to target SNPs within genes that influence growth and development, as well as disease resistance. Moreover, population genetic analyses were carried out in order to determine the correlation between genetic differences among wild populations and bioclimatic variables at the original collection sites.

Material and methods

Selecting germplasm and planting

For this study, 382 accessions of red clover were used that originate from different parts of the Nordic Region of Europe

(Supplementary Table 1). Of these, 294 accessions were obtained from NordGen (a regional genetic resources center for the Nordic countries) and 88 accessions from Lantmännen Seed (a plant breeding and agricultural seed company based in Sweden). The NordGen accessions were selected based on their passport data to represent a variety of available germplasm (cultivars, breeding populations, landraces, and wild populations) representing most of the Nordic region of Europe (i.e., Sweden, Norway, Finland, and Denmark). One Russian accession was also included as it was located at the Finnish border. The Lantmännen varieties include cultivars and synthetic populations from the Scandinavian forage breeding programs (Lantmännen, Sweden; Graminor, Norway; and DLF, Denmark), and hence reflect the variety of cultivated red clover available in northern Europe. These accessions include both diploid and tetraploid types that are categorized either as late or middle-late based on their maturation period.

The accessions were planted and grown for two weeks in a greenhouse at the Swedish University of Agricultural Sciences (SLU, Alnarp, Sweden), as described in Osterman et al. (2021). The BioArk leaf collection kit (LGC Biosearch Technologies) was used to collect ten 6 mm leaf discs (1 leaf disc/plant). One pool of leaf tissue representing ten plants was sampled for each accession separately. DNA extraction and genotyping were conducted at LGC Biosearch Technologies (Berlin, Germany), as described in Osterman et al. (2021).

Genotyping and variant calling

For genotyping, SeqSNP was used with a set of 400 target SNPs developed by Osterman et al. (2021). The SNPs were identified from publicly available red clover genomic resources by targeting coding sequences of genes known to be involved in growth and development as well as in the response to biotic and abiotic stresses. In addition to genotyping the target SNPs, SeqSNP was used to discover novel SNPs in the regions surrounding the target SNPs. In this analysis, 2 × 150 bp reads were used as a sequencing mode. The sequencing depth was set to 50 million read pairs (15 GB raw data) per sample to ensure sufficient coverage of each genotype in each pool thus adjusting the sequencing depth to ×501. SNP calling was performed by aligning the quality trimmed reads to the reference genome using Bowtie2 v2.2.3. For variant discovery, Freebayes2 was used with ploidy set to diploid and minor allele frequency set to 1%. To exclude calls due to sequencing error, allele counts below eight were set to zero as per the recommendations of LGC Biosearch Technologies where genotyping was conducted. The allele frequency of each accession at each locus was calculated based on the read counts.

For determining the validity of converting the read counts of pool-seq into allele frequencies for data analysis, five randomly selected accessions were genotyped at both the individual and pool levels. Following this, the read counts of each pooled sample were converted to allele frequencies. A subsequent step involved converting the genotypic data of the five individuals of each accession into allele frequency data for that particular accession. This was followed by correlation analysis between the allele frequencies obtained from individual genotype sequencing and PoolSeq, which revealed a highly significant correlation ($r > 0.95$; $P < 0.001$) between them. Hence, SeqSNP is a highly reliable method to generate data for allele frequency-based data analyses.

Among the 400 target SNP loci genotyped, 5.5% were mono-allelic, 86% were bi-allelic, 7.5% were tri-allelic, and 0.75% were tetra-allelic across the 382 accessions studied. The remaining 0.25% were INDELs (insertion or deletion of a nucleotide). The *de novo* SNP and INDEL calling generated 347 SNPs and 16 INDELs. Among the 347 novel SNPs, 91% were bi-allelic, 8% were tri-allelic, and 1% were tetra-allelic. Due to the complexity of analyzing pooled sequencing data and a mixture of diploid and tetraploid accessions, only polymorphic bi-allelic markers were used. Additionally, tetraploids were treated as diploids as described in Osterman et al. (2021). Overall, 344 originally targeted and 317 *de novo* discovered bi-allelic SNPs (661 SNPs in total), all with minor allele frequency of 5% or above, were used for data analyses in this study.

Genetic parameter estimation

Tajima's D was estimated using PoPoolation (Kofler et al. 2011b) based on the quality- trimmed reads combined in a sync file using the respective reference sequences to map the reads. The allele counts from Freebayes2 were imported into R (R Core Team, 2013) and the expected heterozygosity for each population (H_S) was calculated using the adegenet package (Jombart, 2008). Using the poolfstat package (Gautier et al., 2022) in R, pairwise F_{ST} was calculated for each pair of SNPs as well as for each pair of accessions. After grouping the accessions according to their origins or types an additional pairwise F_{ST} analysis was performed. Nei's standard genetic distance between populations and between groups (as for the F_{ST} analysis) was calculated using adegenet package. Additionally, Mantel's randomized test comparing Nei's standard genetic distance with the geographic coordinates of the germplasm collecting sites of the wild populations was performed to determine whether isolation by distance (IBD) has a significant effect on the genetic differences between the accessions.

Determining population structure via clustering

Both principal component analysis (PCA) and principal coordinate analysis (PCoA) were used to determine the genetic relationships between the accessions. The PCA was conducted using the pcadapt package (Luu et al., 2017), and SNPs that were most associated with the variation described in the first two principal components were extracted for further analysis. The PCoA was performed using the stats package in R (R Core Team, 2013) based on the Nei's genetic distance. The Nei's genetic distance based relationship between the accessions was further analyzed using ComplexHeatmaps (Gu et al., 2016), which generates heatmaps. These analyses enabled the comparison of the accessions both individually as well as collectively based on their origins and types.

The Nei's genetic distance based relationship between the accessions was further investigated through neighbor-joining (NJ) cluster analysis. The NJ tree was built using the ape package (Paradis et al., 2004) and visualized using the ggtree package (Yu et al., 2017). Incorporating bioclimatic variables to the NordGen accessions of wild populations and maturity types to the Lantmännen accessions into the analyses was made possible using the ggtree package. A map of collection coordinates for landraces, some cultivars and breeding populations as well as wild populations provided by NordGen was constructed using the rnaturalearth package, which uses maps from Natural Earth, in R.

A discriminant analysis of principal components (DAPC) was performed using adegenet with the method described by Jombart et al. (2010); Jombart and Collins (2015) on the allele frequencies. The find.clusters function was used to find the most optimal number of clusters based on the BIC score, and the cluster solution with the lowest BIC score was chosen. The xval function with a test set of 90% with 30 repetitions was used to find the appropriate number of principal components (PCs). This resulted in a five-cluster solution involving 150 PCs.

Environmental data for NJ tree and LASSO models

Bioclimatic variables were retrieved from WorldClim (Fick and Hijmans, 2017) with a spatial resolution of 30 seconds (~ 1 km^2) and imported via the raster (Hijmans et al., 2012) package in R. The coordinates of the germplasm collecting sites of the wild populations were used to extract environmental data with interpolation, hence minimizing the effect of potential recording errors. The bioclimatic variables were evaluated based on how they vary within Scandinavia. Most of the precipitation variables were similar across the sampling sites, and therefore they were excluded. The final set of bioclimatic variables from WorldClim include annual mean temperature and precipitation, as well as isothermal and precipitation seasonality. Annual snow coverage was estimated using snow thickness data retrieved from Climate Data Store (CDS) for months with snow (September to June) from 1980 to 2000. The mean snow thickness for each month at a sampling site during the years 1980 to 2000 was calculated and plotted with months on the x-axis and mean snow thickness on the y-axis. The area under the curve (AUC) of the yearly trend was calculated using the AUC function from the DescTools package (Andri, 2021) in R. The AUC value was chosen to represent mean snow coverage for each collection site.

SNPs with significant contribution to the variation explained by PCA

The pcadapt package (Luu et al., 2017) was used to perform a PCA with the Mahalanobis' method as the function argument. Thus, the Mahalanobis' distance was used to measure the extent to which each SNP is related to the first, in this case, two PCs. A χ^2-test was performed on the SNPs Mahalanobis distance to find those SNPs with significant contribution to the population structure, and the Benjamini–Hochberg correction was applied to control false positive discovery. The significant SNPs were then validated for their biological roles using gene ontology (GO) enrichment to find possible molecular functions differentiating the populations in the PCA clusters.

LASSO models

The least absolute shrinkage and selection operator (LASSO) regression model was used to connect environmental variables to the allele frequencies. As the number of SNPs exceeded the number of populations in this study, LASSO was considered an appropriate model due to the application of penalization and feature reduction. Among the accessions studied, only those representing wild populations were used for the LASSO models. This is because they could be considered to have adapted to the environments of the sampling sites. Considering the number of SNPs available for the data analysis was high (661), a method for selecting only the most relevant ones was devised to increase the biological aspect of model training. The goal of this analysis was to find the SNPs that genetically differentiated two populations. Thus, the F_{ST} values of all 661 SNPs between every pair of populations were calculated and the top 1% values of each pairwise calculation was selected. A final set of 430 SNPs with high F_{ST} values was selected for the LASSO models. The caret package (Kuhn, 2008) in R was used to train and select the LASSO model. A leave one out cross validation (*LOOCV*) approach was used to train a regression model (*glmnet*) and the mean average error (*MAE*) was computed for model selection. The MAE and root mean square error (RMSE) were compared to the variance of each climate variable to validate the final model; and the error was smaller than the standard deviation of the input for all models. To validate the models further, a linear regression analysis was performed on the bioclimatic variables without including the SNP data. In cases where the LASSO model had a lower RMSE than the linear model, the SNPs were considered to have an effect.

Validating selected SNPs in terms of biological functions

The analysis of SNP effect on population differentiation in the PCA and LASSO models resulted in nine sets of SNPs, two from the PCA and seven from the LASSO models. A gene ontology (GO) enrichment analysis was carried out on each set of SNPs to find any biological function underlying the population differentiation. The GO analysis was performed with the workbench at dicots Plaza 5.0 (Van Bel et al., 2021) where the correct names of the genes containing the SNPs were determined via the integrated BLAST function. Then enrichment was performed using all red clover genes as background with p-values showing significant enrichment adjusted following Bonferroni's correction.

Results

The SeqSNP-based sequencing of the 382 red clover accessions resulted in 661 bi-allelic SNP markers, which were then used for population genetics analysis of the accessions (Supplementary Table S2; Osterman et al., 2021). Additionally, 49 tri-allelic, four tetra-allelic and 17 INDELs were identified across the 400 target SNP loci, and 357 SNP loci were discovered *de novo*. Of the 317 *de novo* discovered bi-allelic SNPs, 292 were reported in Osterman et al. (2021) whereas the remaining 25 were specific to this study. It is evident from the number of *de novo* SNPs discovered in this study, compared to that of Osterman et al. (2021), the number of accessions studied had an effect on the number of novel SNPs discovered. Only bi-allelic SNPs were used for the data analyses for the sake of simplicity. At each of the 661 bi-allelic SNP loci, the allele frequency was calculated based on the read counts obtained from the sequencing. The read counts across the 661 bi-allelic SNPs ranged from eight to 4320. Although the range of the allele counts is large, there was no need to scale the frequencies since they were calculated independently for each locus of each accession.

Genetic variation within and among groups

For data analysis, the accessions were grouped based on their origins and population types. The grouping results in nine origin-based groups (Denmark, Finland, Graminor, Lantmännen, Norway, Sweden, DLF, Local population, and Russia) and eight population type-based groups (Breeding population, Cultivar, Diploid, Graminor, Landrace, Tetraploid, Unknown, and Wild Population). Because DLF, Local population, and Russia (among the origin-based groups) and Graminor and Unknown (among the population type-based groups) had only one accession each, they were excluded from some analyses.

The study revealed low genetic diversity and population structure considering the median and mean values of Nei's standard genetic distance and F_{ST} of each group (Table 1, Figure 1). Additionally, the results show a difference in the amount of rare alleles present within groups where wild populations had larger levels of rare alleles than cultivated accessions (Tajima's D in Table 1 and Figure 1). All cultivated groups (breeding populations, cultivars, diploids and tetraploids) had negative F_{ST} mean values. Both negative and zero F_{ST} values indicate lack of genetic variation distinct to each of the populations compared. Only wild populations had a positive mean F_{ST} value, hence, it is the only group (among population types) with any population structure. In the case of origin-based groups, the mean F_{ST} values were negative for Lantmännen, Graminor and Denmark, zero for Finland, and positive for Norway and Sweden. Apparently, the F_{ST} values of the different population type-based groups and origin-based groups were related due to the accessions they shared. The majority of the landrace accessions belong to Finland, and consequently the mean F_{ST} values for both groups were zero. Similarly, most of the accessions from Denmark were cultivars, and hence both Denmark (origin) and cultivars (population type) had a negative mean F_{ST} value. The mean F_{ST} for Sweden and Norway was higher, as they were mainly comprised of wild populations.

The pairwise F_{ST} between groups showed high genetic similarity between the cultivated types (Figure 2A) while the wild population group was divergent from the rest. Among the origin-based groups, Sweden, Norway, and Russia (which are dominated by wild populations) showed significant genetic differentiation from the other origin-based groups (forming a separate cluster in Figure 2B). This suggests a significant difference in allelic states between accessions from these countries and those from the other origins. Further, cultivated types as well as origin-based groups that are largely composed of cultivated types did not appear to have a clear population structure within or between them.

The mean values of Nei's standard genetic distance within the different groups were quite similar (Table 1). Hence, it was further investigated at a population level to illustrate groups of populations with high genetic similarity and had similar genetic relationships with the remaining populations. Only a few groups could be identified in the present study (marked by the blue rectangles in Figures 3A–C). There was a clear separation between clusters containing populations with a relatively high genetic distance (wild populations) and populations with low genetic distance (cultivars, Figure 3A). When the wild populations were separately analyzed, two clusters were identified, where one contained mainly the Swedish and Norwegian populations while the other contained mostly Swedish but also Finnish and Norwegian populations (Figure 3B). The separate analyses of the Lantmännen accessions

TABLE 1 The first column indicates the number of samples in different groups of red clover populations grouped according to their origin or population type.

Grouped by origin	N° samples	H_s	Nei	F_{st}	Tajima's D	B	C	D	G	L	T	U	W	
Denmark[a]	35	0.19	0.02	-0.01	-0.5	3	86						3	
Finland[a]	72	0.18	0.03	0	-0.03		4			88			15	
Graminor[a]	6	0.19	0.02	-0.03	0.02			33	17	50				
Lantmännen[a]	81	0.19	0.02	-0.02	-0.01			44			56			
Norway[a]	92	0.18	0.03	0.03	-0.04	5	1			4			89	
Sweden[a]	95	0.18	0.03	0.03	-0.05	1	7			9		1	81	
DLF[a]	1	0.20	-	-	-0.02						100			
Local population[a]	1	0.20	–	–	0.06		100							
Russia[a]	1	0.16	-	-	0								100	
Grouped by type						**Da**	**F**	**G**	**La**	**N**	**S**	**Df**	**Lo**	**R**
Breeding population[b]	10	0.19	0.02	-0.02	0.07	40				50	10			
Cultivar[b]	41	0.19	0.02	-0.02	-0.05	74	7			2	17			
Diploid[b]	43	0.19	0.03	-0.01	-0.01			5	93				2	
Graminor[b]	1	0.19	–	–	0.01			100						
Landrace[b]	71	0.19	0.03	0	-0.02		81			6	13			
Tetraploid[b]	45	0.20	0.02	-0.04	-0.01	91		7				2		
Unknown[b]	1	0.17	-	-	0.04						100			
Wild population[b]	172	0.18	0.04	0.04	-0.05	0.5	6			48	45			0.5

[a]= a group of accessions belonging to geographic origin; [b]= a group of accessions belonging to population type; H_s, mean expected heterozygosity; Nei, Nei's standard genetic distance; F_{st}, mean fixation index; Tajima's D, Tajima's population genetic test statistic.
The second to fifth column is the mean of the genetic parameter for the group. The last five columns refer to the composition of each group. When grouped by origin it is the percentage of population types, **B**reeding population, **C**ultivar, **D**iploid, **G**raminor, **L**andrace, **T**etraploid, **U**nknown and **W**ild population. When grouped by type the columns refer to the percentage of **D**enm**a**rk, **F**inland, **G**raminor, **La**ntmännen, **N**orway, **S**weden, **D**L**F**, **Lo**cal population or **R**ussian federation.

representing the cultivated gene pool revealed similarly low genetic distance between the accessions, and no clearly defined clusters were found (Figure 3C). Similar results were obtained with the cultivated accessions of NordGen, with the exception of a small cluster formed by the Finish landrace accessions (Figure 3D).

Cluster analysis via PCoA, DAPC and neighbor joining tree

A principal coordinate analysis conducted based on Nei's standard genetic distance showed that the first principal coordinate (PCo1), which accounted for 30.8% of the total variation distinguished most of the wild populations from the cultivated ones (Figure 4A). It was also shown that the landrace populations were placed between the wild and cultivated populations along the PCo1. The second principal coordinate (PCo2), which accounted for 10.9% of the total variation, distinguished wild populations and one landrace population originating from Norway from a group containing wild populations from Sweden, Finland, and Russia, as well as landrace populations from Finland. The results clearly showed that the major contributors to the variation displayed in the first two principal coordinates are wild populations. Thus, to get a better understanding of the main clusters, the 382 accessions were divided into subsets. In the wild population subset, the pattern persisted as expected and the cumulative variance described by the first two PCos decreased only slightly (from 41.7% to 39.1%; Figure 4B).

When the Lantmännen accessions were separately analyzed, the PCoA showed a cumulative variance of 23.9% in the first two PCos (Figure 4C). However, the scatter plot showed no clear partition between diploid and tetraploid accessions in both the PCo1 and PCo2. The separate PCoA of the NordGen accessions revealed a major separation of the landrace accessions from cultivars and breeding populations along the first PCo, which explained 24.8% of the total variation. Furthermore, it showed a separation of populations from Denmark and Finland (Figure 4D). The second PCo described far less variation (8.3%) and did not show a clear separation between any of the different groups.

A DAPC on the 382 accessions explained 79.2% of the total variance and revealed five clusters (Figure 5 and Supplementary Table S1). Clusters 3 and 5 mainly comprised the cultivated types as well as some wild accessions. The major source of variation for the differentiation between clusters 3 and 5 appears to be the accessions' countries of origin, especially Denmark versus Finland. Clusters 1, 2 and 4 differentiated the wild populations from the cultivated types although some landraces were contained in Clusters 1 and 4. Clusters 2 and 4 are

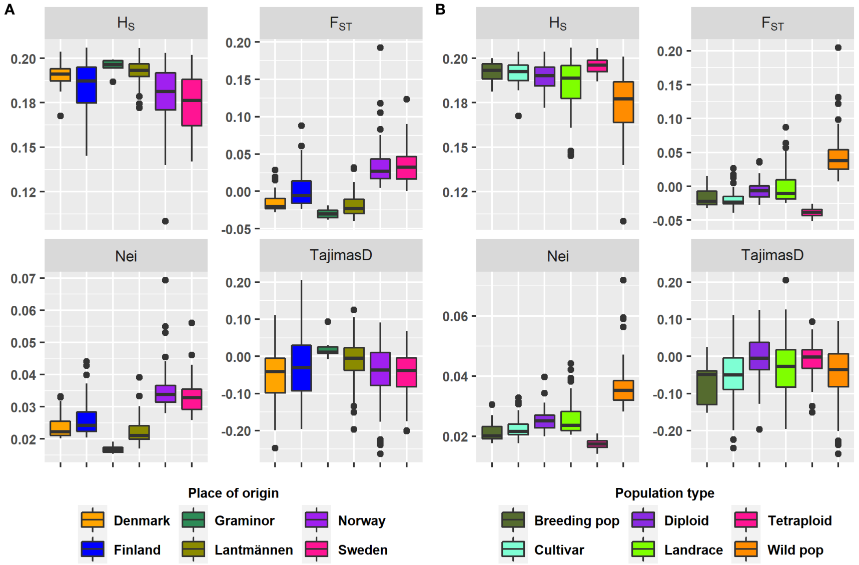


FIGURE 1
A box plot depicting the range and median for the genetic parameters on each group according to **(A)** Origin and **(B)** Type. The genetic parameters were H_S, mean expected heterozygosity; Nei, Nei's standard genetic distance; F_{ST}, mean fixation index; Tajima's D, Tajima's population genetic test statistic.

dominated by wild populations from Sweden and Norway, respectively, while Cluster 1 comprised of wild populations and landraces from Finland and Sweden. This clustering follows the map Nordic Region of Europe from east to west.

Each accession has been assigned to a cluster based on a membership probability, which can be plotted in the same way as the commonly used software STRUCTURE Following the instructions provided in Jombart and Collins (2015) (Figure 6). The membership probabilities were high with some overlaps between cluster 1 and 2, 1 and 3 as well as 1 and 4. Here, the differentiation between Finnish and Danish populations in clusters 1 and 4 is more prominent. It is again shown by the membership of wild populations in all clusters, that the wild populations contain a high genetic variance.

The differentiation between the cultivated and wild populations was again observed in a neighbor-joining cluster analysis based on Nei's standard genetic distance where the 382 accessions formed four major clusters (Figure 7 and Supplementary Table S1). Cluster-1 contained the majority of the wild populations, some cultivated types from both NordGen and the breeding companies. The NordGen cultivated types comprised three breeding populations from Norway, Denmark and Finland and three cultivars from Finland and Sweden. Whereas the Lantmännen cultivated types include one Graminor population and three diploid cultivars from Lantmännen. Cluster-2 and cluster-3 contained the majority of the Lantmännen accessions. Interestingly, cluster-4 contained almost exclusively Finnish accessions with the exception of one diploid cultivar and tetraploid cultivar from Lantmännen and one Norwegian wild population. Wild populations and landraces in cluster-2, cluster-3 and cluster-4 originated from along the coast or near a lake in southern to central Scandinavia. The Mantel test, which compared geographical distances with Nei's genetic distance, revealed that isolation by distance is evident (Supplementary Figure 1), indicating that environmental variance could be linked to genetic variation.

The bioclimatic data from WorldClim successfully described the local environment at each of the wild populations' sampling sites, demonstrating the variation of climate factors in the region. The five bioclimatic variables, shown in columns 4 to 8 in the heatmap of Figure 7, describe the average environment for each sampling site of the wild populations. They show, for example, a connection between lower annual mean temperatures and higher mean snow coverage. The highest snow coverage was recorded in the most northern geographical locations. Additionally, there were several geographical locations with high annual mean temperatures as well as high snow coverage. Similarly, there were multiple sites with high isothermality, i.e. a large difference in day to night

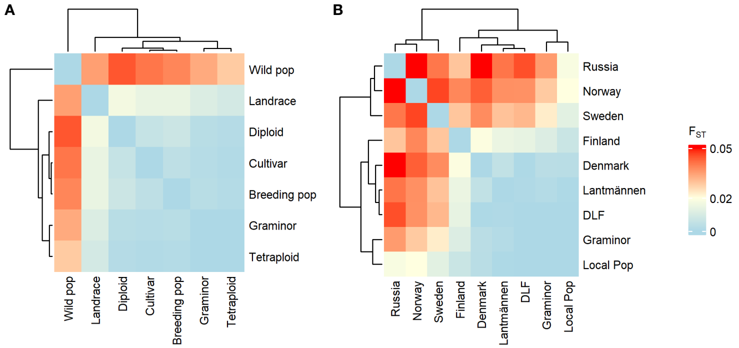


FIGURE 2
Heatmap depicting the pairwise F_{ST} values between groups of red clover populations based on population type **(A)** or origin **(B)**.

temperatures between summer and winter. Wild red clover from these locations would have developed resilience to the harsh winter conditions, to which cultivated red clover is highly susceptible.

Interestingly, despite the fact that the wild populations with close genetic relationship to Lantmännen material do not span the entire geographical area of interest, they still represent most of the climatic conditions. Nevertheless, the main climatic conditions associated with the Lantmännen breeding materials were warmer, steady temperatures and low snow coverage. Additionally, populations with similar maturation periods clustered together in cluster-2 and cluster-3. All populations in cluster-4 with a known maturation period, except one, were late maturing. Contrary to the hypothesis, that late maturing would have larger similarities with northern wild populations, nevertheless there was no clear distinction between the maturity groups.

LASSO models and GO analysis

In the present study, the parameter selection feature of LASSO models was used to estimate the SNPs that were the most informative in predicting the values of bioclimatic variables. A model with a root mean square error (RMSE) smaller than the standard deviation (SD) indicates that there is a predictive effect of the feature (SNP). In other words, there is an effect of the selected markers on the predictive ability of the model. All LASSO models had RMSE smaller or about equal to their respective SD (Table 2). Thus, all models had good predictive ability except for mean snow coverage and isothermality. The goodness of fit of the model was further confirmed by conducting a simple linear regression analysis without using the SNP information and comparing the RMSE. The RMSE of the linear regression was closer to the SD than the RMSE of the LASSO for all models except isothermality and snow coverage, thereby confirming the effect of the SNPs in the model's prediction.

The gene ontology (GO) enrichment analysis was used to validate the model results, in terms of biological functions. The reference gene-coding sequences, of the SNPs selected by the LASSO models were imported into the online tool Plaza workbench via BALST to find the corresponding genes. The models in which SNPs generated a GO enrichment were annual mean temperature, annual precipitation, and annual temperature range (Supplementary Table S3). Interestingly, the set of genes from both the annual precipitation and annual temperature models showed enrichment for genes regulating stomatal opening (Table 2 and Supplementary Table S3). The stomata are known to be involved in the plants regulation of water, oxygen and carbon dioxide, functions that are relevant to changes in temperature and precipitation (Waggoner and Zelitch, 1965; Honour et al., 1995). Furthermore, there was an enrichment of genes coding for kinase binding proteins in the annual mean temperature model (Table 2; Supplementary Table S3). Kinases are a group of enzymes that via post-translational modification plays an important role in plant growth and development. Some kinases are involved in the plant response to changes in both mild and extreme temperatures (Praat et al., 2021). Analogs of the three genes detected in the GO enrichment analysis were, via experimental evidence, connected to heat response (Larkindale et al., 2005; Wu et al., 2014) and to post translational modifications as response to external factors (Colby et al., 2006) in tomato and *Arabidopsis*.

The GO analysis of SNPs from PCA

From the PCA analysis, using the R software's pcadapt package, the SNPs that significantly contributed to the

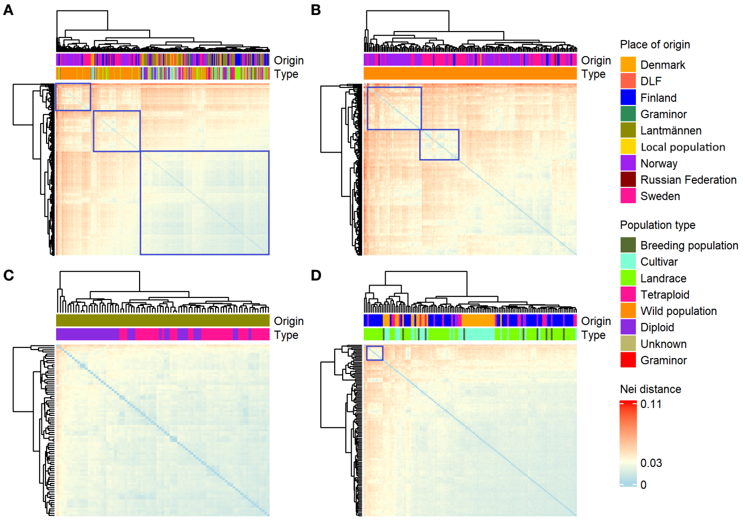


FIGURE 3
A heatmap depicting Nei's standard genetic distances between each pair of populations. The colors indicate high (red), intermediate (yellow) or low (blue) genetic distances. The accessions were clustered according to their pairwise genetic similarities. The accessions included were **(A)** all 382; **(B)** WildOnly those that are wild; **(C)** PopulationsOnly those from Lantmännen; and **(D)** Cultivarsonly cultivars and landraces from NordGen.

clustering of the 382 accessions in the first two PCs were located within the coding regions of 22 and 38 genes, respectively (Supplementary Table S3). A GO enrichment analysis of the 22 genes from PC1 revealed that 10.5% of the genes were enriched for two biological processes, namely, specification of plant organ axis polarity and regulation of root morphogenesis. However, no GO enrichment was observed for the 38 genes from PC2.

Discussion

This study revealed the genetic variation of 382 red clover accessions, including wild and cultivated types representing the red clover gene pool in the Nordic Region of Europe. Among the red clover accessions studied, 45 were known to be tetraploids. However, in order to facilitate the comparison with the diploid populations used in this study they were treated as diploids, following the explanation provided in Osterman et al. (2021). Diploidizing tetraploids is commonly employed to reduce complexity in data analysis and has been implemented in potato for population structure analysis using STRUCTURE and other software developed for diploids (Hirsch et al., 2013; Pandey et al., 2021; Selga, 2022). The genotyping was conducted using a pool-seq approach of the SeqSNP sequencing assay used by Osterman et al. (2021) that targeted genes known to be involved in growth and development as well as stress response. In total, 661 polymorphic, bi-allelic SNPs were detected in the targeted protein coding sequences across the 382 populations, demonstrating the potential of SeqSNP for sequencing the target SNPs as well as for *de novo* SNP discovery. Nevertheless, it should be noted that the novel SNPs identified were within 75 bp of the target SNPs. Therefore, SeqSNP is useful when specific coding regions are targeted, but may not be suitable for the identification of novel SNPs in larger regions, such as quantitative trait loci (QTL) and uncharacterized gene sequences.

The benefit of using pool-seq methods as opposed to individual sequencing methods is the number of populations that can be analyzed. With a pool of 10 individuals, pool-seq can analyze 10 times as many populations as individual genotype analysis for the same cost, assuming their sequencing depth is the same. The main challenge of pool-seq is the data analysis, as the representation of the sampled individuals within a pool can be uneven. The selection of reading depth relative to the number of individuals in each pool is very important. In the present study, using read counts from ten individuals at a sequencing depth of x501 was deemed appropriate, given the result of our in-house analysis that compared data generated through pool-seq and individual genotype sequencing.

The genetic parameters estimated from ten individuals per pool were comparable with the results reported by Jones et al. (2020)

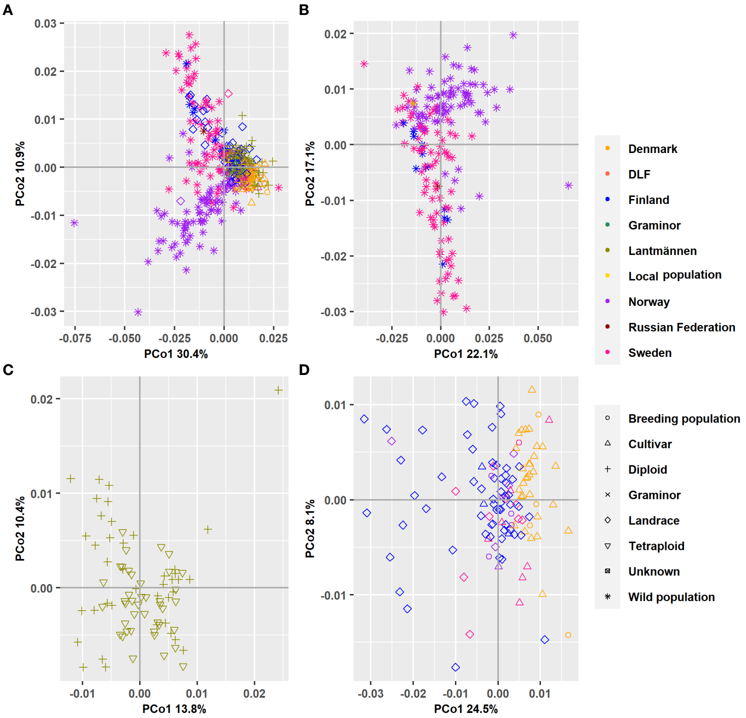


FIGURE 4
A bi-plot of the principal coordinate analysis (PCoA) showing the variation explained by the first two principal components. **(A)** All 382 populations' analyzed together and separate analysis when the accessions have been grouped according to wild populations **(B)** Lantmännen populations **(C)** or Landraces and cultivars held at NordGen **(D)**.

and Osterman et al. (2021). Various bioinformatics software packages have been developed to analyze pool-seq data, including BayPass (Gautier, 2015), PoPoolation (Kofler et al., 2011a; Kofler et al., 2011b), and SelEsim (Vitalis et al., 2014). Pool-seq based study on red clover using BayPass has previously been done to detect genomic signatures of herbicide resistance (Benevenuto et al., 2019). The study used pools of 20 to 40 individuals from 10 populations and data analysis was performed using BayPass, which uses read counts and a hierarchal Bayesian model to estimate genetic variance/covariance and outlier loci. However, in the present study, the number of populations (382) relative to the number of SNPs used was too large to apply BayPass.

The discovered genetic variation differentiated the groups of accessions to different extents based on their population type or origin. The major trend was a separation of the wild populations from the cultivars, with the landraces being represented within both groups. The genetic distance between the cultivated populations was low, and there was no clear separation between them based on their origins. This could be partly due to the strict outcrossing nature of the crop that facilitated a high rate of gene flow, resulting in high heterozygosity with reduced differences in allele frequency between the populations. A high level of heterozygosity has been previously reported in red clover populations analyzed at individual genotypes level, which was attributed to its strict outcrossing reproductive system (Jones et al., 2020; Osterman et al., 2021). Contrary to this, there was a pattern of population structure between the different groups of wild populations. The lower rate of gene flow between the wild populations is probably due to the low level of migration as well as the consequences of geographical distance and terrain.

This study was designed to identify informative SNPs from the perspective of red clover breeding and to generate knowledge regarding the extent to which the genetic material used for breeding reflects available genetic resources in red clover. These objectives are clearly reflected in the NJ tree (Figure 7) as well as

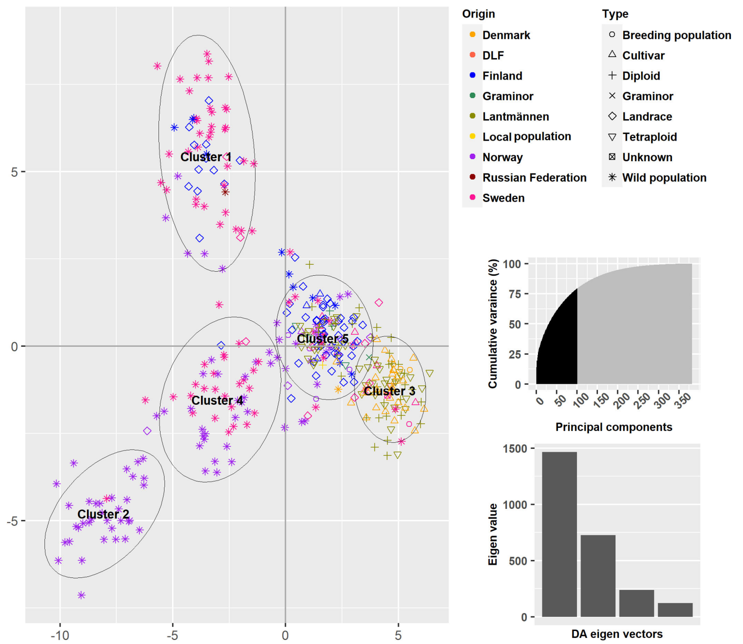


FIGURE 5
A Discriminant Analysis of Principal comonents using 150 Principal components and a five cluster solution.

in the results of LASSO model-based analysis where genetic variation was linked to bioclimatic variables and to relevant biological functions.

Genetic Parameters: H_S, F_{ST}, Nei's standard genetic distance and Tajima's D

The wild populations selected for this study fully spanned the Nordic Region of Europe (Figure 7) with no obvious geographical groupings. The higher F_{ST} (mean of 0.04, Table 1 and Figure 1B) values of the wild populations, compared to the cultivated types, suggests population structure as a consequence of either restricted gene flow, ongoing evolution, or both. In contrast, the cultivated red clover showed low F_{ST} values both within and between different cultivated types and the origins where these groups dominated (Table 1 and Figures 1B, 2A). This indicates a high gene flow within the different types of both the same and different origins. If the F_{ST} between a pair of populations is high, it implies significant differentiation between them. This means that their genetic constitution is significantly different, and hence their crossbreeding may lead to hybrids that are superior to both of them. Since low values of F_{ST} was recorded within cultivated types, crossbreeding with wild populations could lead to further genetic gain.

A lack of genetic differentiation between populations might lead to little to no genetic gain when crossbreeding. This is because the populations possess the same alleles in similar proportions across a majority of loci, and hence crossbreeding does not lead to significant genetic recombination. Even though red clover populations are expected to be highly heterozygous due to their outcrossing nature, variation between populations declines as the majority of their common alleles approach fixation. Tajima's D can be used to measure the amount of rare alleles in a population. Hence, maintaining genetic gain in breeding populations is dependent on the inclusion of new rare alleles. Populations with negative Tajima's D values can be considered to be in expansion following either a bottleneck or selective sweep and thus has an abundance of rare alleles (Tajima, 1989). By selecting such populations further genetic gain can be introduced into the breeding populations. The Tajima's D for the wild populations was negative, which supports the suggestion that an ongoing evolutionary process contributes to their higher genetic differentiation. Hence, the red clover wild populations might have experienced recent selective sweeps and/or events that reduced their population sizes.

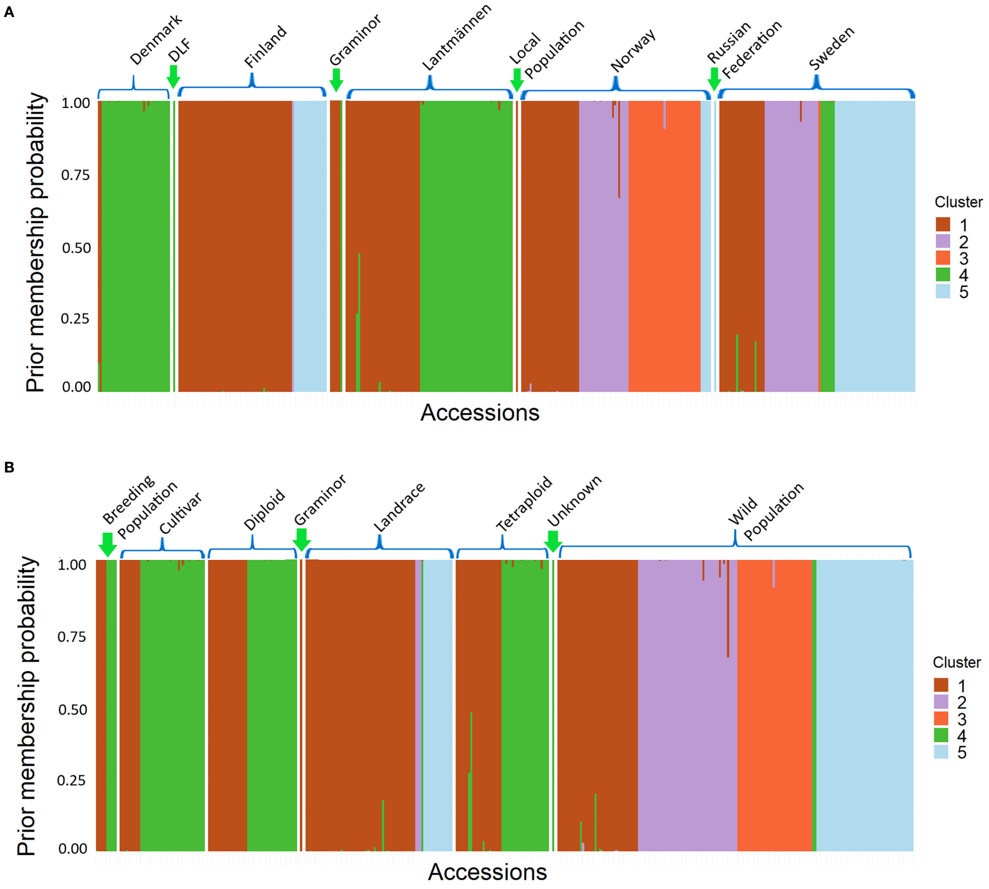


FIGURE 6
A Structure-like membership probability graph showing individual populations of different groups, classified according to **(A)** their origin and **(B)** their population type, assigned to a cluster.

Interestingly, the cultivars and breeding populations from NordGen had lower Tajima's D mean values than the wild populations (Table 1). The lowest Tajima's D values belonged to the Graminor accessions (both as origin and population type, Table 1). This might be due to balancing selection after the development of new cultivars or cultivar types.

Compared to the mean F_{ST} value presented by Jones et al. (2020) of 0.076 for red clover representing Europe and Asia, the mean F_{ST} of the present study (0.022) indicates lower genetic differentiation in Northern European red clover. These results are not surprising since Europe and Asia cover a larger area. Furthermore, red clover was introduced to northern Europe relatively late compared to southern and central Europe (Taylor and Quesenberry, 1996a). Additionally, Jones et al. (2020) used 93.3% ecotypes (here referred to as wild populations) compared to the 45% used in the present study. The results showed that the majority of the genetic diversity was held within the wild populations. Hence, a larger amount of wild populations would increase the F_{ST} in a sample set.

The lowest H_S median was recorded in the wild populations, which also had a higher genetic distance between populations compared to the other population types. The lower Hs values of the wild populations compared to the cultivated types indicate a low gene flow. Of the 382 populations selected for this study, 172 were wild populations. The large proportion of wild populations could be the reason for the relatively lower mean H_S of 0.18 in the present study (Supplementary Table 1) as compared to Osterman et al. (2021) who reported a mean H_S value of 0.21. Interestingly Jones et al. (2020) reported a mean H_S of 0.26. Therefore, the present study may suggest lower within-population diversity in Norther European wild red clover.

Principal coordinate analysis, discriminant analysis of principal components and neighbor-joining cluster analysis

The PCoA scatter plot for the 382 populations showed a separation between cultivated and wild populations on the first principal coordinate (Figure 3A). The second principal coordinate differentiated Swedish and Finnish wild populations from Norwegian wild populations. The landraces from Sweden and Norway clustered together with NordGen and Lantmännen cultivars while Finnish landrace populations were closer to the Swedish and Finnish wild populations. This pattern was also observed in the DAPC, where wild populations and landraces that were not in clusters 1 and 4 exhibited three major trends, Swedish and Finnish, Swedish and Norwegian, and only Norwegian (Figure 5 and Supplementary Table S1). These two main patterns were observed throughout the analysis, the separation of the Swedish wild populations from the Norwegian wild populations and the mixture of the NordGen and Lantmännen cultivars. This pattern was clearly observed in

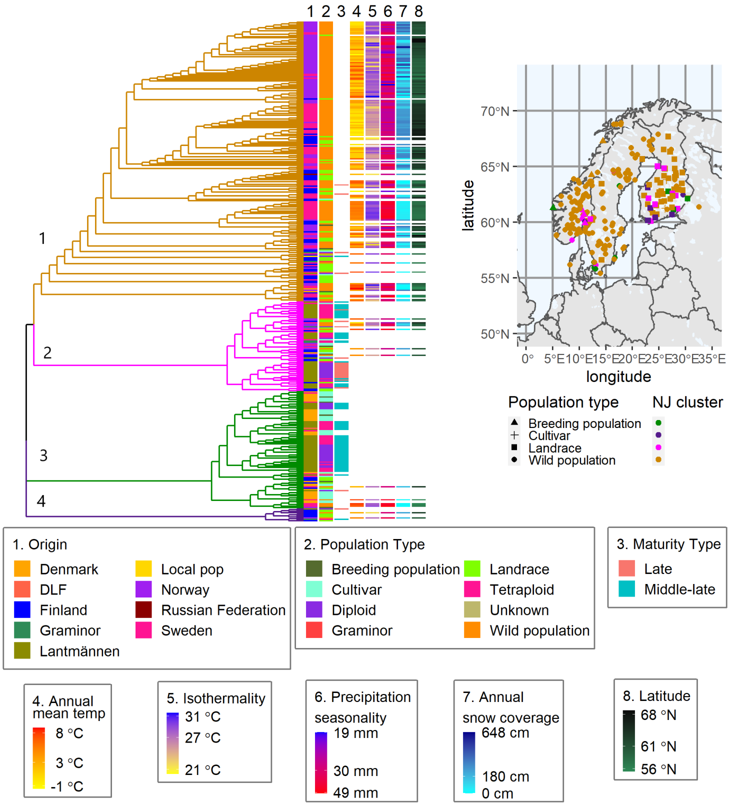


FIGURE 7
A Nei's standard genetic distance based neighbor-joining tree of the 382 populations showing four major clusters. Each column number (1-8) links the column to a specific trait or descriptive data, 1-2 origin and type, 3 maturity types on previously scored populations. Column 4-8 is the bioclimatic variables describing the collection sites of each wild population. The geographical map shows the collection site coordinates for breeding populations, cultivars, landraces, and wild populations obtained from NordGen.

TABLE 2 A summary of the results of best preforming least absolute shrinkage and selection operator (LASSO) model and gene ontology (GO) functional enrichment analysis for the most significant single nucleotide polymorphisms (SNPs).

Environmental parameter	SD	LAMBDA	MAE	RMSE (L)	RMSE (LM)	GO	#	Go genes
Annual mean temperature	2.4	0.2	1.5	1.8	2.4	MF: kinase binding	40	8.3%
Annual precipitation	226	3.7	127	175	227	BP: stomatal opening	88	4.2%
Isothermality	2.3	0.1	1.6	2.2	2.3	None	64	
Latitude	3.1	0.2	1.7	2.1	3	None	63	
Precipitation seasonality	6.1	0.4	4.1	5.3	6.1	None	41	
Annual snow coverage	132	2.5	90	133	131	None	45	
Temperature annual range	4.4	0.1	2.3	3.0	4.4	MF: protein binding BP: stomatal opening	74	50.8%, 4.9%

SD, standard deviation of the input data for each set of bioclimatic variables; RMSE (L), root mean square error of the LASSO model; RMSE (LM), root mean square error of a linear regression; MAE, mean absolute error of the model; GO, gene ontology; #, total number of genes; GO genes, percentage of genes that showed enrichment for each result of LASSO model.

the NJ analysis (Figure 7 and Supplementary Table S1) as well as in Nei's heatmap (Figure 2). This distinction between cultivated and wild red clover suggests that wild populations possess genetic variation that is not represented in the cultivated populations. Hence, by incorporating wild (Norwegian or Swedish) and landrace (Finnish) populations into the breeding programs for red clover, the genetic diversity of the cultivated gene pool can be increased further.

In agreement with the results in Osterman et al. (2021), higher genetic variation differentiated wild populations from Sweden and Norway than the genetic variation that differentiated NordGen and Lantmännen cultivars or diploids and tetraploids. In order to determine the genetic relationship between populations within different groups, various analyses were conducted by grouping the 382 populations according to their origins or types, to reveal any sub-groupings. A separate PCoA analysis for the Lantmännen populations showed no clear differentiation between diploids and tetraploids, in contrast to the low degree of differentiation observed among the NordGen cultivars and landrace populations (Figures 3C, D). Additionally, the lack of clear differentiation between the diploids and tetraploids was evident in the DAPC, where the diploids and tetraploids were assigned to clusters 1 and 2 similarly. However, genetic variation was higher within the diploid group than within the tetraploid group. This can be seen from the PCoA scatter plot where tetraploids were distributed close to the origin while diploids covered the full range of variance described by both PCo1 and PCo2.

In agreement with the results of the PCoA and DAPC, no clear differentiation between diploids and tetraploids was described by Nei's standard genetic distance (Figure 2). However, some sub-clustering of diploids and tetraploids was observed in cluster-2 and cluster-3 of the NJ tree although it was not as clear as their clustering pattern observed in Osterman et al. (2021). The grouping of the tetraploids into different sub-clusters in the present study indicates that they have been derived from chromosome-doubling experiments performed independently on diploids from different genetic backgrounds. Thus, crossbreeding of tetraploids representing different sub-clusters may result in superior cultivars with multiple desirable characteristics. Tetraploid cultivars have higher resilience and biomass yields than diploid cultivars. Hence, genetic differentiation between diploids and tetraploids is expected although it was not the case in the present study. It is likely that the agricultural gain from cultivating tetraploids derives from the molecular genetics of polyploidy rather than from an increased genetic variation since there is no clear genetic variation separating cultivars based on ploidy.

In the case of NordGen germplasm, it is interesting to note that genetic variation was higher among landraces than among cultivars, as clearly depicted in Figure 2D. The two groups also showed significant genetic differentiation, particularly when comparing the landraces from Finland and the cultivars from Denmark (Figures 2D, 5). The distinctness of some Finnish landrace populations was also demonstrated in the heatmap of Nei's genetic distance (Figure 3D) as well as in cluster-4 in the NJ tree (Figure 7). Hence, these Finnish landrace populations might have unique genetic constitution of significant breeding values (agronomic and forage quality) that needs to be explored further. Another interesting finding of the present study was the close genetic relationship between the Danish cultivars from NordGen and the Lantmännen populations (Figures 2, 4). Possibly, this is due to the frequent inclusion of Danish cultivars in Lantmännen breeding programs or to the use of similar genetic resources by different breeding programs to develop cultivars that share similar desirable traits, such as high forage yield.

In order to understand the genetic merit of the gene pool of wild populations, the bioclimatic variables of their respective collection sites were analyzed as a means of examining their respective environments. Wild populations, even naturalized cultivars, are thought to be well adapted to the climate of their natural habitats (Turesson, 1925). Hence, a high genetic similarity between a cultivar and a wild population may indicate that the cultivar is well suited to an environment similar to that of the wild population. Such analyses can provide insight into whether the germplasm under cultivation has sufficient genetic diversity to suit the diverse environments in which they are being cultivated. The present study revealed that cultivated red clover showed a greater tendency to cluster together with wild populations found in the warmer climates of the south and central parts of the Nordic Region with low levels of variation in precipitation and temperature and little to no snow cover. One of the main causes of red clover senescence is repeated freezing and thawing (Smith, 1957; Zanotto et al., 2021). However, such information is difficult to model. Instead, an educated guess can be made using isothermality and snow coverage to identify locations that may have long autumns with frequent fluctuations around the freezing point. Wild populations from northern Norway in cluster 2 (Figure 7), where the snow coverage is high and the annual mean temperature is around zero or negative, shared a close genetic relationship with three middle-late diploid cultivars bred by Lantmännen. Hence, it would be interesting to evaluate the winter hardiness of these cultivars to validate the ideas discussed above.

LASSO prediction and GO functional analysis

This study used least absolute shrinkage and selection operator (LASSO) models to relate the SNP frequency across populations to a specific bioclimatic variable. Due to its ability to rank the importance of variables LASSO models are currently used in multiple fields where the number of samples is less than the number of variables. They are used in gene-based diagnostics (Kohannim et al., 2012; Kim et al., 2018), genome-wide association research (Li et al., 2011), and other forms of unsupervised learning like in chemometrics (Pomareda et al., 2010). In the present study, the objective was to identify highly descriptive markers that can help select suitable germplasm for use in breeding programs. To the best of our knowledge, the LASSO models have not been used to relate a SNP marker to an environmental variable before. In this study, the method was considered successful because it was possible to assess the relationship between the SNPs identified by the LASSO models and the traits appropriate to the bioclimatic variable studied. Hence, this study demonstrates the ability of penalized linear regression models to assess the relationship between SNPs and environments.

Relating SNPs to bioclimatic variables with allele frequencies have previously been done via Bayesian models. However, for these models to converge, the data must fulfill the assumptions of the prior likelihood distribution. If the data does not fit the Bayesian model, a LASSO model could be used as an alternative as they rely on no prior information. LASSO models, however, are not adjusted based on population structure, unlike Bayesian models. As a result, additional steps are necessary in order to exclude possible artifacts due to population structure or statistical false discovery. A GO enrichment analysis was carried out for this purpose in the present study and satisfactory results were obtained.

Four LASSO models showed enrichment for biological processes that can be regarded as plausible responses of plants that are growing in a particular environment (Table 2). For example, the temperature and precipitation models showed enrichment of genes related to stomatal opening, a function known to be involved in the plant response to humidity and temperature (Waggoner and Zelitch, 1965; Honour et al., 1995). Additionally, the enrichment of protein kinases in the annual mean temperature model gives further information on the mechanisms of plant resilience. For a better understanding of persistence in red clover, a study of the stomatal changes and kinase activities in plants bearing different alleles related to their survival would be valuable. Similarly, the knowledge of how wild red clover copes with temperature stress is valuable for breeding since cold resilience is a desirable trait in Nordic breeding programs and beyond. Additionally, a study on root development between cultivated and wild red clover would be of interest given the results of the GO enrichment of auxiliary root development. The establishment of roots could affect persistence, resilience, nitrogen fixation, and nutrient uptake, as well as the establishment of the whole plant. This way of evaluating new germplasm could serve as a key component of red clover improvement.

Conclusion

This study thoroughly described the genetic diversity and population structure of the Nordic red clover genetic resources, which include breeding populations, cultivars, landraces, and wild populations. As shown by this study, further genetic gains are possible by incorporating NordGen cultivars and landraces. Inclusion of selected landraces and wild populations based on the results exhibited in Figure 7 into red clover breeding programs could increase persistence and climate resilience of cultivars and synthetic populations. Furthermore, GO enrichment analysis facilitated the identification of SNPs that may affect the stomatal function and root development in wild populations, thus providing additional knowledge for breeding this forage crop. It would be very interesting to see this method applied in other similar studies involving wild germplasm.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/; PRJNA765476.

Author contributions

MG secured the funding with help from RO, CH, and other project participants. With inputs from CH and RO, MG designed the genotyping part of the study while JO planned the statistical analyses regarding bioclimatic variables. JO, CH, and MG conducted the greenhouse work, including sampling leaf tissue for DNA extraction. JO wrote the manuscript draft and revised it. MG, CH, and RO reviewed the different versions of the draft and assisted in the revision process. All authors contributed to the article and approved the submitted version.

Funding

The study was fully funded by SLU Grogrund – Centre for Breeding of Food Crops, Swedish University of Agricultural Sciences.

Acknowledgments

We would like to thank Linda Öhlund (Lantmännen) for providing us with seeds of red clover cultivars and breeding populations. We would also like to thank NordGen for supplying us with selected red clover germplasm from their collection. We are thankful to Mohammad El-Khalifeh (NordGen) for the additional help in selecting the accessions, and to the greenhouse management staff for keeping the plants healthy. We would like to thank David Parsons (SLU), Elisabet Nadeau (SLU) and Alf Ceplitis (Lantmännen) for constructive comments on the results of the study during the project meetings.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.997860/full#supplementary-material

References

Amdahl, H., Aamlid, T. S., Ergon, Å., Kovi, M. R., Marum, P., Alsheikh, M., et al. (2016). Seed yield of Norwegian and Swedish tetraploid red clover (Trifolium pratense l.) populations. *Crop Sci.* 56, 603–612. doi: 10.2135/cropsci2015.07.0441

Andri, S. (2021) *DescTools: Tools for descriptive statistics*. Available at: https://cran.r-project.org/package=DescTools (Accessed 1st January 2022).

Benevenuto, J., Bhakta, M., Lohr, D. A., Ferrão, L. F. V., Resende, M. F. R., Kirst, M., et al. (2019). Cost-effective detection of genome-wide signatures for 2,4-d herbicide resistance adaptation in red clover. *Sci. Rep.* 9, 20037. doi: 10.1038/s41598-019-55676-9

Colby, T., Matthäi, A., Boeckelmann, A., and Stuible, H.-P. (2006). SUMO-conjugating and SUMO-deconjugating enzymes from arabidopsis. *Plant Physiol.* 142, 318–332. doi: 10.1104/pp.106.085415

De Vega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, Å., et al. (2015). Red clover (trifolium pratense l.) draft genome provides a platform for trait improvement. *Sci. Rep.* 5, 17394. doi: 10.1038/srep17394

Dhamala, N. R., Eriksen, J., Carlsson, G., Søegaard, K., and Rasmussen, J. (2017). Highly productive forage legume stands show no positive biodiversity effect on yield and N2-fixation. *Plant Soil* 417, 169–182. doi: 10.1007/s11104-017-3249-2

Ergon, Å., Skøt, L., Sæther, V. E., and Rognli, O. A. (2019). Allele frequency changes provide evidence for selection and identification of candidate loci for survival in red clover (Trifolium pratense l.). *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00718

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatology* 37, 4302–4315. doi: 10.1002/joc.5086

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* 201, 1555–1579. doi: 10.1534/genetics.115.181453

Gautier, M., Vitalis, R., Flori, L., and Estoup, A. (2022). f-statistics estimation and admixture graph construction with Pool-seq or allele count data using the R package poolfstat. *Molecular Ecology Resources* 32, 1394–1416. doi: 10.1111/1755-0998.13557

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847–2849. doi: 10.1093/bioinformatics/btw313

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Herrmann, D., Boller, B., Studer, B., Widmer, F., and Kölliker, R. (2008). Improving persistence in red clover: Insights from QTL analysis and comparative phenotypic evaluation. *Crop Sci.* 48, 269–277. doi: 10.2135/cropsci2007.03.0143

Hijamns, R. J., Etten, J., Sumner, M., Cheng, J., Baston, D., Bevan, A., et al. (2012) *Raster: Geographic data analysis and modeling*. Available at: https://CRAN.R-project.org/package=raster (Accessed December 3, 2021).

Hirsch, C. N., Hirsch, C. D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., et al. (2013). Retrospective view of north American potato (Solanum tuberosum l.) breeding in the 20th and 21st centuries. *G3 Genes|Genomes|Genetics* 3, 1003–1013. doi: 10.1534/g3.113.005595

Honour, S. J., Webb, A. A. R., and Mansfield, T. A. (1995). The responses of stomata to abscisic acid and temperature are interrelated. *Proc. R. Soc. London. Ser. B: Biol. Sci.* 259, 301–306. doi: 10.1098/rspb.1995.0044

Jombart, T. (2008). Adegenet: A r package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., and Collins, C. (2015). A tutorial for discriminant analysis of principal components (DAPC) using adegenet.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* 11, 94. doi: 10.1186/1471-2156-11-94

Jones, C., De Vega, J., Lloyd, D., Hegarty, M., Ayling, S., Powell, W., et al. (2020). Population structure and genetic diversity in red clover (trifolium pratense l.) germplasm. *Sci. Rep.* 10, 8364. doi: 10.1038/s41598-020-64989-z

Kim, S. M., Kim, Y., Jeong, K., Jeong, H., and Kim, J. (2018). Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. *Ultrasonography* 37, 36–42. doi: 10.14366/usg.16045

Kofler, R., Orozco-terWengel, P., Maio, N. D., Pandey, R. V., Nolte, V., Futschik, A., et al. (2011a). PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS One* 6, e15925. doi: 10.1371/journal.pone.0015925

Kofler, R., Pandey, R. V., and Schlötterer, C. (2011b). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-seq). *Bioinformatics* 27, 3435–3436. doi: 10.1093/bioinformatics/btr589

Kohannim, O., Hibar, D., Stein, J., Jahanshad, N., Hua, X., Rajagopalan, P., et al. (2012). Discovery and replication of gene influences on brain structure using LASSO regression. *Front. Neurosci.* 6. doi: 10.3389/fnins.2012.00115

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28, 1–26. doi: 10.18637/jss.v028.i05.

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743–756. doi: 10.1093/genetics/124.3.743

Larkindale, J., Hall, J. D., Knight, M. R., and Vierling, E. (2005). Heat stress phenotypes of arabidopsis mutants implicate multiple signaling pathways in the acquisition of thermotolerance. *Plant Physiol.* 138, 882–897. doi: 10.1104/pp.105.062257

Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics* 27, 516–523. doi: 10.1093/bioinformatics/btq688

Li, W., Riday, H., Riehle, C., Edwards, A., and Dinkins, R. (2019). Identification of single nucleotide polymorphism in red clover (Trifolium pratense l.) using targeted genomic amplicon sequencing and RNA-seq. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01257

Luu, K., Bazin, E., and Blum, M. G. B. (2017). Pcadapt: An r package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* 17, 67–77. doi: 10.1111/1755-0998.12592

McKenna, P., Cannon, N., Conway, J., and Dooley, J. (2018). The use of red clover (Trifolium pratense) in soil fertility-building: A review. *Field Crops Res.* 221, 38–49. doi: 10.1016/j.fcr.2018.02.006

Öhberg, H. (2008) *Studies of the persistence of red clover cultivars in Sweden*. Available at: https://pub.epsilon.slu.se/1741/ (Accessed July 7, 2021).

Osterman, J., Hammenhag, C., Ortiz, R., and Geleta, M. (2021). Insights into the genetic diversity of Nordic red clover (Trifolium pratense) revealed by SeqSNP-based genic markers. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.748750

Pandey, J., Scheuring, D. C., Koym, J. W., Coombs, J., Novy, R. G., Thompson, A. L., et al. (2021). Genetic diversity and population structure of advanced clones selected over forty years by a potato breeding program in the USA. *Sci. Rep.* 11, 8344. doi: 10.1038/s41598-021-87284-x

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in r language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J. P., et al. (2013). Advances in plant gene-targeted and functional markers: A review. *Plant Methods* 9, 6. doi: 10.1186/1746-4811-9-6

Pomareda, V., Calvo, D., Pardo, A., and Marco, S. (2010). Hard modeling multivariate curve resolution using LASSO: Application to ion mobility spectra. *Chemometrics Intelligent Lab. Syst.* 104, 318–332. doi: 10.1016/j.chemolab.2010.09.010

Praat, M., De Smet, I., and van Zanten, M. (2021). Protein kinase and phosphatase control of plant temperature responses. *J. Exp. Bot.* 72, 7459–7473. doi: 10.1093/jxb/erab345

R Core Team (2013). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: http://www.R-project.org/.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* 15, 749–763. doi: 10.1038/nrg3803

Selga, C., Chrominski, P., Carlson-Nilsson, U., Andersson, M., Chawade, A., and Ortiz, R. (2022). Diversity and population structure of Nordic potato cultivars and breeding clones. *BMC Plant Biology* 22, 350. doi: 10.1186/s12870-022-03726-2

Smith, D. (1957). Flowering response and winter survival in seedling stands of medium red Clover1. *Agron. J.* 49, 126–129. doi: 10.2134/agronj1957.00021962004900030005x

Smith, R. R., Taylor, N. L., and Bowley, S. R. (1985). "Red clover," in *Clover science and technology* (John Wiley & Sons, Ltd), 457–470. doi: 10.2134/agronmonogr25.c19

Sturz, A. V., Christie, B. R., Matheson, B. G., and Nowak, J. (1997). Biodiversity of endophytic bacteria which colonize red clover nodules, roots, stems and foliage and their influence on host growth. *Biol. Fertil Soils* 25, 13–19. doi: 10.1007/s003740050273

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585

Taylor, N. L., and Quesenberry, K. H. (1996a). "Historical perspectives," in *Red clover science current plant science and biotechnology in agriculture*. Eds. N. L. Taylor and K. H. Quesenberry (Dordrecht: Springer Netherlands), 1–10. doi: 10.1007/978-94-015-8692-4_1

Taylor, N. L., and Quesenberry, K. H. (1996b). "Persistence," in *Red clover science current plant science and biotechnology in agriculture*. Eds. N. L. Taylor and K. H. Quesenberry (Dordrecht: Springer Netherlands), 119–129. doi: 10.1007/978-94-015-8692-4_10

Taylor, N. L., and Quesenberry, K. H. (1996c). "Tetraploid red clover," in *Red clover science current plant science and biotechnology in agriculture*. Eds. N. L. Taylor and K. H. Quesenberry (Dordrecht: Springer Netherlands), 161–169. doi: 10.1007/978-94-015-8692-4_13

Thilakarathna, M. S., Papadopoulos, Y. A., Grimmett, M., Fillmore, S. A. E., Crouse, M., and Prithiviraj, B. (2017). Red clover varieties show nitrogen fixing advantage during the early stages of seedling development. *Can. J. Plant Science.* 98, 517–526. doi: 10.1139/cjps-2017-0071

Turesson, G. (1925). The plant species in relation to habitat and climate. *Hereditas* 6, 147–236. doi: 10.1111/j.1601-5223.1925.tb03139.x

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., et al. (2021). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Res.* 50, D1468–1474. doi: 10.1093/nar/gkab1024

Vitalis, R., Gautier, M., Dawson, K. J., and Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics* 196, 799–817. doi: 10.1534/genetics.113.152991

Waggoner, P. E., and Zelitch, I. (1965). Transpiration and the stomata of leaves. *Science* 150, 1413–1420. doi: 10.1126/science.150.3702.1413

Wu, J., Wang, J., Pan, C., Guan, X., Wang, Y., Liu, S., et al. (2014). Genome-wide identification of MAPKK and MAPKKK gene families in tomato and transcriptional profiling analysis during development and stress response. *PloS One* 9, e103032. doi: 10.1371/journal.pone.0103032

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628

Zanotto, S., Palmé, A., Helgadóttir, Á., Daugstad, K., Isolahti, M., Öhlund, L., et al. (2021). Trait characterization of genetic resources reveals useful variation for the improvement of cultivated Nordic red clover. *J. Agron. Crop Sci.* 207, 492–503. doi: 10.1111/jac.12487

Acta Universitatis agriculturae Sueciae
Doctoral Thesis No. 2024:33

Breeding red clover for improved yield and fodder quality is increasingly in demand, as it is a valuable forage legume. To meet this demand, breeding efficiency should be improved. In this thesis, the genetic diversity of its gene pool in Northern Europe was studied and genomic prediction models were developed. Understanding its genetic diversity can help breeders avoid the potential impact of inbreeding depression. Furthermore, genomic prediction models can speed up the breeding cycle and increase parent selection accuracy, helping breeders achieve breeding objectives relatively quickly.

Johanna Osterman received her doctoral education at the Department of Plant breeding, Swedish University of Agriculture in Sweden. She received her MSc in Engineering in biotechnology from Umeå University in Sweden.

ISSN 1652-6880
ISBN (print version) 978-91-8046-330-0
ISBN (electronic version) 978-91-8046-331-7