UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Sede di Piacenza

Dottorato di ricerca per il Sistema Agro-alimentare

Ph.D. in Agro-Food System

Cycle XXXVI

S.S.D. AGR/07. Genetica Agraria

Development and validation of genomic selection models for the improvement of pea grain yield and protein content in Italian environments

Coordinator:

Ch.mo Prof. Paolo Ajmone Marsan

Candidate: Margherita Crosta

Matriculation n: 5014531

Academic Year 2022/2023

UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Sede di Piacenza

Dottorato di ricerca per il Sistema Agro-alimentare

Ph.D. in Agro-Food System

Cycle XXXVI

S.S.D. AGR/07. Genetica Agraria

Development and validation of genomic selection models for the improvement of pea grain yield and protein content in Italian environments

Coordinator:

Ch.mo Prof. Paolo Ajmone Marsan

Tutor:

Dr. Paolo Annicchiarico

Prof. Adriano Marocco

Candidate: Margherita Crosta

Matriculation n: 5014531

Academic Year 2022/2023

INDEX

1. Introduction

1.1. Abstract

Wider pea (*Pisum sativum* L.) cultivation has great interest for European agriculture, owing to its favourable environmental impact and provision of high-protein feedstuff, for which Europe is largely dependent on importations. The main goal of this work was the investigation of genomic selection (GS) potential for the improvement of pea grain yield, protein content and their combination (protein yield) in environments of northern and central Italy, both per se and relative to phenotypic selection (PS). All genomic data were obtained by genotyping-by-sequencing (GBS) based on ApekI restriction enzyme. A Genome-Wide Association Study (GWAS) was performed for grain yield and protein content on three connected Recombinant Inbred Line (RIL) populations and a worldwide germplasm collection, to have a deeper insight into the genetic architecture of these traits. Moreover, the genetic correlation between these traits, and the phenotypic correlation between protein yield and each of its components were assessed in different environments for these material sets. The extent of variation attributable to the genetic and genotype × environment interaction (G × E) components was investigated in three connected RIL populations characterized in three environments. The inter-environment predictive ability of GS models was assessed for the target traits in breeding material, as represented by RIL populations issued by crosses between élite European cultivars, both in an intra- and inter-population prediction scenario (meaning that GS models were applied on the same or a different genetic base relative to that employed for training, respectively). Moreover, GS models for grain yield and protein content were developed on a worldwide germplasm collection and tested for the ability to predict the breeding values of other accessions from the same material or from three connected RIL populations evaluated in three independent environments. A comparison between GS and PS was performed by computing the genetic gains achieved in one selection year by similar budgets relative to the parental lines of each RIL population, either belonging or not to the GS training set. GWAS confirmed the expected polygenic control of grain yield and protein content, by highlighting many significant Single Nucleotide Polymorphisms (SNPs) in different genomic regions. Phenotypic correlation results highlighted a largely predominant role of grain yield on protein yield determination relative to protein content, while the genetic correlation between grain yield and protein content resulted mostly non-significant. Protein content displayed a superior GS predictive ability in all the scenarios, benefiting from a higher within-trial broad-sense heritability and a lower influence of G × E, compared with grain and protein yield. Mean predictive ability values were moderately high

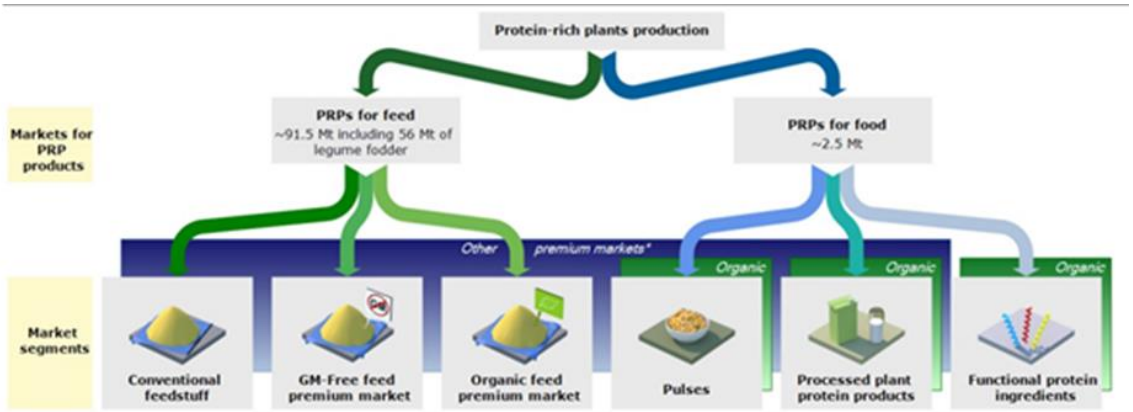
for all the target traits in the intra-population scenario independently from the material set, while a remarkable drop in predictions was observed, especially for grain and protein yield, in the inter-population scenario both for models trained on narrow (three connected RIL populations) or wide (germplasm collection) genetic bases. However, a strong variation in GS predictive ability was detected for all the target traits, especially in the inter-population scenario, depending on the specific material employed for validation, with even grain and protein yield predictions revealing satisfactory for some RIL populations. In the GS inter-population scenarios, the validation sets showing a higher number of polymorphic markers, either computed on all SNPs or on a moderately large subset of top-effect SNPs, tended to display superior predictive ability values for grain yield, while this was not always true for protein content. GS tended towards much higher genetic gains than PS for populations from the GS training set, whereas an opposite scenario characterized the non-training RIL populations, albeit with large between-population variation emerging for both material sets. Overall, our results encourage the adoption of GS for the simultaneous improvement of grain yield and protein content in an intra-population scenario, while showing an interest of inter-population predictions only for protein content or for grain yield on specific material sets.

1.2. Legumes as a resource for the environmental sustainability and food security of European agriculture

1.2.1. Uses and market trends of protein-rich plants in the European Union

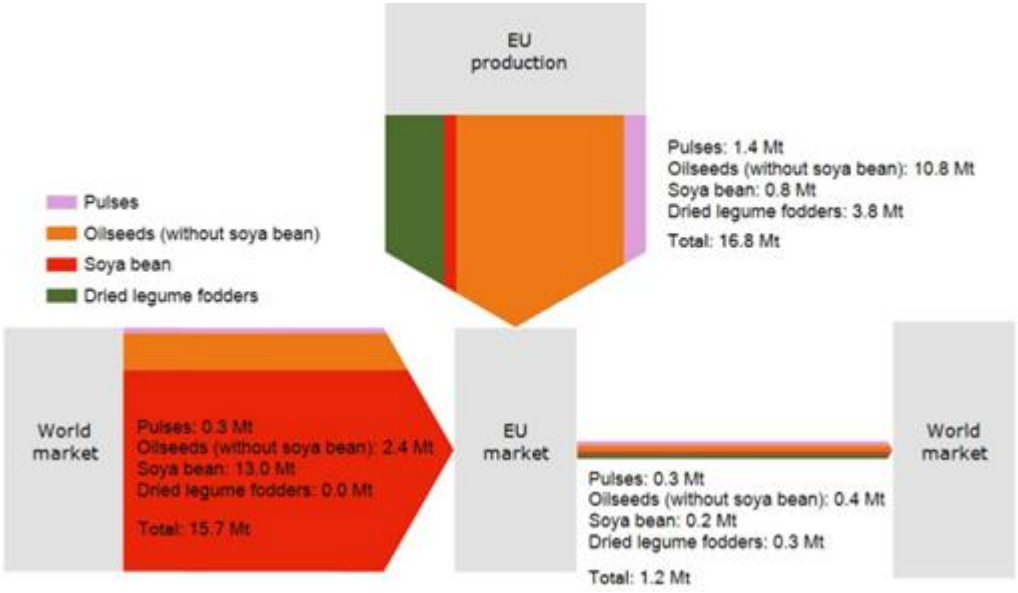
Protein-rich plants, including legumes (consisting of pulses, namely beans, peas, lentils, lupins etc.; fodder legumes, namely mainly alfalfa and clover; and soybean, which is considered as an oilseed) and other oilseeds (rapeseed, sunflower), feature a crude protein content higher than 15% and currently represent about 25% of the total crude plant protein supply in the European Union (EU). Animal feeding is by far the main destination of protein-rich plants and derivatives in the EU, absorbing more than 94% of total consumption, while the rest is employed by the food sector (Figure 1).

Figure 1. Summary of the main market segments using legumes in the European Union (Source: Clément et al., 2018).



Imported oilseeds and derived meals represent the main source of protein-rich plant material in the EU, with a predominant role of soybean (87% of protein-rich plant importations), the rest being essentially sunflower and rapeseed, while legume fodder and pulses are mainly produced internally (Figure 2) and play a minor role, contributing to just 15% and 3% of the total feed protein supplied by protein-rich plants, respectively (Clément et al., 2018).

Figure 2. Protein balance for protein-rich plants in the European Union during 2016 (unit: crude protein, Source: Clément et al., 2018).



EU feed market for protein-rich plants can be divided into three main segments: (1) conventional, (2) free from genetic modification (GM-free), and (3) organic, whose main drivers are resumed in Table 1.

Table 1. Main drivers of protein-rich plants (PRP) feed market segments in the European Union, including conventional, GM-free, and organic sectors. "GM-free" means genetic modification free, "non-GM" non-genetically modified, "GMO" genetically-modified organism, "PDO" Protected Designation of Origin, "PGI" Protected Geographical Indication, and "B2B" business to business, referring to transactions between two companies (Source: Clément et al., 2018).

Conventional feedstuff	GM-Free premium feedstuff	Organic premium feedstuff
- Seek for standard products, with stable quality and availability. constant availability and quality + standard products. - Need for hedging solutions (or price indexation if no futures market). - Feedstuff substitution costs in factories (limited number of silos). - Labour cost (especially for leguminous fodders).	- Social demand - GM-Free soya bean price premium (+ 80-100€) or cost of its substitution. - Availability of non-GM raw materials. - Branding and PDO/PGI. - Price premium for milk. - Additional sourcing, segregation, storage and transaction costs. - Availability of GM-Free soya bean. - GMO regulations.	- Segregation cost of organic materials: storage, handling and pest management. - Agronomic constraints inherent to organic production. - Competition with food outlets. - Availability of organic raw material for feed. High PRP prices in organic. - Lack of B2B services for technological treatments (dehulling, toasting, storing), especially at small scale.

Increasing the cultivation of pulse species adapted to European conditions could be an effective way to reduce EU dependency on protein-rich crop importations, but several barriers impeding the diffusion of these crops still exist (European Commission, 2018). Firstly, exploiting the complementary nutritional profile of oilseed meals, featuring a high protein content, and cereals, featuring a high starch content, is more convenient for the conventional compound feed sector than using crops characterized by intermediate nutritional profiles, such as pulses. In addition, the possibility to employ oilseeds for biodiesel production while obtaining a marketable by-product, namely protein meal, represents a further advantage of these species compared to pulses (Hay, 2019), which are also penalized by the limited market size and agronomic constraints (e.g., pests and low yields). On the other hand, the development of premium markets, such as GM-free and organic ones, and local value chains could contribute to enhance pulse competitiveness, which currently suffers from a high price variability due to the inferior quality consistency, supply steadiness, and hedging opportunities of these species compared to oilseeds. Finally, the growing diffusion of technological treatments such as toasting, dehulling, and protein extraction could increase protein concentration and digestibility in pulse derivatives compared with unprocessed materials, thereby improving their value as animal feed.

Although legume food market represents a niche compared to the feed one, it benefits from higher prices all along the value chain. It consists of three main segments, namely (1) legumes commercialized as whole grains (dry, fresh, canned, or frozen), which can be sold as such or incorporated in ready-to-eat dishes; (2) processed legume-based products (e.g., meat and

dairy alternatives and pulse-based snacks); and (3) functional protein ingredients used by food companies for technical or nutritional purposes. Legumes commercialized as whole grains represent the main segment, accounting for about three quarters of the total market, but the remaining two segments have experienced a significant growth in the last decade and benefit from higher prices compared to legumes sold as such. The legume sourcing strategy depends on the targeted market segment and consumer type, with EU supply relying on campaign contracts with collectors. Whole grains are either sourced locally in the case of more quality-oriented production (e.g., products labelled as Protected Designation of Origin, namely PDO, Protected Geographical Indication, that is PGI, or reporting the country of origin on packaging) or imported in the case of more price-competitive production, while processed legume-based products normally rely on local supply. Different trends characterize soybean-derived compared with pea-derived functional ingredients, with the former being mainly imported, while the latter mostly relying on the internal production. Legume demand for food is affected by different factors depending on the market segment, especially that of whole grain pulses and processed products is mainly driven by the final consumer, while that of functional protein ingredients by the agri-food industry (Table 2).

Table 2. Main drivers of legume food market segments in the European Union (EU), including whole grains, processed plant protein products, and functional protein ingredients. "GM-free" means genetic modification free, and "R & D" research and development (Source: Clément et al., 2018).

Pulses (whole grains)	Processed plant protein products	Functional protein ingredients
- Consumer habits - Image conveyed by the product: positive (traditional, healthy, etc.) or negative (old-fashioned, hard to cook, etc.). - Rise of flexitarian, vegetarian and vegan diets - Availability and stability of supply (implying contracts) - Quality of the grains - EU origin or local sourcing	- Rise of flexitarian, vegetarian, vegan diets, gluten/lactose-free. - Convenience of products/cooking time - Consumer habits - Image of the products (sustainable, healthy, etc.) - Availability and stability of the supply (implying contracts) - Availability of GM-Free supply - Quality of the grains - EU origin and local sourcing	- Functional and nutritional properties of protein-rich plants. - Rise of the demand for meat alternatives and products free from gluten, lactose, etc. - Availability of GM-free supply - Availability and stability of supply (contracts for peas) - Know-how of companies - R&D for new ingredients - Competition with other protein sources (e.g. gluten, whey)

Various factors will likely impact legume food market in the future. First, in the coming years, flexitarian, vegetarian, and vegan diets are expected to increase, and so the demand for legumes and processed plant-based products. Secondly, the growing diffusion of gluten-free foods will probably enhance the demand for legumes, as they naturally contain no gluten and can be combined with rice or corn to create gluten-free products. Additionally, health and environmental considerations are becoming increasingly important for consumer choices, driving the demand for plant proteins as an alternative to animal proteins. Moreover, the growing demand for food featuring local origin and short cooking time, may encourage

legume cultivation and innovation in processing techniques, while diminishing the interest in dry legumes in the EU (Clément et al., 2018). Finally, the establishment of incentives by the EU aimed at rewarding the environmental benefits of legume cultivation, such as nitrogen fixation and the increase of cultivated biodiversity, would contribute to enhance the profitability of this species.

1.2.2. Legume environmental and agronomic benefits

Climate change adaptation and mitigation appear as the greatest challenges humanity is currently facing, implying the necessity to rapidly implement technical, economical, and political changes in the existing systems (Brondizio et al., 2019). Climate change is the result of human activities based on fossil fuel consumption generating an increase in the atmospheric content of a range of gases, defined as greenhouse gases (GHG). These gases, including carbon dioxide (CO_2), accounting alone for 76% of total GHG emissions, methane (CH_4), nitrous oxide (N_2O), and chlorofluorocarbons (Pachauri and Meyer, 2014), trap heat near the Earth's surface causing temperatures to rise and determining the so-called greenhouse effect (NASA, n.d.). Agriculture, forestry, and other land uses accounted for about 15% of global anthropogenic GHG emissions in 2019 (Ghosh, 2022; Figure 3), with the following contributions: (1) 31.5% from livestock enteric fermentation emitting methane, and decomposition of animal manures under low-oxygen conditions producing both nitrous oxide and methane; (2) 22.3% from the application of synthetic nitrogen fertilizers causing nitrous oxide release from soil; (3) 19% from the burning of agricultural residues releasing carbon dioxide, nitrous oxide, and methane; (4) 12% from net carbon dioxide emissions due to changes in forestry cover; (5) 7.6% from net change in carbon stocks due to cropland management; (6) 7.1% from methane emissions due to rice cultivation; and (7) 0.5% from carbon dioxide release due to grassland degradation (Ritchie, 2020; percentages of GHG emissions from the different sources are based on data of 2016). Moreover, in 2016 an additional 1.7% of human GHG emissions was generated by energy consumption in the agricultural and fishing sectors according to Ritchie (2020), while the total contribution of the food sector was estimated around 26% (Ritchie, 2019; Figure 4). Therefore, reducing GHG emissions from the agricultural sector appears crucial to keep the expected global temperature rise for 2030 below 2°C, as stated in the objectives of the Paris Agreement (United Nations, 2015). In this context, legume cultivation represents an outstanding resource for enhancing the sustainability of the agricultural production (Lüscher et al., 2014; Oliveira

et al., 2021) due to the provision of a wide range of agro-ecosystem services, including: (1) biological nitrogen fixation through symbiosis with soil bacteria of genus *Rhizobium* (Rochester et al., 2001; Jensen and Hauggaard-Nielsen, 2003; Crews and Peoples, 2004) and *Bradyrhizobium* (Crews and Peoples, 2004; Jaiswal and Dakora, 2019), (2) increase in cultivated plant and animal (e.g., pollinators) biodiversity with a positive impact on agricultural system resilience, implying lowered weed, pest, and disease risks (Jensen and Hauggaard-Nielsen, 2003; Köpke and Nemecek, 2010; European Commission, 2018); (3) quantitative and qualitative improvement of the performance of subsequent crops in agronomic rotations (European Commission, 2018); (4) extensive soil coverage when combined with cereals in annual or perennial mixtures leading to a reduction of nutrient run-off into groundwater and rivers; (5) soil structure amelioration due to the capacity of some legume species to decrease soil strength (Rochester et al., 2001; McCallum et al., 2004). For these reasons, legume cultivation can contribute to the achievement of crucial sustainability goals, including: (1) diminishing fossil energy consumption and GHG emissions generated by the synthesis, transport, and application of nitrogen fertilizers and, to a lower extent, of agrochemicals (Häusling, 2011; Jensen et al., 2012); (2) increasing soil carbon sequestration thanks to nitrogen fixation favoring humification processes (Christopher and Lal, 2007); (3) providing a healthy and energy-efficient alternative protein source to animal products for human diets (Iannetta et al., 2021), potentially allowing for noticeable land sparing (Searchinger et al., 2019) and deforestation reduction (Ghosh, 2022; Figure 3); (4) providing an energy-efficient biomass source for different purposes, including bio-refineries and chemical industry (Jensen et al., 2012). Furthermore, increasing legume adoption in agronomic rotations appears crucial for tackling the problem of nitrogen deficiency in organic agriculture, representing the main obstacle to its potential widespread diffusion (Barbieri et al., 2021). Finally, enhancing legume production in the EU would contribute to reduce soybean importations from South America mitigating the problem of agriculture-linked deforestation (Ghosh, 2022; Figure 3), and of GHG emissions generated by food transport (Poore and Nemecek, 2018; Figure 4).

Figure 3. Global greenhouse gas emissions by economic sector for year 2019 (Source: Ghosh, 2022).

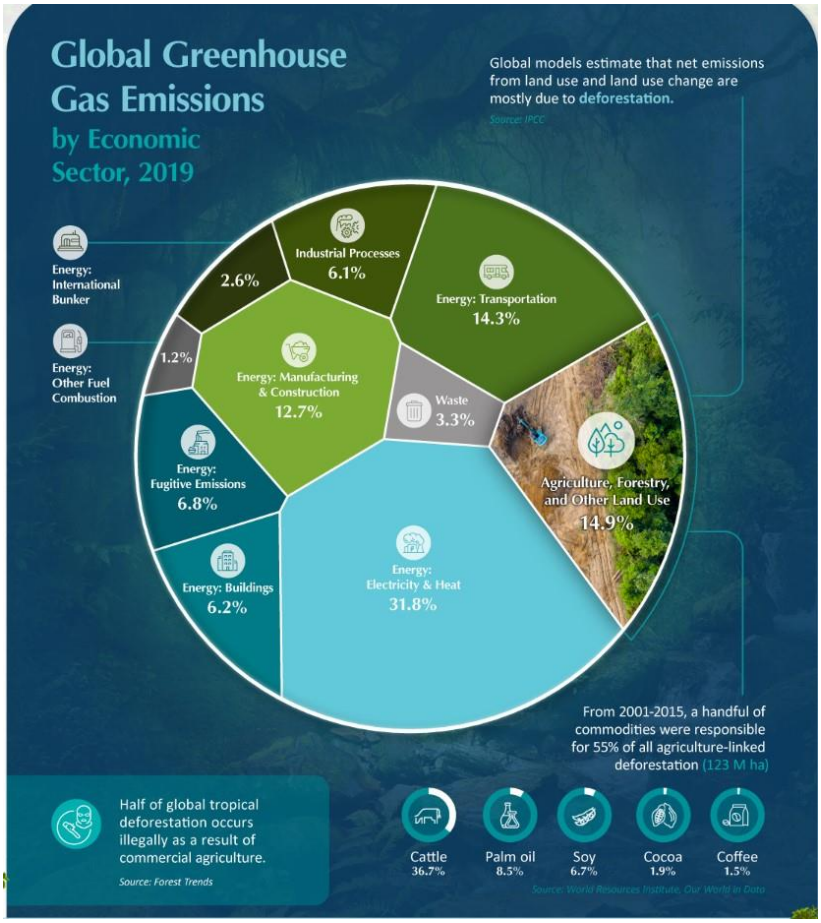
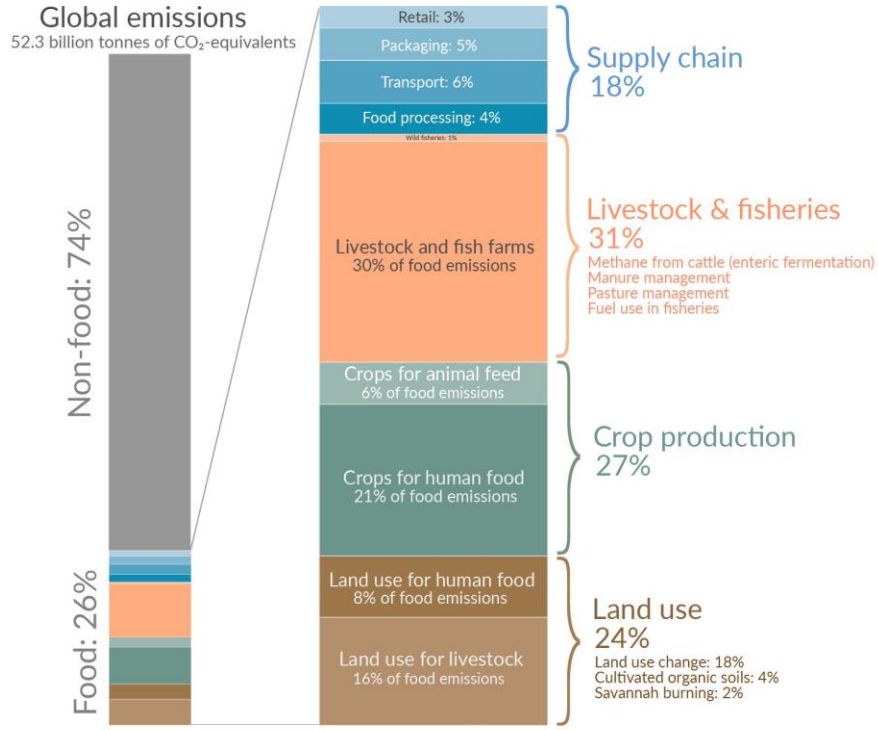


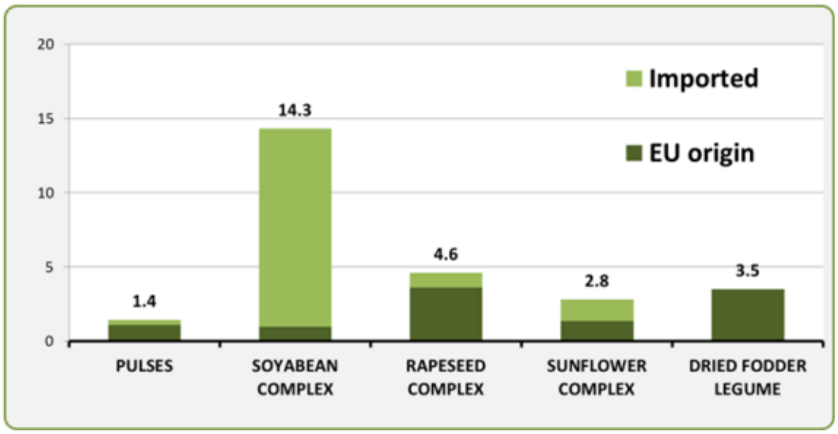
Figure 4. Global greenhouse gas emissions from food production (Source: Poore and Nemecek, 2018).



1.2.3. European Union deficit of feed proteins and related policies

In the last decades, EU livestock industry has expanded to keep up with the increasing demand for animal proteins (FAO, n.d.a), which came at the expense of that for plant-based protein, generating a growing request of high-protein feed (Sepngang et al., 2020). In the same period, EU arable land destined to legume cultivation has declined from 4.7% of 1961 to 1.8% of 2021 (Balázs et al., 2021), under the impulse of international trade agreements. In particular, the General Agreement on Tariffs and Trade (WTO, 1947) and the Blair House Agreement (European Commission, 1993), allowed for the protection of EU cereal production and, at the same time, for duty-free importations of oilseed and protein crops. Obviously, these measures have favored legume importations, implying a substantial reduction of EU investments in legume research and a gradual loss of practical knowledge about the cultivation and processing of these crops by farmers and local businesses, regarding e.g., on-farm selection, storage, processing techniques, and on-farm use as animal feed (Häusling, 2011). Currently, legume cultivars are characterized by a great variability of both yield (Reckling et al., 2015) and economic return (LMC International, 2009), and a considerably lower yield compared with cereal varieties (De Visser et al., 2014; Reckling et al., 2015), stemming from the poor efforts made in legume breeding in the recent decades (Häusling, 2011). In countertendency with other legume species, soybean gross margin has experienced a constant growth in recent years in the EU, while pea, field bean, and alfalfa have featured the lowest gross margins, except for specific value chains generating high prices (e.g., legumes destined to animal feeding for PDO or PGI cheese production). All these factors have contributed to generate the current EU deficit of about 70% for feed protein crops (Figure 5), 87% of which is covered by imported soybean and soybean meal (Clément et al., 2018) mainly from North and South America (Sepngang et al., 2020).

Figure 5. European Union plant protein consumption and sources in million tons of crude protein during 2016-17. "Complex" includes meals, seeds, and beans. (Source: European Commission, 2018).



In South America, massive soybean cultivation has brought about tropical deforestation, contributing to a growth of GHG emissions, a dramatic loss of biodiversity (Wearn et al., 2012), and degradation of soil and water (Maia et al., 2010; Neill et al., 2013). Hence, fostering EU legume production can have a positive impact on the worldwide sustainability of the agricultural sector. Moreover, EU poor self-sufficiency for protein crops implies the exposition of the feedstuff, livestock, meat, and dairy sectors (Häusling, 2011) to trade distortions (Henseler et al., 2013), sustainability (Minderhoud, 2010; Elgert, 2013) and scarcity issues, and price volatility of soybean on the global market (De Visser et al., 2014). In the coming years, the predicted global increase in meat and dairy consumption due to the expected growth in both worldwide population and Gross Domestic Product risks to further exacerbate these problems (De Visser et al., 2014; Clément et al., 2018).

The insufficient internal production of protein-rich feed materials led to an EU Parliament motion in 2011, advocating for long-term policies providing an economic support to legume research and adoption in agronomic rotations. This document identified some prominent reasons, among others, for the competitive disadvantage hitting the EU feed and livestock sectors, including: (1) the different regulatory systems existing for GMOs inside and outside the EU; (2) the need to largely import feed proteins while complying with the zero-tolerance policy on the presence of unapproved GMOs; (3) the under-exploitation of the feed industry potential due to the poor volumes of protein crops produced internally and the severe regulatory constraints on importations (Häusling, 2011). In 2017, the European Soya Declaration signed by fourteen member states highlighted the need to increase the internal production of protein crops and led to the formulation of the EU Protein Plan (European Commission, 2018), aimed at characterizing protein demand and identifying measures to enhance protein crop competitiveness (FEFAC, n.d.). In fact, despite the growing policy interest around legumes (Clément et al., 2018), the EU still lacks specific measures to support their cultivation, which can optionally be established by member states within the Greening Measures (crop diversification), Agri-environment Schemes (i.e., Ecological Focus Areas (EFA) in Pillar 1), or Voluntary Coupled Support (VCS) of Common Agricultural Policy (CAP) (Balázs et al., 2021). However, none of these categories of measures has been really effective in promoting legume adoption, since crop diversification guidelines lack indications about the species to be employed, EFA envisages other options for enhancing biodiversity (e.g., hedgerows, buffer strips, afforested areas) (Bues et al., 2013), and VCS is just a generic incentive to prevent production drops in specific agricultural sectors with an economic, social, or environmental value (Clément et al., 2018). Consequently, national policies

supporting legume cultivation exist in a small number of member states and are relevant only for specific regions within each state (Zander et al., 2016). The new CAP, which has become effective in 2023 and will remain in force until 2027, although still lacking specific measures for legumes, set some mandatory Good Agricultural and Environmental Conditions (GAEC) that may encourage legume adoption in cropping systems. Especially, GAEC on soil protection and quality oblige farms above 10 hectares to adopt crop rotation, which can be substituted by crop diversification only when it contributes to the preservation of soil fertility (European Commission, n.d.), with legumes representing a fundamental resource for both tasks. Anyway, in addition to policies promoting legume cultivation, an increased EU self-sufficiency for protein crops requires the creation of a value chain by infrastructure and market development, sharing of knowledge and best practices, effective market monitoring, and solutions to the fragmentation and inconsistency of the existing funding measures (European Commission, 2018; Balázs et al., 2021).

1.3. Pea (*Pisum sativum* L.)

1.3.1. Taxonomic, botanic, and agronomic characteristics

Field pea (*Pisum sativum* L.) is a self-pollinated, diploid, C3 species from *Pisum* genus belonging to the Fabaceae family (Smýkal et al., 2015), section Vicieae (Baldoni and Giardini, 2001, p. 361). It is native to western Asia, between Turkey and Iraq, and it was likely domesticated between 9,000 and 10,000 years ago as part of the Neolithic crop assemblage (Zohary and Hopf, 1973; Weeden, 2007; Abbo and Gopher, 2017). *Pisum sativum* L. is one of the two species currently attributed to the *Pisum* genus together with *Pisum fulvum*, namely one of its wild relatives, and includes the subspecies *sativum* and *elatius*, containing all cultivated and wild types, respectively (Davis, 1970). Subspecies *sativum* includes two varieties, namely *sativum* and *arvense* (Smartt, 1990, p. 179), which are used for grain and forage production, respectively (Baldoni and Giardini, 1981, p. 320), while subspecies *elatius* contains varieties *elatius*, *pumilio* (or *humile*), and *brevipedunculatum* (Smartt, 1990, p. 179; Table 3). The following information is referred to *Pisum sativum* L. spp. *sativum* var. *sativum*, which includes all the materials employed in the current work.

Table 3. Taxonomy of genus *Pisum* (Davis, 1970; Polhill and Van det Maesen, 1985). (Source: Smartt, 1990, p. 179).

P. *sativum* L.		
	ssp. *sativum*	**var. *sativum***
		var. *arvense*
	ssp. *elatius*	**var. *elatius***
		var. *pumilio* (or *humile*)
		var. *brevipedunculatum*
P. *fulvum*		

Pea plant is characterized by hypogeal germination (Baldoni and Giardini, 2001, p. 361) and is composed by one or several stems, since axillary meristems at the lower nodes can either produce branches or abort (Lake et al., 2021). The leaf consists of two oblong amplexicaul stipules with a waxy cuticle, one or several pairs of oval leaflets with smooth margins, and tendrils at the top of the rachis (Baldoni and Giardini, 2001, p. 361-362; Lake et al., 2021; Figure 6). Field pea displays three main leaf morphologies, that is conventional (Figure 6), semi-leafless, where leaves are replaced by tendrils, or completely leafless, where both leaves and stipules are replaced by tendrils (Uzun et al., 2005; Mikić et al., 2011; Tafesse et al., 2019). The inflorescence is placed at the leaf axil and consists of a pedunculated raceme with 1-3 hermaphrodite flowers displaying small bracts and the typical Fabaceae morphology, that is an orbicular clavate corolla with five petals, the upper one (standard) embracing the two lateral ones (wings), and the bottom one forming a ridge (keel) (Figure 7). The calyx is campanulate with five sepals of different length, the androecium presents ten diadelfian stamens and anthers of uniform length, while the ovary is flattened with an inclined hairy style and 3-10 ovules disposed in two rows (Baldoni and Giardini, 2001, p. 362). Pea is a cleistogamous species, which means it is mostly self-pollinated, even if crossing mediated by pollinators can occasionally occur (Cousin, 1997). Normally, pea has an indeterminate growth habit, but all the varieties employed for industrial cultivation are determinate, which allows for simultaneous maturation (Baldoni and Giardini, 2001, p. 372). In determinate types, leaf formation ceases with onset of flowering (Baldoni and Giardini, 2001, p. 364) and harvesting is performed when plants are 0.35-1.0 m high (Lake et al., 2021). Depending on plant height, pea cultivars can be classified as dwarf, semi-dwarf, or climbing, with the first ones being the most common in intensive cultivation. Pea fruit is a pod of variable length (normally 6-10 cm) and width, hunched or straight, green and sharpened at the ends,

containing 4-10 seeds, and whose section can present different forms. Lower fertile nodes normally produce more numerous and larger pods compared with the upper ones (Baldoni and Giardini, 2001, p. 361-365). Despite the presence of a certain variability for pod indehiscence level (Hradilova et al., 2017), the 'sugar pod' mutation, causing reduced pod parchment and thus contributing to shattering minimization, is largely predominant among cultivated types allowing for flexibility in harvesting time (Siddique et al., 2013; Sadras et al., 2019). The seed colour is determined by the combination of the seed coat and cotyledon colour, normally appearing between green and yellow, while the seed shape is often round, but can also be oval, flattened, squared, hexagonal, or irregular. Dry seed can be smooth, wrinkled, or present an intermediate state (Baldoni and Giardini, 2001, p. 363), and individual seed weight varies approximately from 175 to 300 mg (Sadras, 2007). Pea root system consists in a taproot of variable length from which thin lateral roots depart (Baldoni and Giardini, 2001, p. 361), hosting nodules appointed to biological nitrogen fixation through the symbiosis with *Rhizobium leguminosarum* bv. *viciae* (Jensen, 1986; Voisin et al., 2002), providing nitrogen for about 30-50 kg/ha (Baldoni and Giardini, 2001, p. 366).

Figure 6. Representation of a pea plant. (Source: Alberta Pulse Growers, n.d.).

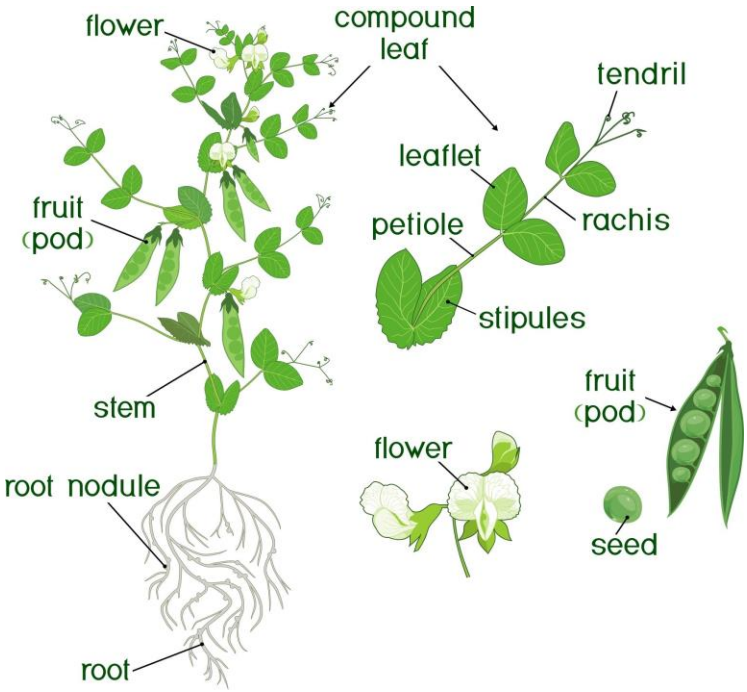
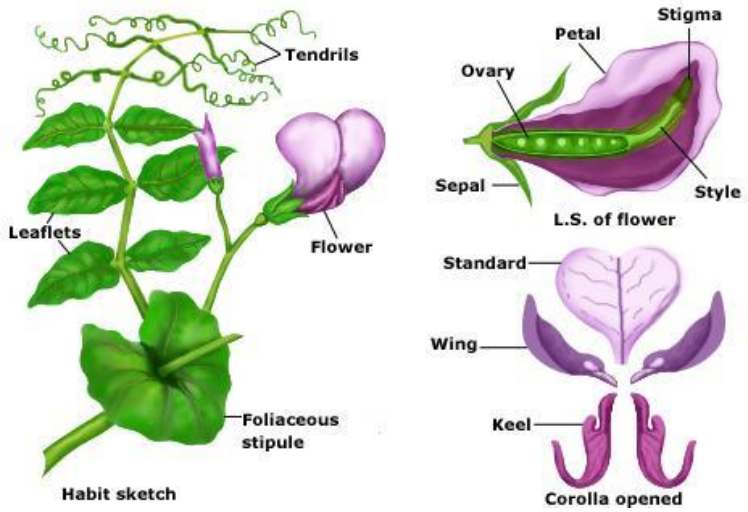


Figure 7. Representation of Fabaceae flower morphology. (Source: Koirala, 2018).



Field pea is cultivated in more than 100 countries (Yadav et al., 2010; Singh et al., 2013), generating a worldwide dry seed production of 12,403,522 tonnes (t, FAO, n.d.b; Figure 8) from 7,978,490 cultivated hectares (ha) in 2021 (FAO, n.d.b), and can rely on over 90,000 accessions conserved in gene banks (Yadav et al., 2010; Singh et al., 2013). The top producers and exporters are Canada and Russia, while the top importers are India, Pakistan, Bangladesh, and China, although the latter is among top producers (FAO, n.d.c; Figure 8). Field pea is mainly grown to produce dry grains that can be used as food or feed for monogastric or polygastric animals, and, to a lower extent, fresh seed (4,017,704 ha in 2021, FAO, n.d.b), which can either be consumed directly or conserved by freezing or canning, and fodder. The main outlet for pea grown in Europe, north America, and Australia is animal feed, but the demand for human food has been expanding considerably since the 2000s (Lake et al., 2021). Cultivars with yellow grains are mostly employed by the feed sector or for functional ingredient production, while varieties with light green grains and wrinkled dark green seeds are preferred for canning and freezing, respectively. In the EU, five pea classes are identified for commercial purposes based on seed diameter, going from “extra-fine” to “medium” with diameter lower than 7.5 mm and higher than 10.2 mm, respectively. Seed composition is influenced by climate and soil (Nikolopoulou et al., 2007) and on average consists in 26% of protein, 67% of starch, 2% of fat, 2% of fibre, and 3% of other elements, including calcium, iron, and phosphorus. Like other legume species, pea is characterized by a low content of sulphur amino acids, namely methionine and cysteine (Baldoni and Giardini, 2001, p. 370-371). Moreover, pea grains contain several active compounds, including polyphenolics, vitamins, saponins, galactose oligosaccharides (Dahl et al., 2012), and tannins, which were found in coloured peas (Bastianelli et al., 1998). Although their

concentration tends to be low in cultivars, pea like other legumes contains trypsin inhibitors (Cousin, 1997), namely proteins suppressing the activity of pancreatic enzymes trypsin and chymotrypsin, thereby causing a reduction in the digestion and absorption of dietary proteins (Gemede and Ratta, 2014). These compounds can be denatured by heat (Lake et al., 2021), but their presence remains problematic especially for animal feeding since the seed is normally employed directly without pre-processing (Duc et al., 2015). Mean pea yield in the last century was estimated at 1.7 t/ha globally, whereas potential yields may exceed 6 t/ha (Smýkal et al., 2015). Mean yield improvement per year amounted to 16 kg/ha between 1961 and 2017 (Figure 9), while that of wheat reached 40 kg/ha (FAO, n.d.c). Noticeably, the mean rate of yield improvement per year between 1961 and 1990 was of 21 kg/ha, afterwards experiencing a considerable decline due to the reduced cultivation interest leading to a drop in research investments (Lake et al., 2021). For pea destined to fresh consumption or conservation by canning or freezing, grain yield can range between 9 and 12 t/ha, whereas the cultivation of indeterminate types by trellis can ensure higher yielding performance (Baldoni and Giardini, 2001, p. 370).

Figure 8. Worldwide pea production of dry seed during 2021 measured in tonnes (t). (Source: FAO, n.d.b).

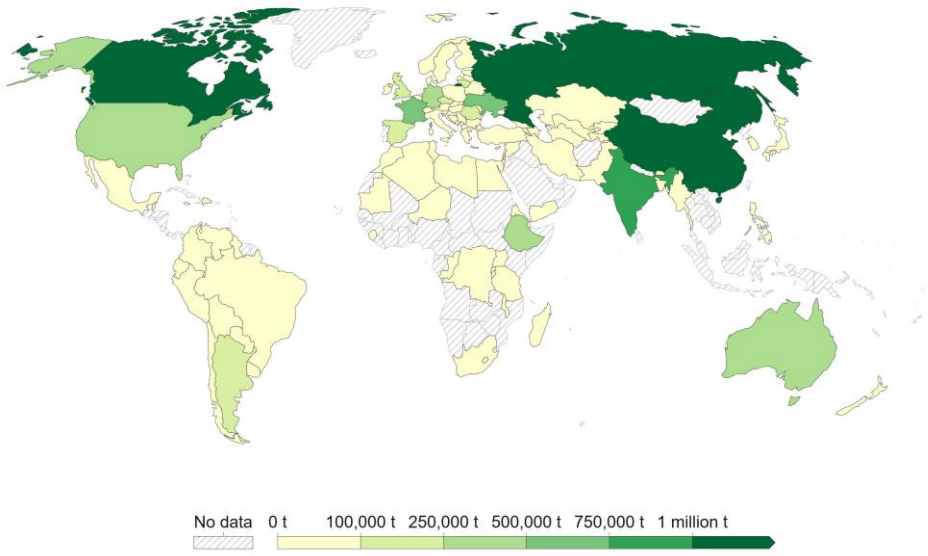
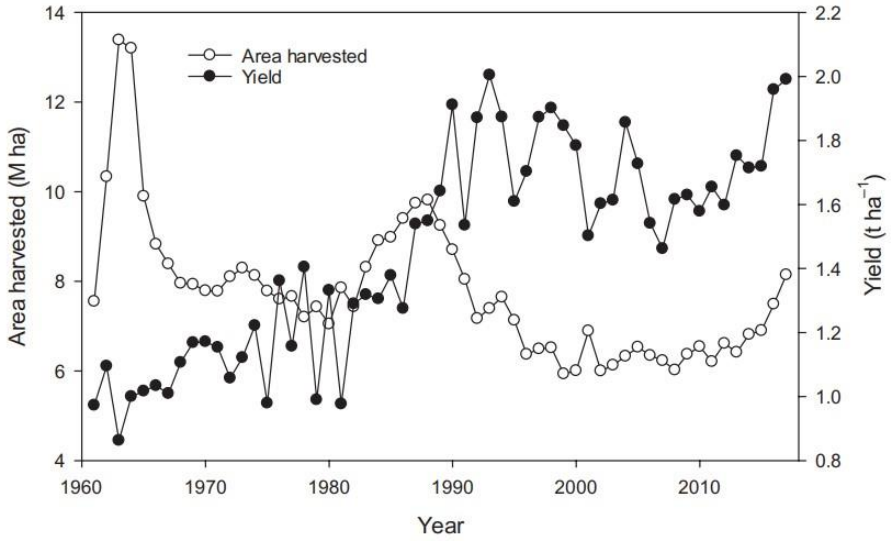


Figure 9. Global pea cultivated surface and grain yield from 1961 to 2017. (Source: FAO, n.d.d).



Pea can be grown in rotation with cereals and/or oilseeds both as a single crop or in multi-crop systems. In temperate regions, it can be cultivated either as a spring-sown crop in environments characterized by very cold winters, such as northern Europe, Canada, and parts of the United States (US), or as an autumn-sown crop in Mediterranean-like environments featuring milder winters, such as southern Europe and Australia (Lake et al., 2021). In India and Bangladesh, it is sometimes grown as a relay crop in the dry season between successive rice crops (Ali and Sarker, 2013). Intercropping with canola, faba bean, wheat, or maize, was often observed to lead to improved yield and yield stability when compared to equivalent areas of the single component crops (Stelling, 1997; Soetedjo et al., 1998; Tan et al., 2020). Field pea, either grown as a single crop or in combination with other species, can be used as a catch or cover crop, or as green or brown manure. Pea inclusion in agronomic rotations based on cereals and/or oilseeds can provide several benefits, such as biological nitrogen fixation, the supply of nitrogen-rich crop residues, improved phosphorus availability (Ha et al., 2007), broader weed control options, reduced pest, disease and weed risk, soil structure improvement, size and diversity increase of microorganism population (Lake et al., 2021). Indeed, field pea was consistently observed to generate remarkable yield gains in the subsequent wheat crop compared with continuous cropping, e.g., in United Kingdom (Vaidyanathan et al. 1987) and Australia (Seymour et al., 2012; Angus et al., 2015). For these reasons, pea adoption in agronomic rotations can lead to a decrease in the use of pesticides and synthetic nitrogen fertilizers, contributing to improve the profitability of the agricultural systems and providing opportunities in food and feed markets (Lake et al., 2021). In this regard, pea cultivation was estimated to be more profitable than that of standard milling wheat

in France due to lower operational costs (30-100 €/ha less in 2017), higher market price (20%–45% more over 2012–16) (Terres Inovia, 2017), and the presence of EU VCS incentive, while in Australia the scenario was the opposite due to wheat yield being about twice that of pea in low-rainfall farms (Lake et al., 2021). In the US, the Farm Act ensures a minimum return for pea cultivation, acknowledging its role in agricultural system sustainability (Yadav et al., 2010). Furthermore, pea is particularly suited to minimum or no tillage systems since it produces a low amount of straw, therefore allowing for an effective establishment of the following crop. Finally, field pea cultivation, especially as a winter crop, may favour work division and diversification of management practices. Indeed, sowing and harvesting are performed later compared with the main autumn-sown crops grown in Europe, and pea variability for cycle length in response to the external conditions enables its cultivation in different environments. For instance, in Australia pea can be sown later than other winter crops facilitating disease and weed management, although late sowing tends to decrease yield potential, and it is harvested earlier than cereals, which can reduce yield losses from terminal drought stress and pod shattering (Lake et al., 2021).

In Europe, soil preparation consists of autumn ploughing to 0.3 m followed by harrowing. Sowing is performed by a cereal seed drill with row spacing of 0.18-0.22 m in the period from September to April (Baldoni and Giardini, 2001, p. 367-368), most frequently around mid-November (Karkanis et al., 2016). The environment determines the sowing time and the consequent choice of cultivar phenological type, that can be (1) spring, (2) classical winter, or (3) winter 'Hr' type, with (1) and (2) being day neutral, while (3) is highly responsive to photoperiod (Lejeune-Hénaut et al., 2008; Bénézit et al., 2017). Autumn-sowing ensures higher yielding potential compared with spring-sowing, leading to an estimated yield increase of 56% in Italy (Annicchiarico and Filippi, 2007), thanks to the longer cropping cycle, higher radiation use efficiency in early spring, and drought escape during grain filling (Stoddard et al., 2006; Urbatzka et al., 2011). Seed dose amounts to 100-130 seeds/m^2 and sowing depth to 3-4 cm. In the case of pea destined to the transformation industry, either several varieties featuring different cycle lengths are sown at the same time or scalar sowing is employed for the same cultivar, to extend the period of product supply to the processing plants. Chemical weeding can be performed before sowing or pre- or post-emergence, while irrigation is normally not employed. Pea meant to dry seed production is harvested when grains reach a humidity of 18-24% by a cereal combine, and then is desiccated to 13% humidity to increase storage life. For crops destined to the transformation industry, harvesting is performed by dedicated machines and the correct timing is established by measuring seed hardness, which

is proportional to starch concentration and inversely correlated to sugar content, due to sugars being converted into starch during seed maturation. A lower grain hardness is required by the freezing compared to the canning industry.

Optimal temperatures in the initial cultivation phase range between 10°C and 20°C (Baldoni and Giardini, 2001, p. 366-370) with base germination temperature being around -1.1°C (Raveneau et al., 2011), whereas higher temperatures can cause cycle acceleration leading to seed quality deterioration. Spring frosts featuring temperatures lower than -4°C can damage pea vegetative system possibly causing plant death (Baldoni and Giardini, 2001, p. 366), especially at the seedling stage (Meyer and Badaruddin, 2001), while after winter hardening pea can tolerate rhizosphere temperatures as low as -8.5°C (Murray et al., 1988, p. 831-843). Like most legumes, pea is sensitive to freezing temperatures, particularly at the flowering, early pod formation, and seed filling stages (Maqbool et al., 2010). Moreover, it is susceptible to soil compaction (Siczek et al., 2013), water logging, and calcareous or saline soil, with optimal pH being around 5.5-6.5. Drought and heat stress occurring from onset of flowering to the end of grain filling can severely diminish yield potential, due to reduced flower fertilization (Baldoni and Giardini, 2001, p. 364-366) and/or photosynthesis rate (Annicchiarico et al., 2017a).

Several biotic stresses can limit pea cultivation and productivity. The main fungal diseases are anthracnose, which is caused by three diverse types of fungi (*Ascochyta pisi*, *Mycosphaerella pinodes*, and *Ascochyta pinodella*) and is transmitted by infected seeds or crop residuals, and *Fusarium oxysporum pisi*, a terricolous fungus for which resistant varieties are available causing plant death or yield losses by early or late infections. Other common fungal pathogens are *Pythium debaryanum* and *ultimum*, causing plant rot at early stages, *Botrytis cinerea*, which can be particularly detrimental during pod formation and filling, *Erysiphe polygoni*, *Septoria pisi*, and *Uromyces pisi*. All these fungal diseases are favoured by the presence of humid conditions. Among bacterial pathogens, *Pseudomonas pisi* is especially harmful since it can damage all the vegetative organs and the seeds, through which it can be transmitted to the following generations. Seed coating, rotations, and the use of sane seed and resistant varieties are the most common means to counteract these diseases. Several viruses can infect pea, the more frequent being Pea Common Mosaic Virus and Top Yellow Virus. Three species of insect pests are especially relevant for pea, that is pea weevil (*Bruchus pisorum;* Figure 10) and *Laspeyresia nigricana*, namely a coleopter and a budworm whose larvae feed on seeds, and pea aphid (*Acyrthosiphon pisum*), which can transmit viruses

and cause yield losses by damaging flowers (Baldoni and Giardini, 2001, p. 373). Among holoparasitic plants, crenate broomrape (*Orobanche crenata* Forsk.) can severely hamper pea cultivation in the Mediterrean region and eastern Asia (Rubiales et al., 2003). Chemical treatments and warehouse disinfestation, and pre- or post-emergence weeding are the most common control measures employed for insect pests (Baldoni and Giardini, 2001, p. 373; Ferrari et al., 2006) and broomrape (Rubiales et al., 2003), respectively.

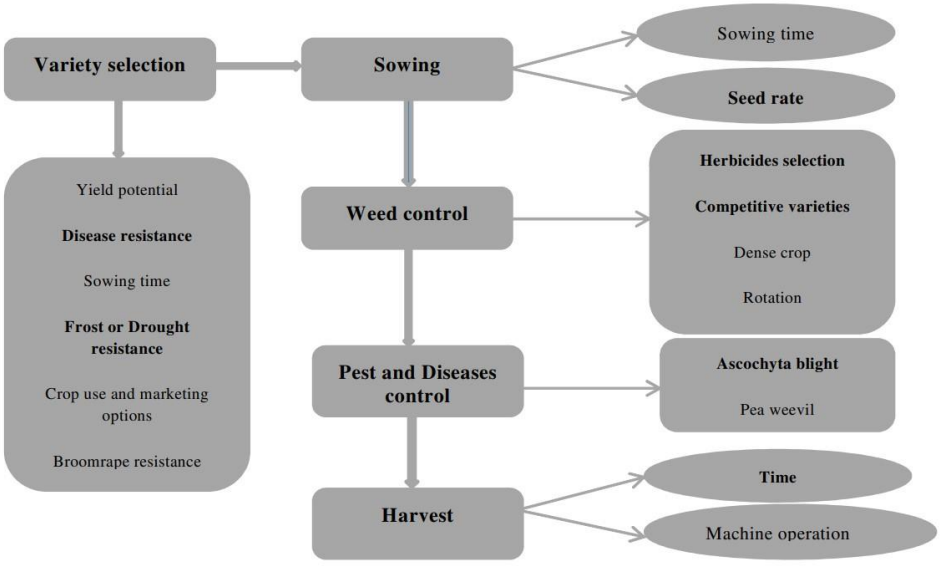
Figure 10. Pea weevil **(***Bruchus pisorum*, left) and its damage (right, from left to right: opened exit hole, emerging adult, and unopened exit hole). (Source: Cesar Australia, n.d.).

1.3.2. Pea potential for European agriculture and major breeding goals

Field pea is the most widely grown grain legume in Europe due to its higher yielding potential compared with other cool-season grain legumes in the western (Carrouée et al., 2003) and southern region (Annicchiarico, 2008), and a moderately good rate of yield genetic progress (Annicchiarico, 2017). Pea diffusion in this area is motivated, among others, by its adaptation capacity to a broad range of climates, from continental to semiarid, thanks to different cultivar phenological types that allow for spring-sowing in central and northern Europe minimizing the risk of frost damage, and autumn-sowing in the south (Karkanis et al., 2016). In Europe, human consumption of pulses is lower than in other regions of the world (Schneider, 2002), with pea being one of the most popular legumes in local diets (Karkanis et al., 2016). Furthermore, pea use as livestock feed is common for monogastric animals in Europe (Brenes et al., 1989; Gatel and Grosjean, 1990), and can further expand, since its potential to effectively substitute soybean in animal diets was demonstrated for beef (Anderson et al., 2007; Soto-Navarro et al., 2012) and dairy cattle (Khorasani et al., 2001; Volpelli et al., 2009), and meat lambs (Lanza et al., 2003). An even partial replacement of soybean by pea for animal feeding in Europe would help to mitigate the several issues related to soybean cultivation and importation, especially from south America, as described in paragraph *1.2.3*.

Moreover, pea cultivation may contribute to alleviate soil organic matter deficiency in southern Europe thanks to legume capacity to improve soil fertility and properties (Carranca et al., 1999; Piotrowska and Wilczewski, 2012), considering that in this area yield gap with cereals is less pronounced compared with the rest of the continent (Nemecek et al., 2008; Stoddard, 2013) due to autumn-sowing extending cycle length (Karkanis et al., 2016). In addition, pea lower yielding ability compared to other crops, such as wheat, may contribute to generate higher prices, without considering other long-term benefits of its inclusion in agronomic rotations (Karkanis et al., 2016). The growing market interest for organic agriculture represents a further opportunity to expand pea cultivation (Barbieri et al., 2021), despite a yield penalty of 10-14% compared with conventional management (Gopinath et al., 2009) due to increased pest, disease, and weed infestation risk (Corre-Hellou and Crozat, 2005). The main factors affecting pea cultivation in Europe are summarized in Figure 11.

Figure 11. Main factors affecting pea cultivation in Europe. (Source: Karkanis et al., 2016).



Breeding progresses achieved in the past decades have contributed to increase the agronomic interest around pea, e.g., by improving standing ability through the incorporation of recessive *afila* alleles converting leaflets into tendrils (Snoad, 1974; Davies, 1977; Hedley and Ambrose, 1981), introducing resistance to powdery mildew (*Erysiphe pisi*, Harland, 1948), and developing determinate genotypes (Marx, 1977) with a high number of flowers per node (Hardwick et al., 1979). Nevertheless, yield instability due to biotic and abiotic stresses (Sagan et al., 1993; Cousin, 1997) together with the lower yield compared with other crops, such as wheat, which was reported to produce more than twice in Germany, Spain, and UK (Nemecek et al., 2008; Stoddard, 2013), represent major constraints to pea extensive adoption

(Karkanis et al., 2016). As regards plant architecture, pea variable standing ability still represents the overriding problem, despite the improvement generated by the introduction of semi-leafless genotypes. In this sense, a deeper understanding of the factors affecting stem mechanics may possibly lead to further advancements (Ambrose, 2008, p. 14-15). Among abiotic stress factors, drought, which is often combined with heat stress (Ranalli, 1995), constitutes the main limitation to pea cultivation worldwide, being particularly harmful in the Mediterranean area during seed development phases (Annicchiarico et al., 2017a). Drought stress is expected to become more frequent in the Mediterranean area and to expand northward and eastward due to climate change (Alessandri et al., 2014), enhancing the importance of finding tolerant genotypes that can rely either on drought escape via early flowering (Fang and Xiong, 2015), which is crucial in Mediterranean-like environments allowing autumn-sowing (Turner et al., 2001), or intrinsic tolerance (Fang and Xiong, 2015). Autumn-sown cultivars, besides ensuring a higher yielding potential by cycle extension and terminal drought escape, may also contribute to increase pea adoption in regions characterized by a continental climate (Duc et al., 2015). In this context, cold tolerance emerges as a major breeding goal not only for inland areas, but also for the Mediterranean region, where plants partially hardened or de-hardened may suffer from frost events of even limited duration and severity (Annicchiarico and Iannucci, 2007), advocating for cultivars with an elevated freezing tolerance and a slow de-hardening (Vocanson and Jeuffroy, 2007). As regards biotic stresses, paramount breeding targets are represented by resistance to *Aphanomyces euteiches*, *Peronospora viciae*, anthracnose, aphids (*Acyrthosiphon pisum*), and bruchids (*Bruchus pisorum, Bruchus affinis*) (Ambrose, 2008, p. 15). Since grain protein content of pea commercial cultivars is modest, usually ranging between 22 and 26% on a dry-matter basis, its increase represents a major breeding objective for both animal feeding and human consumption (Burstin et al., 2011; Duc et al., 2015). Besides the common need to minimize trypsin inhibitor content (Duc et al., 2015), other seed qualitative traits may be relevant for specific market outlets, such as a high sucrose concentration and its maintenance over a long time for varieties meant to the freezing industry.

Due to pea strict self-pollination, the pedigree and single seed descent systems have been the most common breeding methods employed to generate pure line varieties, while bulk selection, which exploits natural selection to produce pure line cultivars, mixtures, or evolutionary populations, has been used less frequently. Recurrent backcross and, after the advent of molecular markers, marker-assisted selection (MAS), have largely been used for the introduction and selection of favourable alleles for monogenic or oligogenic traits, e.g.,

disease or lodging tolerance and seed quality (Ambrose, 2008, p. 18). For the improvement of polygenic traits, such as grain yield, PS has been used traditionally, while GS (Meuwissen et al., 2001) was introduced more recently. Although a specific legal framework for cultivars produced by gene editing is still missing in the EU (Vanderschuren et al., 2023), this technology (Doudna and Charpentier, 2014; Li et al., 2023) together with pea genome sequencing (Kreplak et al., 2019; Yang et al., 2022) has opened new breeding scenarios by greatly enhancing the potential for molecular marker identification, and gene detection and mutation.

1.3.3. Genetic and genomic resources

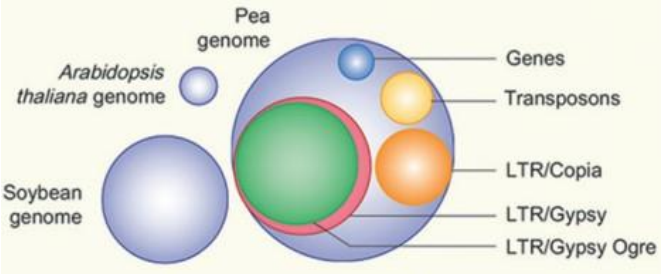
A broad range of public genetic resources are available for pea due to the ease in the production and maintenance of inbred lines, the early domestication process, and the widespread cultivation, including several *ex-situ* germplasm collections (Table 4) that, in the absence of a dedicated CGIAR global institution, are coordinated by the International Consortium for Pea Genetic Resources. Several core collections have been established with different objectives and selection criteria, e.g., the distance of geographical origin according to passport information (Knüpffer and Van Hintum, 1995) or the dissimilarity based on molecular marker data (Elshire et al., 2011; Taranto et al., 2018; Singh et al., 2019). Moreover, the early interest in pea morphological variability has resulted in large mutant collections, such as that initiated by Lamprecht (Blixt, 1963; Lamprecht, 1974), to which mapping populations, near isogenic lines, and ancient material selected by seed saver organisations (Seed Savers Exchange, n.d.) have added more recently (Ambrose, 2008, p. 10-11). Despite this remarkable body of genetic variation, its availability in a form that can be readily used in breeding programs, and the clarification of the underlying structure, drivers (Ambrose, 2004), and relationship with phenotypic variation for traits of interest (Crosta et al., 2023) still represent relevant open issues.

Table 4. *Ex-situ* germplasm collections with more than 1000 *Pisum* accessions. (Source: Ambrose, 2008, p. 10).

FAO Institute code	Country	Number acces-sions	Web site for Germplasm searches
ATFC	Australia	6567	http://www2.dpi.qld.gov.au/extra/asp/AusPGRIS/
SAD	Bulgaria	2787	http://www.genebank.hit.bg/
ICAR-CAAS	China	3837	http://icgr.caas.net.cn/cgris_english.html
GAT	Germany	5336	http://fox-serv.ipk-gatersleben.de/
BAR	Italy	4297	http://www.ba.cnr.it/areagg34/germoplasma/2legbk.htm
CGN	The Netherlands	1008	http://www.cgn.wur.nl/pgr/
WTD	Poland	2899	http://www.ihar.edu.pl/gene_bank/
VIR	Russia	6790	http://www.vir.nw.ru/data/dbf.htm
ICARDA	Syria	6105	http://singer.grinfo.net/index.php?reqid=1151843332.3126
NGB	Sweden	2724	http://www.ngb.se/sesto/index.php?scp=ngb
JIC	UK	3194	http://www.jic.ac.uk/GERMPLAS/pisum/index.htm
USDA	USA	3710	http://www.ars-grin.gov/npgs/searchgrin.html

Pea genome is diploid, spans over the considerable length of 4.45 gigabases, and is organised into seven chromosomes (2n = 14) showing extended primary constrictions due to the presence of multiple domains of centromeric chromatin. Repeated sequences are largely diffused, since transposable elements constitute more than 83% of the genome, with long-terminal repeat retrotransposons (LTR) representing 72.7% (Figure 12). The most recent whole genome duplication event in pea history is that associated to the crown Leguminosae divergence dating back to about 55 million years ago. Since pea divergence from other tribes, its genome has been subject to higher point mutation and gene rearrangement rates compared with the other sequenced legume genomes, which is likely due, at least in part, to the abundance of LTR mostly from Gypsy Ogre group (Burstin et al., 2020). In addition to the first reference genome published by Kreplak et al. (2019), a second *de-novo* assembly was generated by Yang et al. (2022), featuring better continuity and quality in repeated regions, together with a pangenome of 118 wild and cultivated accessions, while exome resequencing was performed for 240 genotypes either selected from a wider germplasm collection (Siol et al., 2017) or chosen based on phenotype by Aubert et al. (2023).

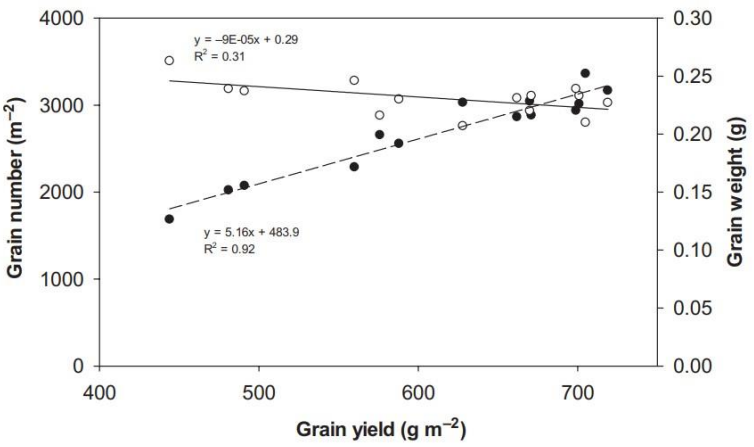
Figure 12. Representation of pea genome size (proportional to circle area) compared to other species, and of the contribution of different genetic components in terms of genome length. (Source: Burstin et al., 2020).



1.3.4. Grain yield and protein content: physiological and genetic control, and relationship

Pea grain yield appears more closely related to seed number compared with seed weight (Doré et al., 1998; Poggio et al., 2005; Sadras, 2007; Sadras et al., 2013; Figure 13), with the former trait showing a linear (Guilioni et al., 2003) or curvilinear (Sadras et al., 2013) relationship with growth rate during a critical period including flowering (Lake et al., 2021). Nonlinearity may suggest a decoupling between vegetative and reproductive biomass growth under conditions favouring early canopy closure (Sadras et al., 2013), possibly due to altered light quality inside a dense canopy causing flower abortion (Lake et al., 2021). The number of seeds produced for each fertile node, determined by both pod number per node and seed number per pod, was observed to affect yield more when pea was subject to drought stress during flowering compared with regular water availability conditions, due to a reduction in the number of fertile nodes under drought (Munier-Jolain et al., 2010). Any factor decreasing the growth rate during the period from onset of flowering to the beginning of seed filling, including water deficit, heat and cold stress and nitrogen deficiency, can diminish the final seed number (Guilioni et al., 2003). Seed weight, which is essentially determined by seed growth rate and duration, is more affected by the genetics and less responsive to the environment compared with grain number (Lemontey et al., 2000). Seed growth rate is determined by the number of cotyledon cells (Munier-Jolain and Ney, 1998), which is fixed during embryogenesis and depends on the embryo trophic conditions (Weber et al., 1996; Lemontey, 1999), while seed filling stops when the available resources are exhausted or maximal seed size is reached (Burstin et al., 2007). As seed cell divisions span from flowering to the beginning of seed filling, any stress occurring in this period can impair seed weight (Lake et al., 2021).

Figure 13. Least squares regression lines for grain yield vs. seed number (closed circles) or seed weight (open circles) for the semi-leafless pea cultivar Nitouche cultivated in southern Chile. (Source: Sandaña and Calderini, 2012).



Seed protein content depends on the relative accumulation of starch and protein in grains (Lhuillier-Soundele et al., 1999) that is affected by both nitrogen availability during seed filling (source limitation), and embryo capacity to accumulate storage compounds (sink limitation) (Burstin et al., 2007). In legumes, nitrogen source capacity is influenced by both symbiotic fixation and soil assimilation (contributing 80% and 20% of the nitrogen acquired under favourable conditions, respectively, according to Salon et al., 2001) in proportions that vary with environmental conditions, such as water and mineral nitrogen availability, temperature and soil structure (Sprent et al., 1988; Salon et al., 2001), impacting both the plant status and the activity of nitrogen-fixing bacteria (Voisin et al., 2003a, b). During seed filling, assimilates are mainly destined to grains, while afterwards nitrogen is remobilized from vegetative organs (Schiltz et al., 2005) in correspondence of photosynthetic machinery degradation (Sinclair and de Witt, 1975, 1976) accounting for a considerable proportion of total seed nitrogen (Sinclair and de Witt, 1975; Schiltz et al., 2005). An example of mechanism controlling seed nitrogen storage capacity is that regulated by the *R* gene, which, if present in the recessive form, causes subnormal grain starch accumulation (Wang and Hedley, 1985; Turner et al., 1990; Craig et al., 1998, 1999), together with increased protein content, and wrinkled shape (Burstin et al., 2007). Finally, seed protein concentration showed a positive correlation with temperature sum (Karjalainen and Kortet, 1987), and phosphorus and nitrogen fertilization (Sosulski et al., 1974), while resulting somehow negatively correlated with precipitation during summer (Karjalainen and Kortet, 1987).

Since the selection for enhanced grain protein content is likely to be performed concurrently with that for higher grain yield, determining the genetic relationship between these traits is crucial for breeding purposes. To our knowledge, genetic correlation was only assessed by Crosta et al. (2022), resulting mostly non-significant and thus encouraging the simultaneous improvement of these traits, while phenotypic correlation estimates varied from negative (around -0.4 in Tar'an et al., 2004 and Krajewski et al., 2011; −0.11 according to Klein et al., 2020, but with large variation across populations and environments) to null (Cousin et al., 1985; Bărbieru, 2021).

Although many Quantitative Trait Loci (QTLs) have been detected in several studies for both grain protein content (Irzykowska and Wolko, 2004; Tar'an et al., 2004; Burstin et al., 2007; Krajewski et al., 2011; Klein et al., 2014; Gali et al., 2019; Klein et al., 2020) and yield or its components (Irzykowska and Wolko, 2004; Tar'an et al., 2004; Burstin et al., 2007; Krajewski et al., 2011; Klein et al., 2014; Gali et al., 2018; Gali et al., 2019; Klein et al.,

2020), the frequently inconsistent genomic positioning and the small or unstable proportion of variance explained support the hypothesis of a polygenic control, in accordance with the complexity of the underlying physiological mechanisms (Burstin et al., 2007).

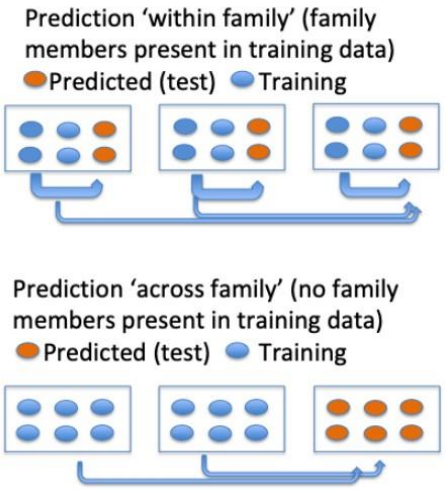
1.4. Genomic selection

1.4.1. The technique

GS was initially conceived by Meuwissen et al. (2001) for animal breeding and only subsequently adopted by plant breeders. The core ideas behind GS are: (1) the formulation of a prediction model based on all the available genome-wide markers, whose effects can be estimated even if their number exceeds that of observations by considering them as random effects, and (2) the estimation of individual breeding values (genomic estimated breeding values or GEBV), namely trait values attributable solely to the genetic component that are computed as the sum of the products between marker genotypes and the relative effects. The logics behind (2) is that, if marker density is high enough relative to linkage disequilibrium (LD) decay pattern in the population, there must be a marker linked to each QTL for a specific trait, so that all QTLs can be tracked simultaneously. GS model performance must be assessed on an independent dataset compared with that used for model training, a process that is defined as cross-validation. Indeed, when the predictors are more numerous than the observations, as in the case of most GS models, a perfect fit can be obtained due to an excessive adaptation of the model to the training data, a problem that is referred to as over-fitting (Meuwissen et al., 2001; Heffner et al., 2009). An important issue in cross-validation is the correction for environmental effects, which allows to reinforce the intensity of the "genetic signal". This is even more important when breeding data is unbalanced over environments for genetic effects, meaning that the genotypes are not the same in all the environments (e.g., because only the material selected basing on previous year data is tested the following year), so that environmental effects can mask or bias genetic effects. In plant breeding, the most common procedure consists in employing a model with genotype effect, which can either be fixed or random with no correlation structure, and environmental effects (e.g., year and location) that are usually considered as random if the genotype is taken as fixed or vice versa. Best linear unbiased estimators (BLUEs) and best linear unbiased predictions (BLUPs) are the solutions obtained when genotype is fixed or random,

respectively, and are employed for GS model training and validation. An alternative, when using statistical software that accept replicate data, is estimating environmental and genetic effects simultaneously to correct for the former without biasing the latter. In this case, the correction for environmental effects can be performed only in the training set or both in the training and validation sets, with the second option resulting in higher predictive ability values. Another important step is the definition of the cross-validation scheme, which determines the criterion for the division of data in training and validation set that can be either stratified or random, meaning that it can consider or not data grouping based on, e.g., environment, population etc. Three criteria can be employed, namely single data split following a stratified scheme by using, e.g., one or more environments or populations for validation, k-fold cross-validation, envisaging data division in k, random or stratified, groups of equal size with each group serving by turn as the validation set (Figure 14), or Leave-One-Out schemes using by turn single individuals for validation. The optimal cross-validation scheme depends on the breeding application, with random schemes being usually less relevant for practice, while structured ones can be more representative of real-life scenarios, considering that the presence of relatives of validation individuals in the training set and the availability of larger training sets normally improve predictive ability. GS model performance is normally evaluated by predictive ability, namely the correlation between GEBV and phenotypic data expressed as BLUEs or BLUPs or corrected for environmental effect, while predictive accuracy represents the correlation between GEBV and the true breeding values and is obtained by dividing predictive ability by its potential maximum (which is equal to the square root of heritability when a single phenotype is associated to each genotype). GS models can account for G × E, dominance or epistatic effects, or covariance of genomic effects for multiple traits (multi-trait GS models), if these components are considered relevant for a specific breeding scenario. The main advantages of GS compared with PS can be expressed by referring to breeders' equation parameters as: (1) reduction of the time necessary to perform one selection cycle, thanks to early selection relying on genomic information and accelerated generation advancement; (2) increase of selection intensity thanks to the lower individual evaluation time and, possibly, cost; (3) improvement of selection accuracy for similar times and costs due to the use of genomic information (Janss and Ramstein, 2023).

Figure 14. Representation of two stratified cross-validation schemes, in the former two individuals from each population are included in the validation set, in the latter one population is employed as the validation set. (Source: Janss and Ramstein, 2023).



As regards the possible pipelines for GS implementation in breeding programs featuring a low or intermediate investment level, it is crucial to consider that the development of the training set and that of the population to be evaluated should proceed in parallel to maximize the time gain relative to PS. For instance, two scenarios can be hypothesized, one assuming intermediate or high economic means in which one or more growth chambers and greenhouses are available, and the other one featuring a lower investment level and relying only on greenhouses. In the first scenario, the subset of lines forming the training set can be advanced through generations by using growth chambers allowing to perform at least four generations per year, while the rest of material can be advanced at the rate of two generations per year in greenhouses. In the lower budget scenario, greenhouses can be employed to advance the GS training set, whereas the other material can be grown in the field performing one generation per year. Moreover, mass selection based on agronomic, or disease resistance traits can be applied on the selection target material after the initial selfing generations of RIL population construction to eliminate genotypes with undesirable characteristics. If enough resources for genotyping are available, GS based on Optimal Haploid Value (Daetwyler et al., 2015) can be applied on segregating individuals from early generations in RIL population construction, to minimize the duration of the selection cycle. Finally, having a genetically wide GS training population, e.g., formed by factorial crossings between a moderately large set of parental lines, as performed in pioneer studies in white lupin (Pecetti et al., 2023), may reduce the need for population-specific models possibly allowing for GS continuous application relying on regular model updating based on new phenotypic data.

1.4.2. Statistical models

The first GS model was introduced by Meuwissen et al. (2001) under the name of BLUP, and now is still commonly used and referred to as whole-genome random regression or ridge-regression BLUP (rrBLUP). This model can be written in the most general form as:

$$y = Xa + Wb + e$$

where X is a matrix of eventual environmental covariates with one row per individual and one column per covariate, a is the vector of eventual environmental effects, W is the matrix of individual marker genotypes with one row per individual and one column per marker, b is a vector of marker effects, and e is a vector of residuals. All marker effects b_i are fitted simultaneously, since they are considered as random effects coming from the same normal distribution and thus featuring a common variance, which can be written as:

$$b_i \sim N(0, \sigma^2_b)$$

where σ^2_b represents mean marker variance (Janss and Ramstein, 2023).
Since a model with a large set of predictors can be transformed into a model based on a similarity matrix, this property was exploited to reformulate rrBLUP into a different form named as genomic BLUP (GBLUP) that directly models the individual breeding values ($g = Wb$) as random effects distributing according to:

$$g \sim N(0, WW'\sigma^2_b)$$

Since the explained variance of GBLUP is the same of rrBLUP, these two models are equivalent and the only difference lies in the computational effort, which is inferior for GBLUP due to the lower number of parameters estimated (VanRaden, 2008).
A variant of GBLUP is the so called weighted GBLUP (WGBLUP) envisaging marker weight adjustment based on effect estimates from a preliminary GWAS, which can be advantageous compared with rrBLUP or GBLUP for traits featuring major QTLs (Wang et al., 2012).
Bayesian models can benefit from an enhanced flexibility compared to the above-mentioned linear mixed models, since marker effects can follow distributions other than normal. For instance, distributions featuring larger tails allow to put more weight on specific markers,

which can be advantageous for traits featuring major QTLs or when model prediction is performed on more advanced generations compared to training (Janss and Ramstein, 2023). Common Bayesian models are BayesA, Bayesian Lasso, and Power Lasso envisaging t, double exponential, and exponential power prior distribution of marker effects, respectively (Tibshirani, 1996; Meuwissen et al., 2001; Park and Casella, 2008). Variable selection models are a special class of Bayesian models that estimate for each marker the probability of being relevant or not, meaning that marker effects can follow a prior normal distribution with a large or small variance (George and McCulloch, 1993). The probability of each marker to be relevant or not is also considered as a model parameter characterized by its own prior distribution, which is constructed to allow only a few effects to be large, while the great majority is forced to be small (Janss and Ramstein, 2023). Common GS models belonging to this class are BayesB and BayesC, with the former featuring zero variance for the non-relevant markers and marker-specific variance for the relevant ones according to a t prior, while the latter envisages zero effect for non-relevant markers (Meuwissen et al., 2001; Habier et al., 2011).

1.4.3. Comparison with marker-assisted selection

MAS relies on gene mapping to identify the markers significantly associated to traits of interest to be used in the breeding process. Gene mapping can consist of either linkage or association mapping, where the former requires the creation of experimental populations usually from crossings between accessions with extreme phenotypes, while the latter relies on natural populations or germplasm collections. On one side, linkage mapping is more suitable to investigate the effect of rare alleles thanks to the balanced allelic frequencies characterizing artificial populations, but its results are less transferable to breeding material. On the other hand, association mapping ensures higher allelic diversity and mapping resolution due to several generations of recombination rearranging the original haplotypes. Besides the long times required by gene mapping and validation, MAS features low statistical power for the detection of small effect QTLs characterizing especially polygenic traits, which leads to the identification of a small fraction of the total QTLs (those featuring large effect) whose effect tends to be overestimated due to positive bias (Beavis effect). MAS limited statistical power largely depends on the need to adopt a stringent significance threshold to counteract the false positive increase due to the high number of markers that are usually tested, implying a reduction in the chance of finding true positive effects (multiple testing problem). Moreover, when several QTLs are found, pyramiding their alleles by MAS can

become time-consuming. For these reasons, MAS is normally successful for traits controlled by a small number of genes featuring a large effect, while it displays poor performance for polygenic traits. GS is much more effective than MAS when breeding for polygenic traits, since the inclusion of all the genotyped markers in the prediction model, without any previous selection step, allows to account for all the relevant effects if the marker density is sufficient (Meuwissen et al., 2001; Heffner et al., 2009).

1.4.4. State of the art for grain yield and protein content improvement in pea

Previous studies highlighted encouraging results for genomic prediction of pea grain yield or its components in different water availability conditions, cross-validation scenarios, including inter-population and inter-environment predictions, despite with some variability between environments and populations, and materials, such as breeding lines and germplasm collections (Burstin et al., 2015; Tayeh et al., 2015; Annicchiarico et al., 2017a, 2019, 2020; Al Bari et al., 2021). Moreover, a superior predicted efficiency of GS relative to PS for grain yield improvement was repeatedly reported in different water availability conditions and cross-validation scenarios (Annicchiarico et al., 2017a, 2019, 2020), although field assessments supporting this information are missing. A great influence of G × E interaction emerged for grain yield, especially across southern European environments, where it appeared more affected by year-to-year climatic variation than by geographical distance. This suggested the convenience of breeding for wide adaptation in sub-regions belonging to this area, such as Italy (Annicchiarico and Iannucci, 2008; Pecetti et al., 2019), which motivates our choice of considering environments of northern and central Italy as part of the same target region. To our knowledge, no information is currently available about GS model predictive ability or G × E interaction size relative to the genetic effects for grain protein content, with G × E that was reported as modest or non-significant in some studies (Matthews and Arthur, 1985; Krajewski et al., 2011) and significant in others (Burstin et al., 2007).

1.5. Research objectives

The general objective of the thesis work was the investigation of GS potential for the improvement of grain yield, protein content, and especially protein yield in pea, both per se and relative to PS in environments of northern and central Italy. The special focus on protein yield is motivated by its major importance for pea use as a protein source in animal feeding.

In this context, the main specific objectives of the work described in the following chapters are:

- Investigating the genetic architecture of grain yield and protein content
- Investigating the potential of GS for the prediction of grain yield, protein content, and protein yield in several scenarios differing for cross-validation configuration, with a particular interest for inter-environment and inter-population predictions
- Assessing the rate of genetic progress achieved for grain and protein yield by GS and PS

2. Pea grain protein content across Italian environments: genetic relationship with grain yield, and opportunities for genomic selection for protein yield

2.1. Objectives

This study is complementary to that conducted by Annicchiarico et al. (2019) for grain yield and is aimed at assessing (1) the extent of G × E for grain protein content, (2) grain yield and protein content genetic correlation and genetic architecture, (3) GS predictive ability for protein yield and its predicted efficiency relative to PS, both in intra-population and inter-population prediction scenarios.

2.2. Materials and methods

2.2.1. Plant material

This study was based on the same plant material and test environments described in Annicchiarico et al. (2019) for grain yield data. It included 306 genotypes belonging to three Recombinant Inbred Line populations issued by connected crosses between three cultivars, i.e., the European Attika and Isard and the Australian Kaspa, which featured high and stable grain yield across Italian environments in a previous cultivar assessment (Annicchiarico, 2005; Annicchiarico and Iannucci, 2008). Attika × Isard (A × I) included 102 lines, Kaspa × Attika (K × A) 100, and Kaspa × Isard (K × I) 104. The parental genotypes and the cultivar Spacial, which was used as a control because of its high yielding ability across Italian environments (Pecetti et al., 2019), were also included in field experiments, bringing the total tested genotypes to 310. DNA for genotyping was extracted from four F6 plants for each genotype grown in a non-heated glasshouse, while phenotyping was carried out on F7 plants.

2.2.2. Phenotyping

All field experiments were autumn-sown, rain-fed, and designed as a randomized complete block with three replicates and were identified hereafter by the combination of location and growing season as Lodi 2013–14, Lodi 2014–15, and Perugia 2013–14 (Picture 1). Lodi

(45°19′N, 9°30′E) is in northern Italy and is characterized by a subcontinental climate, whereas Perugia (43°06′N, 12°23′E) features a cool Mediterranean climate. An organic management was adopted in Lodi and Perugia 2013–14, while Lodi 2014–15 was managed conventionally. Additional details regarding experiment set up and grain yield assessment can be found in Annicchiarico et al. (2019). Grain protein content was determined on 100 g of dry seed per plot, which were previously ground by a cutting mill (Pulverisette 19, Fritsch GmbH, Germany) equipped with a 1 mm mesh sieve, by near-infrared spectroscopy (NIRS) with a Nirflex 500 spectrometer (Büchi, Italy) working in the 1,000–2,500 nm range. 348 plot samples were selected based on spectral information, of which 245 belonged to the current material set and 103 to the germplasm collection described in paragraph *3.2.1.*, according to a Kennard Stone multivariate design. The reference data were obtained by duplicate analysis of total nitrogen content by Dumas' method with a ThermoQuest NA1500 elemental analyser (Carlo Erba, Milano, Italy) and atropine as a standard. Partial Least Squares method within PLS Toolbox 8.9 (Eigenvector Research Inc.) was employed to develop prediction models after applying a pre-processing to NIRS spectra consisting in 2nd derivative computation and mean centring. In addition, a ten-fold venetian blind cross-validation was performed to determine the optimal number of principal components (PCs) to be included in the prediction models. Two models were developed, either envisaging external parameter orthogonalization in the pre-processing of NIRS spectra or not, and predictions were based on the mean between their results. The mean across models for cross-validation R^2 was equal to 0.78, and that for calibration R^2 to 0.93, with minor differences between models. Both models showed an estimated error of prediction of 0.12 g of nitrogen every 100 g of sample. Grain protein content was obtained by multiplying the NIRS-estimated nitrogen content by 6.25. Protein yield was computed on a plot basis by multiplying dry grain yield by grain protein content.

Picture 1. Field trial.

2.2.3. Statistical analysis of phenotypic data

The following analyses were performed for grain yield, protein content, and protein yield of the lines belonging to the RIL populations, unless otherwise specified, by SAS/STAT® or R Studio. In all the analyses, variance components were estimated by restricted maximum likelihood (REML) method. RIL populations were compared for mean trait value in each environment by Duncan's test based on the output of a mixed model with the RIL population as a fixed factor and the genotype within RIL population and the replicate as random factors. A mixed model with genotype and replicate as random factors was applied to data from each RIL population and environment to assess the significance of within-population variation and its extent as genetic coefficient of variation computed as $CV_g = S_G / m \times 100$, where S_G is the square root of genotype variance (S_G^2), and m is trait mean value. A mixed model including the environment as a fixed factor and the genotype, its interaction with the environment, and the replicate as random factors was employed to test the significance of the variance components relative to genotype and G × E. Another mixed model with the RIL population, the genotype, their interaction with the environment, and the replicate as random factors, and the environment as a fixed factor was employed to test the significance of the variance components relative to the first two random factors and their interaction with the environment. G × E was further investigated by computing the genetic correlation of line trait values for pairwise combinations of environments as $r_g = r / H_1H_2$, where r is the Pearson's correlation of line trait values for a given combination of environments, and H_1 and H_2 are the square root of the broad-sense heritability on a genotype mean basis in each environment, computed as $H^2 = S_G^2 / (S_G^2 + S_e^2 / n)$, with S_G^2 and S_e^2 representing variance components relative to genotype and error, and n the replicate number. Broad-sense heritability values were used to calculate BLUPs (DeLacy et al., 1996), which served as phenotypic data for GS and GWAS analyses. The genetic correlation between grain yield and protein content was estimated in each environment according to Piepho (2018) by using the freeware implementation in R proposed by Onofri (2019). The phenotypic correlation between protein yield and each of its components was estimated in each environment to assess the impact of grain yield and protein content on this trait. A mixed model including all genotypes (lines and parent/control cultivars), with genotype and environment as fixed factors and replicate as a random factor, was employed to assess the number of lines outyielding the control variety Spacial and the top-performing parent cultivar for all the target traits.

2.2.4. Genotyping and genomic data processing

Information about DNA isolation and GBS can be found in Annicchiarico et al. (2017a). Raw reads for library construction were demultiplexed using axe demultiplexer (Murray and Borevitz, 2018). Trimming for restriction enzyme remnants, alignment on reference genome version 1a (Kreplak et al., 2019), and SNP calling were performed according to the dDocent pipeline (Puritz et al., 2014). The final genotype matrix, in the form of a vcf file, was filtered for quality by vcftool (Danecek et al., 2011), with parameters -minQ 30, -max-non-ref-af 1, and -non-ref-af 0.001. The resulting dataset was filtered for increasing levels of maximum missing per marker (mpm) values, amounting to 5%, 10%, 15%, 20%, and 30%. Markers that were monomorphic or with minor allele frequency (MAF) lower than 5% were removed. Afterwards, filtering according to maximum missing per sample (mps) thresholds of 10%, 25%, and 50% was performed. Missing data imputation was performed according to the Random Forest method by R package MissForest (Stekhoven and Bühlmann, 2012), with the configuration ntree = 100, maxiter = 10, and genotypes defined as factors.

2.2.5. Genomic selection

The intra-population, inter-environment prediction scenario was assessed by a ten-fold stratified cross-validation scheme with ten repetitions, using two environments for training and one for validation according to all the possible combinations. Predictive ability was estimated separately for each RIL population to investigate within-population GS model performance. Results were averaged across repetitions, training environment sets, and RIL populations. Due to the satisfactory combination of computational performance and predictive ability that emerged from previous studies (Annicchiarico et al., 2019, 2020), rrBLUP (Meuwissen et al., 2001) was employed in this scenario to define the optimal mpm and mps thresholds to be employed for all the other GS analyses. Intra-population, inter-environment predictions for the optimal mpm and mps thresholds were assessed by four GS models, namely rrBLUP, BayesA (Meuwissen et al., 2001), BayesC (Habier et al., 2011), and Bayesian Lasso (Park and Casella, 2008). Because of its good predictive ability and computational efficiency, rrBLUP was selected for assessing the inter-population, inter-environment scenario, envisaging model training on data of one RIL population averaged across two environments and validation on data of each of the other two RIL populations in the remaining environment. All populations and pairs of environments were used by turn for

model training and the results were averaged across training sets. All the GS analyses were performed by the R package GROAN (Nazzicari and Biscarini, 2017).

2.2.6. Comparison of genomic vs. phenotypic selection

The correlation of each of mean phenotypic data across the two GS training environments and GEBV with phenotypic data in the validation environment was computed by averaging the results across all environment combinations and RIL populations, providing a comparison of PS vs. GS predicted performance for all the target traits. In addition, a comparison of GS vs. PS in terms of predicted efficiency, namely by accounting for differences in selection cycle duration and costs, was carried out for protein yield according to both the intra-population inter-environment, and the inter-population inter-environment scenarios. GS predictive accuracy was estimated according to Lorenz et al. (2011) as $r_{Ac} = r_{Ab} / H$, where r_{Ab} is the predictive ability averaged across RIL populations and training environments, and H the square root of the broad-sense heritability on a genotype mean basis in the validation environment. The expected genetic gain per GS cycle was computed according to Heffner et al. (2010) as $\Delta G_G = i_G\, r_{Ac}\, s_A$, where i_G is the standardized GS differential and s_A the standard deviation of phenotypic values. To get GS expected genetic gain per year, ΔG_G was divided by $t_G = 0.5$, namely the duration in years of one GS cycle under the hypothesis of two selection cycles per year. PS expected genetic gain per year was estimated according to Falconer (1989) as $\Delta G_{P'} = i_P\, H\, s_A / t_P$, where i_P is the standardized PS differential and t_P the duration in years of one selection cycle, which was hypothesized as equal to 1 in the case of two locations tested during the same year, or 2, if the same or different locations are tested in two different years. For the assessment of the inter-population inter-environment scenario, broad-sense heritability on a genotype mean basis was estimated for each RIL population and pairwise combination of selection environments as $H^2 = S_G^2 / (S_G^2 + S_{GE}^2 / e + S_e^2 / e\, n)$, with S_{GE}^2 representing G × E variance component, e the number of environments, and other terms corresponding to previous definitions. Consequently, comparing GS with PS in terms of predicted efficiency means accounting for the gap between i_G and i_P arising from the different evaluation cost per genotype, which was hypothesized as equal to € 220 for PS and € 60 for GS. This means that for a given budget GS would allow to evaluate 3.7 times the genotypes that could be assessed by PS. Since the ratio of i_G to i_P varies between 1.316 and 1.445, hypothesizing to select either the top 2.7% of accessions by GS and 10% by PS, or 5.5% by GS and 20% by PS (Falconer, 1989), an intermediate ratio of 1.381 was considered.

2.2.7. Genome-wide association study and linkage disequilibrium decay

GWAS was preferred to composite interval mapping (CIM), which is normally employed to identify marker-trait associations in experimental populations, because it allowed for a joint analysis of all the RILs. This ensured a much higher statistical power compared to that achievable by CIM, which would have relied on three separate within-population analyses, each based on about one third of the total number of individuals (simulations were run for QTLs controlling 1%, 5%, and 10% of phenotypic variance based on Wang and Xu, 2019). Population structure information to be included in the GWAS model was obtained by a Discriminant Analysis of Principal Components (DAPC; Yendle and Macfie, 1989) performed on genotype data pruned for excess of LD to avoid the strong influence of SNP clusters when estimating genetic relatedness (Laurie et al., 2010). Pruning was performed on SNPs of known genomic position by snp.pruning() function from R package ASRgenomics with a maximum r^2 threshold of 0.2, a window size of 50 SNPs, and an overlap of 5 SNPs between consecutive windows, generating a set of 5,094 SNPs. For DAPC, the k-means clustering algorithm was run iteratively for increasing values of K (i.e., numbers of genetic clusters) from 1 to 30, to identify its optimal value according to differences between successive values of the Bayesian information criterion. The analysis was performed on the output of a principal component analysis (PCA) to benefit from dimensionality reduction but keeping all the PCs to avoid information loss. The final DAPC was performed by using the optimal K value. The number of PCs to be retained for DAPC, and that of discriminant functions to be used as covariates in GWAS models, were determined by visual inspection of plots of PC cumulative variance and discriminant function eigenvalues, respectively. Based on this operation, 150 PCs were considered for DAPC, and 2 discriminant functions were employed as GWAS covariates. The procedure was implemented by using the functions find.clusters() and dapc() from R package adegenet (Jombart and Ahmed, 2011).

LD was estimated as r^2 value for pairwise combinations of SNPs within a 100 kb window by LD.decay() function from R package sommer (Covarrubias-Pazaran, 2016). The r^2 values were plotted against physical distance and fitted by a polynomial curve as described in Marroni et al. (2011). The 90th percentile of the r^2 distribution for pairwise combinations of SNPs located on different chromosomes was estimated by setting argument unlinked to true in LD.decay() function, to assess the most meaningful LD decay threshold for candidate gene research in our dataset.

A GWAS was performed for grain yield and protein content averaged across the three test environments by the Blink model (Huang et al., 2019) in R package GAPIT3 (Wang and Zhang, 2021). To investigate eventual differences in significant SNPs depending on the level of winter cold stress, GWAS was conducted also on grain yield data from each of Lodi 2013-14 and Lodi 2014-15. The examination of quantile-quantile (QQ) plots highlighted an appropriate compensation of population structure by the first two DAPC discriminant functions for all the datasets (Appendix, Figure 1). A Bonferroni threshold of 5% was employed to select significant SNPs.

2.3. Results

2.3.1. Phenotypic variation, genotype × environment interaction, and trait interrelationships

On average, Lodi 2013–14 featured higher grain yield, protein content, and protein yield compared to Lodi 2014–15 (Appendix, Table 1), in accordance with the more favourable conditions provided by a milder and wetter winter (Appendix, Table 2). Perugia 2013–14 showed intermediate grain protein content along with the lowest grain and protein yield (Appendix, Table 1), which were probably due to strong weed diffusion (Annicchiarico et al., 2019). The range of variation for trait values of the 306 RILs averaged across environments was 1.79–7.77 t/ha for grain yield, 21.7–26.6% for protein content, and 0.46–1.95 t/ha for protein yield. Several RILs outperformed the best parent cultivar for grain yield, protein content, and protein yield, with nine lines resulting significantly superior to the best parent cultivar (Isard) and six overcoming even the elite commercial variety Spacial for protein yield. Significant differences in RIL population means occurred for most traits and environments (Table 5) with grain and protein yield following a similar pattern, and RIL population × environment interaction resulted significant for all traits ($p < 0.01$; Appendix, Table 3). K × I tended to display the best trait value within single environments (Table 5), in accordance with the trend towards higher grain yield characterizing Kaspa and Isard, and with the higher protein content of Kaspa relative to the other parental lines (Appendix, Table 4). Significant variation was found for all traits and RIL populations in each environment according to CV_g, which were much smaller for protein content compared with both grain

and protein yield, with the latter two traits displaying similar values (Table 5). The greater CV_g observed for grain and protein yield of all RIL populations in Lodi 2014–15 were caused by a definitely colder winter compared with the other environments (Appendix, Table 2), leading to large variation between genotypes for winter survival. Variance component assessment in the whole dataset revealed over two-fold larger genetic relative to G × E variance for protein content, while opposite results were found for grain and protein yield (Appendix, Table 3). For all traits, within-population genetic variation resulted much larger than between-population one, whereas G × E was somewhat more affected by RIL population × environment than by genotype within population × environment component (Appendix, Table 3). Line genetic correlation for grain and protein yield between pairs of environments was much lower between different years in Lodi than between different locations during 2013–14 (Table 6), thereby confirming the greater extent of genotype × year compared with genotype × location variance component in the Italian target region. Although statistically significant, G × E for protein content did not imply marked inconsistency of genotype responses between environments, as revealed by the high genetic correlation values ($r_g \geq 0.73$; Table 6). Together with the greater size of genotype relative to G × E variance component, a much larger within-trial broad-sense heritability was detected on average for protein content ($H^2 = 0.82$) compared with grain and protein yield ($H^2 = 0.52$ and $H^2 = 0.54$, respectively). Grain yield and protein content exhibited a slightly positive genetic correlation in all environments, which resulted significant at $p < 0.05$ only in Lodi 2014–15 (Table 6). Protein yield was much more affected by grain yield than by protein content, as revealed by phenotypic correlation results (Table 6).

Table 5. Mean value and genetic coefficient of variation (CV_g) of three traits measured in three test environments on pea lines from three connected RIL populations (102 lines from A × I, 100 from K × A, and 104 from K × I). Row means followed by different letters differ at $p < 0.05$ according to Duncan's test. CV_g (computed as S_G / m, where S_G = square-root of genotype variance and m = trait mean value) resulted always significantly different from zero at $p < 0.01$.

Trait	Environment	Mean value				CV_g (%)		
		A x I	K x A	K x I	SE	A x I	K x A	K x I
Grain yield (t/ha)	Lodi 2013-2014	5.99 a	6.33 a	6.54 a	0.14	10.1	17.5	18.2
	Lodi 2014-2015	5.80 a	2.52 b	5.78 a	0.18	28.0	51.3	33.0
	Perugia 2013-2014	2.61 b	2.77 b	3.31 a	0.08	24.8	20.7	14.8
Protein content (%)	Lodi 2013-2014	24.72 b	25.55 a	25.69 a	0.1	3.7	3.9	3.3
	Lodi 2014-2015	23.23 a,b	23.03 b	23.37 a	0.1	3.9	3.6	3.9
	Perugia 2013-2014	23.29 b	24.82 a	24.68 a	0.11	3.9	4.5	3.4
Protein yield (t/ha)	Lodi 2013-2014	1.48 b	1.62 a	1.68 a	0.03	11.1	18.0	18.5
	Lodi 2014-2015	1.34 a	0.58 b	1.35 a	0.04	30.6	53.5	34.0
	Perugia 2013-2014	0.61 c	0.69 b	0.82 a	0.02	25.6	21.8	14.3

Table 6. **A)** Genetic correlation of line values across pairs of test environments featuring the same location (same loc.) or year for three pea traits. Genetic correlation was always significantly different from zero at $p < 0.01$. **B)** For each environment, genetic correlation between grain yield and protein content (r_g) with the relative standard error (SE), and phenotypic correlation between protein yield and its component traits (grain yield and protein content) were displayed. Genetic correlation was significantly different from zero only in Lodi 2014-15 ($p < 0.05$), while phenotypic correlation in all environments ($p < 0.01$). All the results refer to 306 pea lines from three connected RIL populations.

A) Trait	Same loc.	Same year	B) Environment	r_g	SE	Grain yield	Protein content
Yield (t/ha)	0.35	0.79	Lodi 2013-14	0.12	0.08	0.98	0.30
Protein content (%)	0.73	0.92	Lodi 2014-15	0.18	0.07	1.00	0.25
Protein yield (t/ha)	0.34	0.80	Perugia 2013-14	0.14	0.08	0.99	0.29

2.3.2. Genomic selection

GBS produced, on average, 551,210 reads per sample. The number of polymorphic SNPs was severely affected by the mpm and mps thresholds applied at filtering. Mpm below 5% always implied very few polymorphic SNPs (< 500), so models with mpm and mps in the range 5–30% and 10-50%, respectively, were tested, producing from 2,297 to 30,464 polymorphic SNPs. Slight predictive ability differences were found for these combinations

of mpm and mps for all traits, with a trend towards lower protein content predictive ability for the threshold combination of mpm = 0.3 and mps = 0.1 in all environments. Thresholds of mpm = 0.2 and mps = 0.25 were adopted for the following GS analyses since they ensured a good compromise between model predictive ability and number of markers and genotypes retained in the dataset with 18,674 polymorphic SNPs and 276 RILs. The four GS models performed very similarly in the intra-population inter-environment scenario for all traits, with a very slight advantage for grain and protein yield of rrBLUP, which therefore was employed for subsequent analyses. The mean predictive ability in this scenario resulted high for protein content (r_{Ab} = 0.53), and moderately high for grain and protein yield (r_{Ab} = 0.40 and r_{Ab} = 0.41, respectively; Table 7). Intra-population inter-environment predictions did not differ markedly for the single validation environments (Table 7), although a somewhat lower predictive ability was achieved for grain and protein yield in Perugia 2013-14, and for protein content in Lodi 2014–15 (Table 7). The inter-population inter-environment scenario implied a predictive ability decrease of about 50% for all traits compared to the intra-population inter-environment scenario, with model training on A × I leading to distinctly inferior predictions for grain and protein yield (Table 7). The mean inter-population predictive ability for protein content was not only higher, but also less affected by the specific RIL population used for GS model training compared with the other two traits (Table 7).

Table 7. Predictive ability range for the four models (rrBLUP, BayesA, BayesC, Bayesian Lasso) tested in the intra-population inter-environment scenario and predictive ability values for each validation environment (Lo = Lodi, Pg = Perugia) in the same scenario, or training population in the inter-population inter-environment scenario. All the predictive ability values were obtained by rrBLUP, using a ten-fold stratified cross-validation scheme with ten repetitions in the intra-population scenario. The results were averaged across validation environments in both scenarios, and populations in the inter-population scenario. The genetic base consisted of 276 genotypes from three connected RIL populations.

	Intra-population					**Inter-population**			
Trait	**Range**	**Lo 2013-14**	**Lo 2014-15**	**Pg 2013-14**	**Mean**	**A × I**	**K × A**	**K × I**	**Mean**
Grain yield (t/ha)	0.39-0.40	0.39	0.45	0.36	0.40	0.08	0.28	0.27	0.21
Protein content (%)	0.52-0.53	0.60	0.45	0.53	0.53	0.27	0.21	0.32	0.27
Protein yield (t/ha)	0.40-0.41	0.40	0.46	0.36	0.41	0.08	0.25	0.27	0.20

2.3.3. Comparison of genomic vs. phenotypic selection

Based on correlation results, the ability of PS and GS relying on two selection or training environments to predict RIL phenotypic data in a third environment resulted similar, with a

modest advantage of PS for protein content (0.75 vs. 0.70), and of GS for grain and protein yield (0.48 vs. 0.46, and 0.51 vs. 0.49, respectively). GS predicted efficiency relative to PS was largely dependent on the assumed GS scenario and duration of one PS cycle, resulting over four-fold larger for GS intra-population prediction and a two-year PS cycle, equal for GS inter-population scenario and a one-year PS cycle, and over two-fold larger in the other cases (Table 8).

Table 8. Ratio of genomic selection (GS) to phenotypic selection (PS) efficiency for protein yield based on predicted genetic gains per unit time for similar evaluation costs assuming two environments for PS and for GS model training in intra-population (GS_A) or inter-population (GS_B) inter-environment scenarios. H_C is the square root of the broad-sense heritability on a genotype mean basis; r_{Ac} is GS predictive accuracy; t_P is the duration in years of one cycle of PS. Efficiency ratios were averaged across validation environments and RIL populations.

		GS_A/PS efficiency ratio			**GS_B/PS efficiency ratio**		
Trait	H_C	$GS_A r_{Ac}$	$t_P = 1$	$t_P = 2$	$GS_B r_{Ac}$	$t_P = 1$	$t_P = 2$
Protein yield (t/ha)	0.676	0.511	2.192	4.383	0.252	1.084	2.167

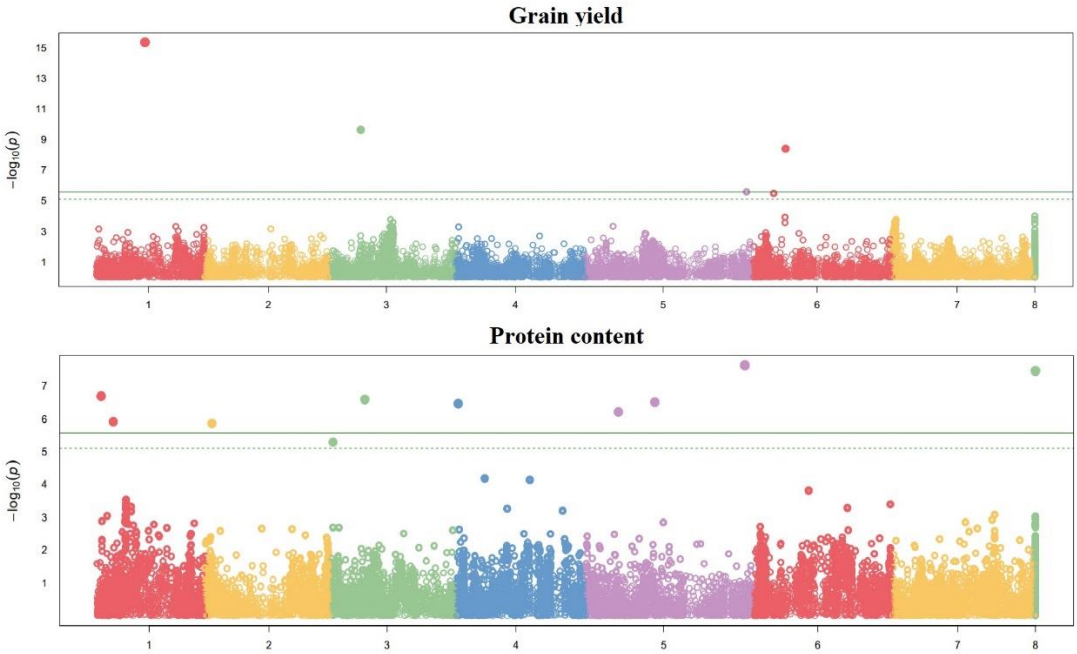
2.3.4. Genome-wide association study and linkage disequilibrium decay

Genomic data employed for GWAS and LD analysis were obtained by filtering the original dataset according to mpm < 0.2, mps < 0.25, MAF > 0.05, and SNP heterozygosity < 0.3, retaining 18,674 polymorphic SNPs and 276 RILs. On average, LD reached $r^2 = 0.2$ at 99,884 bp, with single chromosome values ranging from 99,685 bp for chromosome 2 to 99,997 bp for chromosome 1 (Appendix, Figure 2). The 90th percentile of the r^2 distribution for pairwise combinations of SNPs located on different chromosomes resulted equal to 0.08 and was reached at 99,885 bp on average. Based on these results, a 100 kb genomic region was scanned in both directions from each significant SNP to look for candidate genes.

The DAPC was performed by adopting K = 3, since it resulted as the optimal group number in all the iteration rounds, in accordance with the presence of three RIL populations. Grain yield and protein content averaged across the three environments displayed several significant associations spread across the genome, confirming their expected polygenic control. Five significant SNPs mapping on chromosomes 1, 3, 5 and 6 were found for grain yield, and ten significant SNPs were identified for protein content, of which nine mapped on chromosomes 1, 2, 3, 4 and 5, and one on scaffolds (Figure 15). For grain yield, four significant SNPs were detected on chromosomes 2, 3, and 6 in Lodi 2013-14, and eight on chromosomes 1, 2, 3, 5, and 6 in Lodi 2014-15, of which two SNPs mapping on chromosomes 1 and 5 resulted significant also for trait mean across environments. In addition, significant

SNPs were detected for grain yield in corresponding genomic regions by GWAS conducted on different datasets, namely two SNPs on chromosome 3 for trait mean across environments and Lodi 2014-15, and three on chromosome 6, one for each dataset. The list of significant SNPs detected for grain yield and protein content for each dataset was provided in Appendix Table 5 along with the estimated effect, while a list of the candidate genes relative to the analysis performed on mean trait values across environments was reported in Appendix Table 6.

Figure 15. Manhattan plots showing the association scores of 18,674 SNPs with two traits averaged across three test environments for a GWAS based on Blink model and performed on 276 lines belonging to three connected RIL populations. The continuous and dashed lines represent Bonferroni threshold at 1% and 5%, respectively.



2.4. Discussion

This study showed that the improvement of pea protein content is less challenging compared with that of grain yield, independently from the selection method (PS or GS). The reason is a lower influence of G × E for protein content, which simplifies both PS and GS by reducing the number of experiments needed to get reliable data for selection or model training. Moreover, the absence of a strong genetic correlation between grain yield and protein content is encouraging for their simultaneous improvement, as confirmed by the presence of different genomic regions controlling these traits in the GWAS. This result is in accordance with

phenotypic correlation values reported by some earlier studies (Cousin et al., 1985; Klein et al., 2020; Bărbieru, 2021), but not by others (Tar'an et al., 2004; Krajewski et al., 2011). The 5% phenotypic variation detected for grain protein content falls in the range of what reported for breeding material (Cousin et al., 1985; Tar'an et al., 2004; Burstin et al., 2007; Jha et al., 2015), while being lower than that found by Ferrari et al. (2016) for the same genetic base, probably due to single-environment evaluation in the latter study. The much higher genetic variation found within compared to between RIL populations for all the target traits, according to variance component estimates, emphasized the importance of within-population selection. Protein yield was predominantly affected by grain yield and far less by protein content according to phenotypic correlation results, as confirmed by the very similar genetic variation, variance components, and genomic prediction results obtained for the first two traits. The greater size of genotype × year compared with genotype × location interaction that emerged for all traits, likely due to variation in winter cold stress severity (Annicchiarico et al., 2019) agreed with grain yield results based on larger environment samples (Annicchiarico and Iannucci, 2008; Pecetti et al., 2019), supporting selection for wide adaptation in the target region.

The slight differences in predictive ability found between statistical models for all traits in the intra-population inter-environment scenario were in accordance with the results reported in earlier pea studies for genomic prediction of grain yield or other traits (Burstin et al., 2015; Annicchiarico et al., 2017a, 2019). Major findings of this study were the high (r_{Ab} = 0.53) and moderately high (r_{Ab} = 0.41) GS predictive ability values in the intra-population inter-environment scenario obtained for protein content and protein yield, respectively, with a limited influence of the specific training environments. The higher predictive ability found for grain yield relative to Annicchiarico et al. (2019) was likely due to the greater number of GS training environments employed by the current study (two vs. one), without excluding the effect of SNP calling based on pea sequenced genome instead of mock genome. The greater GS predictive ability detected for protein content compared with grain or protein yield is likely due to its higher within-trial broad-sense heritability and lower influence of G × E. The decrease of GS predictive ability passing from the intra-population to the inter-population scenario approached 50% for all traits, but its value varied remarkably for grain and protein yield depending on the RIL population used for model training. Indeed, the use of A × I as a training set implied a lower predictive ability for these traits compared with the other two populations, in accordance with what reported for the same materials for grain yield under severe drought and onset of flowering (Annicchiarico et al., 2017a). This may be due

to a lower number of polymorphic markers characterizing A × I, whose parents are both of European origin and showed the highest genetic similarity according to Nei's (1972) genetic distance compared with the other pairs of parental lines (Annicchiarico et al., 2019).

The observed LD decay was much faster than that reported by Alemu et al. (2022) for a collection of 188 vining pea varieties and breeding lines provided by a single company (where $r^2 = 0.2$ was reached at 6,930,000 bp on average vs. about 100,000 bp in our study), while being considerably slower compared to the findings of Pavan et al. (2022) and Crosta et al. (2023) for two worldwide germplasm collections ($r^2 = 0.2$ was reached at 30 bp and 1,445 bp on average, respectively). These results are in line with the expectations, considering that the first genetic base was likely narrower, while the other two considerably wider and resulting from a much higher number of meiotic events relative to the current population. However, the only study using our same LD decay fitting method was that by Crosta et al. (2023), while different methods were employed in the other works, with a possible effect on the results. GWAS results confirmed the polygenic control of grain yield and protein content by highlighting many significant markers spread across the genome for trait mean data across environments, thereby supporting the interest of developing GS models for both traits and their combination. QTLs for grain yield were detected in the same genomic regions of our significant SNPs on chromosome 1 (Gali et al., 2019), 5 (Klein et al., 2014; Gali et al., 2018), and 6 (Klein et al., 2020; Crosta et al., 2023), and for protein content on chromosome 2 (Gali et al., 2018), 3 (Gali et al., 2019), 4 (Klein et al., 2014), and 5 (Klein et al., 2014; Gali et al., 2018; Gali et al., 2019; Klein et al., 2020). Many candidate genes of possible interest emerged for both grain yield and protein content. For instance, for the first trait, Psat1g096760 encodes a phosphatidylethanolamine-binding protein that can be involved in flowering control in response to the environmental conditions (Książkiewicz et al., 2016), while Psat3g051840, Psat3g051880, and Psat5g289760 code for transcription factors whose families (RING for the first two and BZIP for the last one) play a role both in plant growth and abiotic stress response (Dröge-Laser et al., 2018; Han et al., 2022). Moreover, Psat5g289640 encodes an electron transfer flavoprotein that regulates the flux to the mitochondrial transport chain under carbohydrate-limiting conditions (Brito et al., 2022). For protein content, Psat5g132320 may participate in plant symbiosis with *Rhizobia*, since it encodes a lysin motif domain that is known to play a key role in plant-microbe interaction (Gust et al., 2012), while Psat2g022320 codes for an ethylene insensitive 3 protein, which is involved in leaf senescence and nitrogen metabolism in wheat (Sultana et al., 2021). The predominant influence on grain yield mean data across experiments of Lodi 2014-15, namely the

environment featuring the strongest winter cold stress, resulted evident from the fact that all the significant SNPs detected for trait mean, except for one on chromosome 6, were either significant or close to significant SNPs in this environment. Moreover, our hypothesis of a different genetic control of grain yield depending on the intensity of winter cold stress was confirmed by the differences in significant SNPs and relative genomic regions found between Lodi 2013-14 and Lodi 2014-15.

This study provided an unprecedented comparison of GS vs. PS in terms of predicted efficiency for protein yield improvement in pea. Its results indicated an advantage of intra-population GS over all PS scenarios, and of inter-population GS on PS relying on two-year data, which represents the most realistic scenario due to the sizeable genotype × year interaction detected for grain yield. A crucial confirmation of the advantage of GS over PS for pea protein yield improvement will be provided by future research work comparing these selection methods in terms of actual genetic gains.

Note: the work presented in this chapter was mentioned in the other sections by referring to the relative publication (Crosta et al., 2022).

3. Genomic prediction and allele mining for the improvement of grain yield and protein content in a pea germplasm collection

3.1. Objectives

The objectives of the current work are: (1) testing the ability of GS models developed on a germplasm collection to predict grain yield and protein content both in germplasm accessions and breeding material, as represented by three connected RIL populations that were evaluated in three independent Italian environments in earlier studies (Annicchiarico et al., 2019; Crosta et al., 2022); (2) performing a GWAS for seed protein content and grain yield under severe terminal drought to identify eventual QTLs for these traits; (3) investigating the genetic relationship between grain yield and protein content and the phenotypic correlation between protein yield and each of its component traits in a highly diversified genetic base.

3.2. Materials and methods

3.2.1. Plant material and phenotyping

The study was based on 220 cultivated pea landraces and old cultivars belonging to 19 regional germplasm pools and 11 modern cultivars bred in France (Attika, Cartuce, Dove, Enduro, Genial, Isard, Messire, Spirale), Spain (Cigarron, Viriato) or Germany (Santana) (Appendix Table 7). This collection was set up by pooling selected accessions that were provided by IPK (Gatersleben), INRAE UMRLEG (Dijon), John Innes Centre (Norwich), CNR-IGV (Bari) and ICARDA's gene bank. These institutions were asked to provide accessions which, according to the available knowledge, were able to maximize the genetic diversity within the gene pool of each country that was addressed by our request. A previous study (Pavan et al., 2022) confirmed the wide genetic variation and the absence of duplicates among the accessions included in this collection. This material was evaluated by Annicchiarico et al. (2017b) in Lodi, northern Italy (45°19'N, 9°03'E), in a spring-sown rain-fed field experiment designed as a randomized complete block with two replicates. This experiment was characterized by substantial terminal drought, as provided by a rainfall amount of 178 mm over the crop cycle. Grain yield and protein content were determined on a plot basis and NIRS method was employed for protein content measurement based on the

same calibration and models described in paragraph *2.2.2*. Further details about the experiment can be found in Annicchiarico et al. (2017b). The three RIL populations employed for GS model validation were issued by connected crosses between three parent cultivars (Attika and Isard, of European origin, and Kaspa, bred in Australia) that featured high and stable grain yield across Italian environments in earlier variety testing (Annicchiarico, 2005; Annicchiarico and Iannucci, 2008). This set included 306 lines that were evaluated for grain yield by Annicchiarico et al. (2019), and for protein content by Crosta et al. (2022) in three environments of northern or central Italy. These environments differed from that of germplasm collection evaluation in various respects: they were autumn-sown, which implied substantial winter cold stress (particularly in one environment), more favourable in terms of water availability (featuring at least 500 mm rainfall over the crop cycle) and managed organically. Further details about these experiments can be found in the relative studies (Annicchiarico et al., 2019; Crosta et al., 2022) or in Chapter 2.

3.2.2. Statistical analysis of phenotypic data and trait interrelationships

Broad-sense heritability was estimated by the formula presented in paragraph *2.2.3*. A linear mixed model was applied on ecotypes with germplasm pool and accession within pool as random factors, to assess the significance and relative size of between and within pool variance components estimated by REML method (Annicchiarico et al., 2017b). Phenotypic correlation between protein yield and each of its two components was estimated, and the genetic correlation between grain yield and protein content was computed according to Piepho (2018) by using the freeware implementation in R proposed by Onofri (2019).

3.2.3. Genotyping and genomic data processing

For DNA extraction, one plant per accession was selected that was morphologically representative based on visual evaluation. Information on DNA isolation and GBS can be found in Pavan et al. (2022) for the 231 accessions of the germplasm collection, and in Annicchiarico et al. (2019) for lines belonging to the three connected RIL populations. GBS was performed according to the protocol of Elshire et al. (2011) modified by using the ApeKI restriction enzyme and KAPA Taq polymerase. The raw reads of accessions from the germplasm collection were pre-processed by Trimmomatic Version 0.39 (Bolger et al., 2014), aligned against pea reference genome v1a (Kreplak et al., 2019) by Burrows-Wheeler

Aligner (Li and Durbin, 2009), and subjected to quality control and SNP calling according to the dDocent pipeline (Puritz et al., 2014). Monomorphic markers were eliminated from the resulting dataset, which was filtered by MAF > 5%, mpm < 20%, SNP heterozygosity rate < 30%, and mps < 25%, retaining a total of 223 accessions and 41,114 SNPs. Information about demultiplexing, alignment to the reference genome, and quality filtering can be found in paragraph *2.2.4.* for the three connected RIL populations used for validation. RIL genomic data were merged with molecular data from the germplasm collection retaining only polymorphic SNPs. The resulting dataset was filtered by MAF > 5%, mpm < 20%, SNP heterozygosity rate < 30%, and mps < 25%, retaining a total of 276 RILs and 4,929 SNPs in common with the germplasm collection. Missing data were estimated by the k nearest neighbour imputation method (Andridge and Little, 2010).

3.2.4. Genomic selection

Three GS models were tested in two scenarios for each of grain yield and protein content, namely rrBLUP (Meuwissen et al., 2001), BayesC (Habier et al., 2011), and Bayesian Lasso (Park and Casella, 2008), by the R package GROAN (Nazzicari and Biscarini, 2017). The first scenario was based on 41,114 SNPs and consisted in a ten-fold non-stratified cross-validation performed on germplasm collection data with fifty repetitions for rrBLUP and ten for Bayesian models, whose results were averaged to get a unique predictive ability value. The second scenario envisaged an inter-population inter-environment validation of GS models developed on the germplasm collection on both separated and pooled data of the three connected RIL populations. In this case, GS models included only 4,929 SNPs and phenotypic data of the validation set were averaged across the three evaluation environments. Filtering retained 77 lines and 4,784 polymorphic SNPs for population A × I, 96 lines and 4,846 polymorphic SNPs for K × A, and 103 lines and 4,922 polymorphic SNPs for K × I. Moreover, the number of polymorphic SNPs among those featuring the highest 100, 300, or 1,000 effects in absolute value according to GS models for grain yield and protein content, was computed for each RIL population in the validation set to investigate its relationship with the within-population predictive ability. The choice of considering a maximum of 1,000 SNPs was due to most plants or livestock breeding simulations assuming 1,000 or less QTLs for polygenic traits (Brito et al. 2011; Yin et al. 2014; Wientjes et al. 2015; Yao et al. 2018; Strandén et al. 2019; Peters et al. 2020).

3.2.5. Genome-wide association study and linkage disequilibrium decay

Population structure information to be included in the GWAS models was obtained by a DAPC (Yendle and MacFie, 1989) performed on genotype data pruned for excess of LD to avoid the strong influence of SNP clusters when estimating genetic relatedness (Laurie et al., 2010). The snp.pruning() function from R package ASRgenomics was employed on SNPs of known genomic position with a maximum r^2 threshold of 0.2, a window size of 50 SNPs, and an overlap of 5 SNPs between consecutive windows, generating a set of 11,072 SNPs. The k-means clustering algorithm was run iteratively for increasing values of K (i.e., cluster number) from 1 to 30, to identify its optimal value according to differences between successive values of the Bayesian information criterion. The analysis was performed on the output of a PCA to benefit from dimensionality reduction but keeping all the PCs to avoid information loss. The final DAPC was performed by using the optimal K value. The number of PCs to be retained for DAPC, and that of discriminant functions to be used as covariates in the GWAS models, were determined by visual inspection of plots of PC cumulative variance and discriminant function eigenvalues, respectively. Based on this operation, 150 PCs were considered for DAPC, and 8 discriminant functions were employed as GWAS covariates. The whole procedure was implemented by using the functions find.clusters() and dapc() from R package adegenet (Jombart and Ahmed, 2011). LD was estimated as r^2 value for pairwise combinations of SNPs within a 100 kb window by LD.decay() function from R package sommer (Covarrubias-Pazaran, 2016). The r^2 values were plotted against physical distance and fitted by a polynomial curve as described in Marroni et al. (2011). The 90th percentile of the r^2 distribution for pairwise combinations of SNPs located on different chromosomes was estimated by setting argument unlinked to true in LD.decay() function, to assess the most meaningful LD decay threshold for candidate gene research in our dataset. A GWAS was performed on 41,114 polymorphic SNPs according to the Blink model (Huang et al., 2019) by the R package GAPIT3 (Wang and Zhang, 2021). Visual examination of QQ plots highlighted an appropriate compensation of population structure by DAPC discriminant functions (Appendix Figure 3). A Bonferroni threshold of 5% was employed to select significant SNPs.

3.3. Results

3.3.1. Phenotypic variation and trait interrelationships

Phenotypic variation within all the regional pools resulted significant at $p < 0.01$ for both grain yield and protein content in Annicchiarico et al. (2017b), to which we refer for further details about the variability within and between pools. The high impact of terminal drought was confirmed by a low mean grain yield (about 1.1 t/ha), with modern cultivars displaying lower grain yield, despite an earlier flowering, and a similar protein content compared with the traditional germplasm (Table 9). The range of phenotypic variation for both target traits was remarkably larger for the landraces and old cultivars compared with the improved varieties (Table 9). Grain yield and protein content featured modest and moderately high broad-sense heritability, respectively (Table 9), while no significant genetic correlation emerged between these two traits. Finally, phenotypic correlation results highlighted a predominant influence of grain yield compared with protein content on protein yield determination ($r = 0.99$ with $p < 0.001$ vs. $r = 0.15$ with $p = 0.03$, respectively).

Table 9. Mean, variation range, and broad-sense heritability estimated on a genotype mean basis for three pea traits measured on a worldwide germplasm collection of 220 landraces from 19 regional pools and 11 improved varieties.

	Landraces		Improved varieties		
Trait	**Mean**	**Range**	**Mean**	**Range**	H^2
Grain yield (t/ha)	1.11	0.16-3.30	0.85	0.31-2.03	0.47
Protein content (%)	22.8	17.5-27.28	22.8	20.4-23.2	0.68
Onset of flowering (days from 1/1)	133	113-154	131	127-136	0.87

3.3.2. Genomic selection

The GS models trained and validated on the germplasm collection displayed moderately high predictive ability for both grain yield and protein content ($r_{Ab} = 0.435$ and 0.549 for the best model, respectively), mostly with slight differences between statistical models (Table 10). As expected, the inter-population inter-environment GS scenario produced a substantial decrease in predictive ability, whose extent varied largely with RIL population. For both traits, K × I showed the highest predictive ability, which for grain yield was almost equal to that of the intra-population scenario, while for protein content resulted about one third of it. On the other hand, K × A and A × I displayed intermediate and zero predictive ability for grain yield, while both showed null predictive ability for protein content (Table 11). When pooled lines of the three RIL populations were used as the validation set, predictive ability dropped by about

50% compared with the intra-population scenario for protein content (0.281 vs. 0.549 considering top-performing models), while it reached zero for grain yield (Table 11). The number of top-effect polymorphic SNPs for protein content was very similar in the three RIL populations, independently from the number of SNPs considered, although K × I and A × I tended to display the highest and lowest number, respectively, in the 300 and 1,000 SNP scenarios (Figure 16). The same trend characterized grain yield data, with the difference between A × I and the other two populations becoming quite evident in the 1,000 SNP scenario (Figure 16).

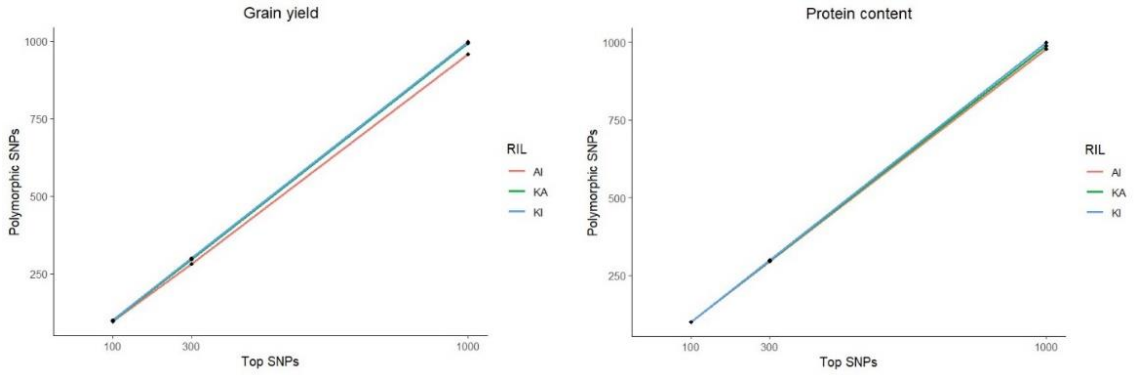
Table 10. GS predictive ability in an intra-population intra-environment scenario based on a ten-fold non-stratified cross-validation for two quantitative traits and three models relying on 41,114 SNPs and 223 accessions represented by 212 landraces from 19 regional pools and 11 modern cultivars from a worldwide pea germplasm collection.

Trait	rrBLUP	BayesLasso	BayesC
Grain yield	0.435	0.431	0.426
Protein content	0.549	0.540	0.539

Table 11. Phenotypic variation range in the validation set for two pea traits and predictive ability values based on three GS models trained on 212 landraces from 19 regional pools and 11 modern cultivars from a worldwide germplasm collection characterized in a single environment and validated on 276 RILs from three connected populations characterized in three environments.

			Predictive ability		
Trait	**Validation set**	**Range**	**rrBLUP**	**BayesLasso**	**BayesC**
Grain yield (t/ha)	RILs A × I	2.79 - 6.79	−0.236	-0.237	-0.246
Grain yield (t/ha)	RILs K × A	2.09 - 6.05	0.270	0.258	0.264
Grain yield (t/ha)	RILs K × I	3.08 - 7.60	0.446	0.439	0.443
Grain yield (t/ha)	All RILs	2.79 - 7.60	−0.025	-0.038	-0.022
Protein content (%)	RILs A × I	21. 7 - 25.8	-0.225	-0.240	-0.190
Protein content (%)	RILs K × A	22.0 - 26.6	0.028	0.024	-0.013
Protein content (%)	RILs K × I	22.5 - 26.4	0.184	0.185	0.157
Protein content (%)	All RILs	21.7 - 26.7	0.281	0.263	0.255

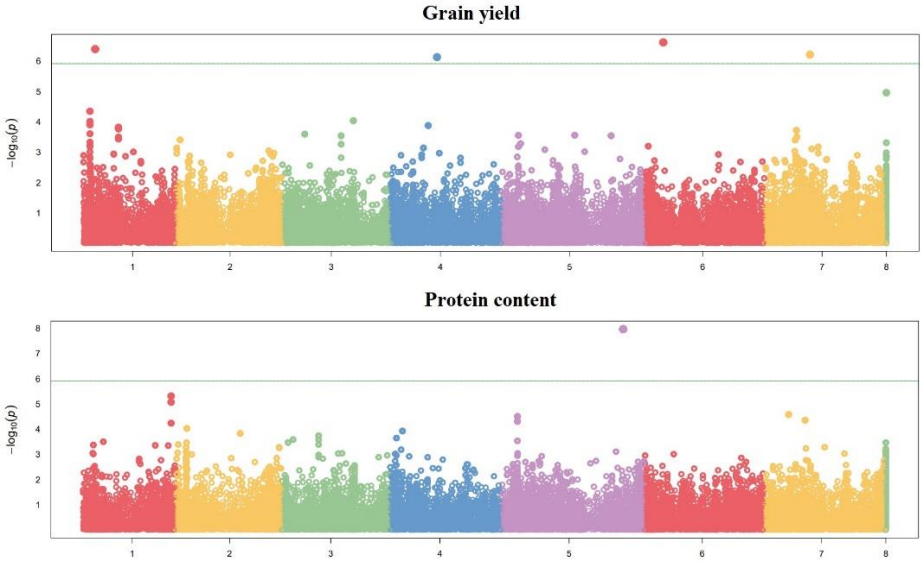
Figure 16. Number of polymorphic SNPs among the 100, 300, or 1,000 featuring the highest effect in absolute value according to GS models for pea grain yield and protein content displayed separately for the three RIL populations in the validation set (A × I, K × A, K × I), which was formed by 276 lines. GS model training was performed on 212 landraces from 19 regional pools and 11 modern cultivars from a worldwide germplasm collection.



3.3.3. Genome-wide association study and linkage disequilibrium decay

On average, LD reached half of its 90th percentile (r^2 = 0.38) at 217 bp, with single chromosome values ranging from 146 bp for chromosome 2 to 326 bp for chromosome 4 (Appendix Figure 4). The 90th percentile of the r^2 distribution for pairwise combinations of SNPs located on different chromosomes resulted equal to 0.05 and was reached at 10,140 bp on average (Appendix Figure 4). The mean distance at which r^2 dropped to 0.05 on a specific chromosome was scanned in both directions from each significant SNP on that chromosome to look for candidate genes. The DAPC was performed by adopting K = 16 as the optimal cluster number. The list of significant SNPs detected for grain yield and protein content is provided in Appendix Table 8 along with additional information about their estimated effect, while a list of the relative candidate genes is reported in Appendix Table 9. Both protein content and grain yield under severe terminal drought displayed significant associations. Four significant SNPs mapping on chromosomes 1, 4, 6 and 7 were found for grain yield, and one significant SNP was detected for protein content on chromosome 5 (Figure 17).

Figure 17. Manhattan plots showing the association scores of 41,114 SNPs with two traits along pea chromosomes for a GWAS based on Blink model and performed on 212 landraces from 19 regional pools and 11 modern cultivars from a worldwide germplasm collection. The green line represents Bonferroni threshold at 5%.



3.4. Discussion

The predominant role of grain yield compared with protein content in protein yield determination, and the absence of significant genetic correlation between grain yield and protein content reported by Crosta et al. (2022) for breeding material were confirmed in our diversified germplasm collection in the presence of substantial terminal drought. This emphasized the importance of improving grain yield to enhance protein yield, while encouraging simultaneous breeding for higher grain yield and protein content.

Intra-population, intra-environment predictive ability values of 0.43 for grain yield, which was nearly identical to that reported for the USDA pea collection (Al Bari et al., 2021), and of 0.55 for protein content support the use of GS models to select superior genotypes for these traits in germplasm collections. The application of GS models trained on the germplasm collection to predict breeding values of lines from the three RIL populations was challenged by the much lower genetic diversity of the validation set compared with the training set, the limited SNP number (4,929), and the large differences in evaluation environments for sowing time (spring vs. autumn) and drought stress extent (severe vs. limited). In this context, the null predictive ability that emerged for grain yield, when using the pooled lines of the three

RIL populations for validation, was not surprising. The same applies to protein content, for which the 50% predictive ability loss compared with the intra-population scenario was even lower than expected, considering that a comparable decrease was observed for inter-population inter-environment predictions relative to RIL populations having one parent in common and evaluated in similar test environments (Crosta et al., 2022). As regards the variation in predictive ability between RIL populations, the fact that the extent of phenotypic variation within populations was similar for both traits (Table 11), suggests that other factors may account for these differences. For both grain yield and protein content, K × I, K × A and A × I featured decreasing predictive ability values, in accordance with the declining number of polymorphic markers found for these populations, either considering the whole SNP set or the SNPs with the highest 300 or 1,000 effects, with K × I showing a surprisingly high predictive ability for grain yield (0.446). The large predictive ability differences observed between populations for grain yield may suggest the importance of all or most polymorphic SNPs for the prediction of this trait, considering the small differences detected in the number of polymorphic markers even in the 1,000 SNP scenario for populations showing very different predictive ability values (e.g., K × A and K × I). Indeed, the total number of markers included in the model was largely inferior compared to the minimum estimated based on LD decay in the training set (4,929 vs. about 37,300), which means that populations with more polymorphic markers probably benefited from tracking a higher QTL number. Moreover, it should be considered that the effect of the tracked QTLs can vary depending on the specific validation population, further contributing to predictive ability differences between populations. Ultimately, for protein content, the inter-population GS model can still represent an interesting option to perform predictions on pooled breeding lines in the absence of models trained on closer genetic bases, while for grain yield its interest was limited to specific RIL populations. In this sense, in presence of a suboptimal marker number, investigating the number of polymorphic SNPs on the whole marker set or on a moderately large subset of top-effect markers in each test population may provide useful information about the potential of GS models trained on a different genetic base to predict grain yield and, to a lower extent, protein content breeding values.

The observed LD decay was much faster than that reported by Alemu et al. (2022) for a collection of 188 vining pea varieties and breeding lines from a single company, and by Crosta et al. (2022) for three connected RIL populations (r^2 = 0.2 was reached on average at 6,930,000 bp and about 100,000 bp, respectively, vs. 1,445 bp in our study), while being slower compared to the findings of Pavan et al. (2022) for a larger germplasm collection

(where r^2 = 0.2 was reached at 30 bp on average). These results are in line with the expectations, considering that the first two genetic bases were likely much narrower and underwent a considerably lower number of meiotic events, while the last one was probably wider relative to our germplasm collection. However, the only study using our same LD decay fitting method was that by Crosta et al. (2022), while other methods were employed in the other works, which may have affected the results. Despite the somewhat suboptimal sample size, the GWAS was able to detect significant associations for both grain yield and protein content, which, if validated by further studies, may be exploited for breeding purposes. Klein et al. (2020) identified significant QTLs for protein content in the same genomic region of chromosome 5 in which we found the only significant SNP for this trait. Significant loci for grain yield were detected in the same genomic regions of our significant SNPs by Gali et al. (2018; 2019) on chromosome 1 and by Crosta et al. (2022) on chromosome 6. The very fast LD decay possibly led to an underestimation of the number of significant SNPs due to the relatively low marker density, but at the same time it ensured an almost single gene resolution, which helped in the identification of candidate genes. Candidate genes of possible interest emerged for grain yield encoding proteins with regulatory functions (Psat1g031400, Psat4g098400, and Psat6g064800 coding for a protein kinase domain, a RNA recognition motif, and a helix-loop-helix DNA-binding domain, respectively) and a no apical meristem protein (Psat7g111400) that affects the position of meristems and primordia in other species (Souer et al., 1996; Sablowski and Meyerowitz, 1998), while the only candidate gene for protein content (Psat5g246720) codes for a protein from the rhomboid family.

In conclusion, our study produced GS models sufficiently accurate to enable the screening of germplasm resources for grain yield and protein content, with a potential interest also for the application to specific breeding materials. In addition, information about genomic areas involved in the control of the two traits was generated and, if confirmed by further studies, can be used in the selection process.

Note: the work presented in this chapter was mentioned in the other sections by referring to the relative publication (Crosta et al., 2023).

4. Genomic selection for pea grain yield, protein content, and protein yield: predictive ability in independent Italian environments for target and non-target genetic bases

4.1. Objectives

The main objective of this work was the investigation of GS predictive ability for pea grain yield, protein content, and protein yield in different environments, and on both the same and a different genetic base compared with those employed for model training.

4.2. Materials and methods

4.2.1. Plant material and phenotyping

GS model training set consisted of 276 RILs, of which 77 were issued from cross A × I, 96 from K × A, and 103 from K × I according to previous definitions, characterized in three autumn-sown environments of northern and central Italy (Lodi 2013-14, Lodi 2014-15, and Perugia 2013-14) for grain yield (Annicchiarico et al., 2019) and protein content (Crosta et al., 2022). Further details about the experimental setting, materials, environmental conditions, and phenotyping procedures can be found in these reports. GS model validation set relied on 131 RILs, which were randomly sorted from six populations and evaluated in Lodi (northern Italy) during the cropping seasons 2018-19 and 2019-20 (Picture 2). 19 RILs belonged to A × I, 23 to K × A, and 22 to K × I, while the other lines originated from three additional connected crosses, i.e., 23 from each of Dove × Attika (D × A) and Attika × Guifilo (A × G), and 21 from Alliance × Isard (C × I). All the parent lines were selected from a larger group of international cultivars because of high and stable grain yield, and moderate phenological differences between environments of northern and southern Italy (Annicchiarico, 2005; Annicchiarico and Iannucci, 2008). The large use of Attika as a parent in these crosses was due to its elevated competitive ability against weeds (Annicchiarico and Filippi, 2007), which is often advantageous under organic management. An autumn sowing (October 25) was adopted in cropping season 2018-19, while a winter sowing (December 10) was employed in 2019-20. The first validation cropping season, compared with the second one, featured greater winter cold stress and more rainfall, especially during late spring

(Appendix Table 11). Further details about the experimental setting and grain yield measurement can be found in Annicchiarico et al. (2021), while grain protein content was measured by NIRS method based on the calibration and models described in paragraph *2.2.2.*, and protein yield was obtained by multiplying grain yield by protein content plot values.

Picture 2. Validation field trial.

4.2.2. Heritability estimate

Broad-sense heritability for grain yield across validation trials was computed from variance components relative to genotype (S_G^2), genotype × year interaction (S_{GY}^2), and experimental error (S_e^2) estimated by REML method, according to the formula: $H^2 = S_G^2 / (S_G^2 + S_{GY}^2 / y + S_e^2 / y\ n)$, where y represents the number of cropping seasons, and n that of replicates in each experiment. Since protein content was measured on pooled replicate samples of each genotype in each experiment, it was not possible to compute broad-sense heritability for this trait and protein yield.

4.2.3. Genotyping and genomic data processing

For the GS training set, detailed information about DNA isolation, GBS, and quality filtering can be found in Annicchiarico et al. (2017a) or in paragraph *2.2.4*. The genotype dataset was filtered by MAF > 5%, mpm < 20%, mps < 25%, and SNP heterozygosity < 30%. For the GS validation set, GBS data were generated by the Elshire Group Ltd. according to the protocol established by Elshire et al. (2011) with some modifications, as described by Annicchiarico et al. (2021). Library sequencing was performed by using the Illumina HiSeq X platform and paired-end runs (2 × 150 bp). The SNP calling was performed according to the dDocent pipeline (Puritz et al., 2014), by aligning reads with pea reference genome (Kreplak et al.,

2019) v1a. The resulting vcf file was filtered for quality using vcftools (Danecek et al., 2011) with options -remove-indel, -minQ 30, -non-ref-af 0.00, -max-non-ref-af 0.999, and -max-missing 0.3, transformed in a 012 SNP matrix, and further filtered for MAF > 5%, mpm < 10%, mps < 25%, and SNP heterozygosity < 30%. Missing data were imputed according to k-nearest neighbours imputation method (Batista and Monard, 2002).

4.2.4. Genomic selection

GS models for grain yield, protein content, and protein yield were based on rrBLUP and 5,537 SNPs. Training was performed on phenotypic data of 276 RILs averaged across the three evaluation environments. Validation relied on 131 RILs not included in the training set, of which 64 originated from the same (hereafter named as target, and represented by lines from crossings A × I, K × A, and K × I) and 67 from a different (hereafter named as non-target, and represented by lines from crossings D × A, A × G, and C × I) genetic base compared with that used for training, and was performed both on data from the single evaluation environments and their mean. The filtering procedure retained 3,460 polymorphic SNPs for A × I, 4,740 for K × A, 4,487 for K × I, 2,981 for D × A, 3,217 for A × G, and 3,848 for C × I. Predictive ability was estimated for both the single RIL populations and the pooled lines of all the populations within the target and non-target genetic bases. Moreover, the number of polymorphic SNPs among those featuring the highest 100, 300, or 1,000 effects in absolute value according to GS models for grain yield and protein content was computed for each RIL population in the validation set to investigate its relationship with the within-population predictive ability. The choice of considering a maximum of 1,000 SNPs was due to most plants or livestock breeding simulations assuming 1,000 or less QTLs for polygenic traits (Brito et al. 2011; Yin et al. 2014; Wientjes et al. 2015; Yao et al. 2018; Strandén et al. 2019; Peters et al. 2020).

4.3. Results

4.3.1. Genomic selection

GS models validated on the same genetic base used for training displayed moderately high predictive ability values for all traits and estimation methods (within or across RIL populations), with a higher predictive performance observed for protein content (Table 12).

The use of data from different validation environments or their mean displayed an impact of variable size and direction on predictions, depending on the trait and predictive ability estimation method considered (Table 12). For the non-target genetic base, predictive ability was basically null for grain and protein yield independently from the estimation method, while resulting intermediate and modest for protein content according to within and across RIL population estimates, respectively (Table 12). Substantial predictive ability differences were detected between RIL populations of both genetic bases for most traits and validation datasets. K × I always showed by far the highest predictive ability for grain and protein yield in the target genetic base, while displaying similar results compared with A × I for protein content, whereas predictions for K × A resulted null for all traits (Table 13). For the non-target genetic base, C × I most often showed the highest predictive ability for grain and protein yield, despite usually modest, while the predictive performance for the other two RIL populations was low or null. For protein content, the predictive ability detected for each population tended to vary largely with the validation dataset (Table 13). For grain yield, the population ranking for the number of polymorphic SNPs remained quite constant across the three scenarios, with the differences between populations increasing with the number of top SNPs considered (Figure 18). Moreover, a trend towards a higher number of polymorphic SNPs in RIL populations belonging to the target genetic base was observed (Figure 18). On the other hand, for protein content the population ranking for the number of polymorphic SNPs was largely influenced by the number of top-effect SNPs considered (Figure 18).

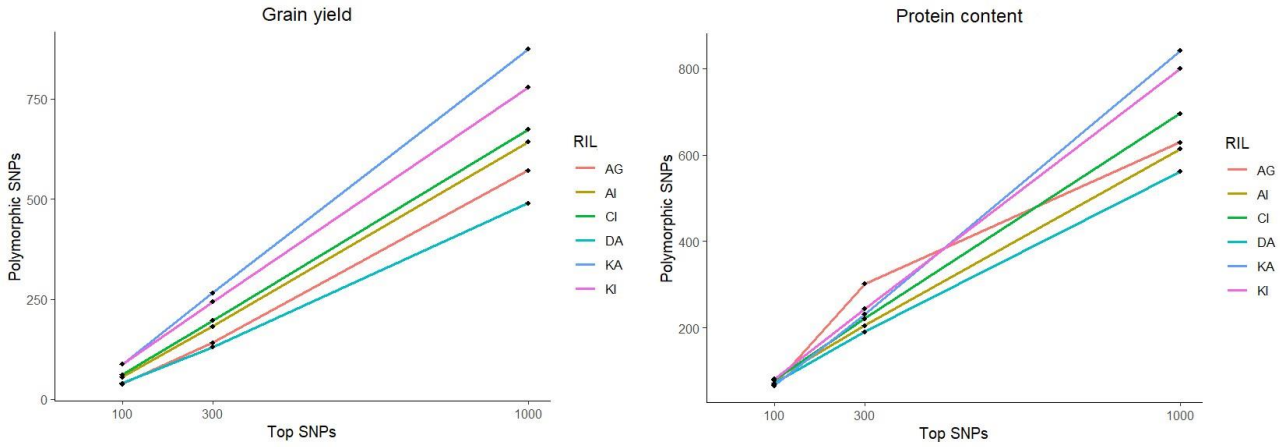
Table 12. Within and across RIL population predictive ability values for three pea traits obtained by rrBLUP GS models trained on 276 lines from three RIL populations issued by connected crosses and relying on 5,537 SNPs. Validation was performed on data of 131 lines from the same (target GB) or a different (non-target GB) genetic base compared with that used for training from two evaluation environments or their mean. Within RIL population predictive ability was computed by averaging the results across RIL populations forming each genetic base. Broad-sense heritability was reported only for grain yield because protein content was measured on pooled replicate material of each genotype in each environment.

		Predictive ability				
		Within-RIL		**Across-RIL**		
Trait	**Season**	**Target GB**	**Non-target GB**	**Target GB**	**Non-target GB**	H^2
Grain yield	2018-19	0.256	0.113	0.355	-0.087	0.70
Grain yield	2019-20	0.258	0.011	0.444	-0.100	0.72
Grain yield	mean	0.292	0.079	0.399	-0.110	
Protein content	2018-19	0.313	0.372	0.390	0.295	
Protein content	2019-20	0.425	0.314	0.449	0.117	
Protein content	mean	0.403	0.360	0.419	0.229	
Protein yield	2018-19	0.245	0.085	0.308	-0.155	
Protein yield	2019-20	0.267	-0.089	0.447	-0.269	
Protein yield	mean	0.279	0.003	0.378	-0.256	

Table 13. Predictive ability values for three pea traits displayed for each RIL population and obtained by rrBLUP GS models trained on 276 lines from three connected RIL populations and relying on 5,537 SNPs. Validation was performed on data of 131 independent lines from the same (target GB) or a different (non-target GB) genetic base compared with that used for training from two evaluation environments or their mean.

		Predictive ability					
		Target genetic base			Non-target genetic base		
Trait	Season	AI	KA	KI	AG	CI	DA
Grain yield	2018-19	0.368	-0.111	0.510	0.147	0.256	-0.063
Grain yield	2019-20	0.303	-0.142	0.613	0.047	0.136	-0.149
Grain yield	mean	0.407	-0.135	0.603	0.104	0.233	-0.100
Protein content	2018-19	0.575	-0.128	0.492	0.721	0.195	0.202
Protein content	2019-20	0.636	-0.075	0.714	0.288	0.331	0.322
Protein content	mean	0.708	-0.138	0.639	0.663	0.030	0.385
Protein yield	2018-19	0.387	-0.170	0.518	0.111	0.237	-0.094
Protein yield	2019-20	0.319	-0.181	0.662	0.030	-0.111	-0.186
Protein yield	mean	0.412	-0.191	0.616	0.074	0.101	-0.167

Figure 18. Number of polymorphic SNPs among the 100, 300, or 1,000 featuring the highest effect in absolute value according to GS models for grain yield and protein content displayed separately for the six RIL populations in the validation set (A × I, K × A, K × I, D × A, A × G, and C × I), which was formed by 131 lines. GS model training was performed on 276 lines from RIL populations A × I, K × A, and K × I.



4.4. Discussion

Despite moderately good, within RIL population predictive ability values for the target genetic base were inferior to those reported by Crosta et al. (2022) for the same materials and traits in a similar prediction scenario. However, while in the study by Crosta et al. (2022) the same genotypes were employed for training and validation, in the current work the lines in the training and validation set were different, which makes it closer to a real-life scenario.

For the non-target genetic base, the predictive performance remained almost constant relative to the target genetic base for protein content, while it basically dropped to zero for the other traits, in contrast with the modest predictive ability values reported by Crosta et al. (2022) for grain and protein yield in a similar scenario. This supports the hypothesis of a simpler genetic control of protein content compared with grain and protein yield, which seems in line with its higher within-trial broad-sense heritability and lower size of G × E relative to genetic effects detected by Crosta et al. (2022). While for the target genetic base across RIL population predictive ability always resulted superior to within population values, which was in accordance with the expectations, since the former estimation method can account for differences in population means, an opposite scenario characterized the non-target genetic base. This could be due to the exclusion of the non-target genetic base from the training set preventing GS models from predicting trait means of the relative RIL populations, while the likely lower variance of predicted and observed breeding values may have contributed to the advantage of within relative to across population predictions. The null predictive ability found for K × A for all traits and validation environments was completely unexpected and contrasted with previous results for the same material and prediction scenarios (Annicchiarico et al., 2019; Crosta et al., 2022, unpublished results), and with the fact that this population displayed the highest number of polymorphic SNPs. These contradictions, together with the fact that most of the lines from this population were genotyped by a different company compared with the other populations, may suggest the occurrence of some undefined problem during the genotyping procedure, possibly causing these unexpected results. The intention would be to repeat the genotyping of K × A lines soon to be more certain about the reliability of the genomic data available for this population. Despite less striking, also the superior predictions observed for K × I for grain and protein yield were surprising and may be attributable to the higher number of polymorphic markers characterizing this population relative to the others, except for K × A. In line with this rationale, the population ranking for the number of polymorphic SNPs, either computed on the whole marker set or on the SNPs featuring the highest 300 or 1,000 effects, perfectly reflected that for grain yield predictive ability, except for K × A. For protein content, the results of the 100 and 300 SNP scenarios were the most representative of the predictive ability values. Indeed, the population ranking for the number of polymorphic markers in the 100 SNP scenario reflected that for predictive ability in Lodi 2019-20, while the 300 SNP scenario highlighted a superior number of polymorphic SNPs in A × G, which was someway reflected by the elevated predictive ability detected for this population in Lodi 2018-19. Investigating

the number of polymorphic SNPs based on the whole marker set or on a moderately large subset of top-effect markers in each test population may provide useful information about the potential of GS models to predict grain yield, while this appears more difficult for protein content, which was likely controlled by a lower number of loci. Overall, our results appear promising for the use of GS in an inter-environment intra-population scenario for all the target traits, while the interest of inter-environment inter-population predictions seems to be limited to protein content and to specific germplasm sets for grain yield.

5. Comparison of genetic gains obtained by phenotypic and genomic selection on target and non-target genetic bases for pea grain and protein yield in Italian environments

5.1. Objectives

The main goal of this work was comparing PS and GS in terms of achieved genetic gains for grain and protein yield in Italian environments independent from the training ones, on both the same and a different genetic base relative to that employed for GS model training.

5.2. Materials and methods

5.2.1. Plant material, phenotyping, and selection process

Plant material consisted in the three or two top-performing lines selected for grain and protein yield, respectively, by GS and PS from each of six RIL populations, namely A × I, K × A, K × I, A × G, D × A, and C × I as previously defined, and the relative parental lines for a total of 43 genotypes (Appendix Table 10). GS was applied on an initial set of 63-93 RILs from each population (differences in line number were due to the filtering process) that were not included in the training set, while PS on a subset of 23 lines from each population. The over five-fold higher number, on average, of lines evaluated by GS compared with PS is due to the use of the same budget for both selection types and the higher individual cost of PS based on two-year data relative to GS (Annicchiarico et al., 2017c). These accessions were characterized by using a randomized block design with three replicates in three environments of northern and central Italy, of which two employed a late-winter sowing, namely Lodi 2022 (February 1) and Lodi 2023 (February 7; Picture 3), and one was sown during early-winter, namely Perugia 2022-23 (December 1, defined as Perugia 2023 hereafter) (Appendix Table 11). GS models based on BayesC were trained on mean genotype data of 306 RILs from A × I, K × A, and K × I (hereafter defined as GS target genetic base, as opposed to A × G, D × A, and C × I, defined as GS non-target genetic base) collected in Lodi 2013-14, Lodi 2014-15, and Perugia 2013-14, as described by Crosta et al. (2022). Filtering was performed separately for each material set, which was formed by validation RILs from a specific population and the whole training set, by using thresholds of MAF > 5%, mpm < 10%, and mps < 50% for

all populations, except for A × I, for which mpm < 15% was employed to ensure a SNP number comparable with the other populations. 4,831 markers were retained for A × I, 4,325 for K × A, 4,872 for K × I, 4,010 for A × G, 4,492 for C × I, and 4,867 for D × A. PS was based on mean genotype data of Lodi 2018-19 and Lodi 2019-20, as explained by Annicchiarico et al. (2021). Several traits were recorded on a plot basis, including: (1) dry grain yield, (2) grain protein content, which was measured by NIRS method based on the calibration and models described in paragraph *2.2.2.*, (3) onset of flowering, determined as the number of days since March 1 at which half of the plants displayed at least one open flower, and (4) farmer acceptability score, expressed on a 1-9 preference scale.

Picture 3. Field trial in Lodi 2023.

5.2.2. Statistical analyses of phenotypic data

The first analysis included only the RILs and was based on an ANOVA model with selection type (PS or GS), GS set (GS target or non-target genetic base), environment, RIL population within GS set, all the interactions of 2nd and 3rd degree between these factors, genotype nested within the crossing of GS set and selection type, G × E interaction, and replicate as fixed factors. The top-performing genotypes for grain yield were analysed by using this trait as the response variable, while those selected for protein yield were analysed by using grain and protein yield, and protein content by turn as the response variables. The second analysis included both the RILs and the parental lines and was based on an ANOVA model with genotype group (with four groups represented by the combination of selection type and GS set levels, and the fifth by parental lines), environment, genotype within genotype group, all the interactions of 2nd degree between these factors, and replicate as fixed factors. The top-performing genotypes for grain yield were analysed by using this trait, onset of flowering, and farmer acceptability score by turn as the response variables, while those selected for

protein yield were analysed by using grain and protein yield, and protein content by turn as the response variables. All the post-hoc comparisons between means for factors with more than two levels were performed, when the relative factor resulted significant in the ANOVA (alpha = 0.05), according to Duncan's test (alpha = 0.05) by duncan.test() function from R package agricolae. The sum of squares reported for the ANOVA were of type III. The rate of genetic progress achieved in one year by PS or GS for each RIL population was estimated by multiplying the percentage genetic gain relative to the parent mean by the number of selection cycles performed in one year, namely 2 for GS, and 1 or 0.5 for PS based on one- or two-year data, respectively.

5.3. Results

5.3.1. Statistical analyses of phenotypic data

ANOVA results (Appendix Table 12) of top-performing lines for grain yield highlighted a significant advantage of RILs selected by PS compared with GS, as well as of lines from the GS target relative to the non-target genetic base (Table 14). In line with ANOVA results (Appendix Table 12), according to which RIL population interaction with selection type was not significant, all RIL populations showed a better performance of PS lines compared with GS ones, as reflected also by the genetic gains of the two selection types with respect to parent mean (Table 15). Indeed, the genetic gains obtained by PS were much higher compared with GS for all RIL populations except for K × I, for which the two values were almost equivalent (Table 15).

Table 14. Grain yield mean of top-performing lines selected for this trait by either GS or PS, and of lines from RIL populations belonging (target) or not (non-target) to the GS training set with the associated standard error (SE). Different letters indicate a significant difference between means at alpha = 0.05 according to ANOVA results (Appendix Table 12).

Grain yield (t/ha)					
Selection type	**Mean**	**SE**	**GS set**	**Mean**	**SE**
PS	4.12 a	0.076	target	4.23 a	0.076
GS	3.63 b	0.076	non-target	3.52 b	0.076

Table 15. Grain yield mean of top-performing lines selected for this trait by either GS or PS from each RIL population (RIL), associated standard error (SE), and percentage genetic gains obtained by GS and PS with respect to parent mean (gain GS/PS %).

Grain yield (t/ha)					
RIL	**GS**	**PS**	**SE**	**Gain GS %**	**Gain PS %**
A × I	3.54	4.23	0.186	8.5	29.8
K × A	4.38	4.84	0.186	11.7	23.4
K × I	4.17	4.22	0.186	19.1	20.4
A × G	3.22	3.87	0.186	-4.8	14.4
C × I	3.23	3.76	0.186	1.3	17.7
D × A	3.23	3.79	0.186	-7.3	8.7

A very similar pattern was observed for the lines selected for protein yield by using this trait as the response variable, namely a significantly superior performance of PS compared with GS RILs, as well as of lines from the GS target relative to the non-target genetic base (Table 16; Appendix Table 13). Moreover, the same trend was observed for grain yield of these lines, confirming the strict relationship occurring between grain and protein yield, while protein content showed no significant difference between the two selection types, but a significant superiority of the lines from the GS target over those from the non-target genetic base (Table 16; Appendix Table 14 and 15). Like the top-performing lines for grain yield, also the best lines for protein yield showed no significant interaction in the ANOVA between RIL population and selection type for both the target trait and grain yield (Appendix Table 13; Appendix Table 14). Indeed, PS outperformed GS in terms of achieved genetic gains for protein yield in all RIL populations except for K × I and A × G, for which GS and PS displayed similar values (Table 17). Differently, the interaction between RIL population and selection type resulted significant for protein content (Appendix Table 15), for which the difference between PS and GS means was largely dependent on the RIL population (Table 17). When genetic gains were referred to one selection year, the superiority of PS over GS was dramatically reduced or nullified for RIL populations in the GS target genetic base, with GS displaying a marked advantage over both PS scenarios in K × I for grain and protein yield and in K × A for protein yield, and over the two-year PS scenario for both these populations and traits (Table 18). In contrast, PS still performed considerably better than GS in all cases for RIL populations out of the GS target genetic base, except for protein yield in A × G, for which GS ensured higher genetic progress per year than PS in both scenarios (Table 18).

Table 16. Mean of top-performing lines selected for protein yield by either GS or PS and belonging (target) or not (non-target) to RIL populations included in the GS training set for three pea traits with the associated standard error (SE). Different letters indicate a significant difference between means at alpha = 0.05 according to ANOVA results (Appendix Table 13, 14, and 15).

	Protein yield (t/ha)		Grain yield (t/ha)		Protein content (%)	
Selection type	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**
PS	1.00 a	0.025	4.04 a	0.101	24.83 a	0.053
GS	0.90 b	0.025	3.64 b	0.101	24.70 a	0.053
GS set	**Mean**	**SE**	**Mean**	**SE**	**Mean**	**SE**
target	1.02 a	0.025	4.12 a	0.101	25.00 a	0.053
non-target	0.87 b	0.025	3.56 b	0.101	24.52 b	0.053

Table 17. Mean of top-performing lines selected for protein yield by either GS or PS from each RIL population (RIL) and associated standard error (SE) for three pea traits. Percentage genetic gains obtained by GS and PS with respect to parent mean were displayed for protein yield. Different letters indicate significant differences between selection type means within each population at alpha = 0.05 according to ANOVA results (Appendix Table 13, 14, and 15) and comparisons based on least significant difference.

	Protein yield (t/ha)					Grain yield (t/ha)			Protein content (%)		
RIL	**GS**	**PS**	**SE**	**Gain GS %**	**Gain PS %**	**GS**	**PS**	**SE**	**GS**	**PS**	**SE**
A × I	0.84	1.03	0.061	8.15	32.32	3.49	4.22	0.247	24.30 a	24.45 a	0.130
K × A	1.10	1.17	0.061	14.26	21.33	4.36	4.73	0.247	25.29 a	24.86 b	0.130
K × I	1.01	1.00	0.061	18.13	16.25	4.03	3.92	0.247	25.55 a	25.58 a	0.130
A × G	0.94	0.92	0.061	12.90	11.24	3.76	3.74	0.247	24.61 a	24.61 a	0.130
C × I	0.76	0.97	0.061	-1.51	25.52	3.19	3.97	0.247	24.12 a	24.78 b	0.130
D × A	0.74	0.90	0.061	-12.37	7.77	3.04	3.69	0.247	24.33 a	24.69 b	0.130

Table 18. Percentage genetic gains relative to parent mean obtained in one year by GS and PS for each RIL population (RIL) according to two PS scenarios relying on experiments performed during a single (t_P = 1) or two (t_P = 2) years. The assumed duration of one GS cycle was of half a year.

	Grain yield (t/ha)				Protein yield (t/ha)			
	t_P = 1		t_P = 2		t_P = 1		t_P = 2	
RIL	**GS**	**PS**	**GS**	**PS**	**GS**	**PS**	**GS**	**PS**
A × I	17.0	29.8	17.0	14.9	16.3	32.3	16.3	16.2
K × A	23.5	23.4	23.5	11.7	28.5	21.3	28.5	10.7
K × I	38.2	20.4	38.2	10.2	36.3	16.3	36.3	8.1
A × G	-9.6	14.4	-9.6	7.2	25.8	11.2	25.8	5.6
C × I	2.5	17.7	2.5	8.8	-3.0	25.5	-3.0	12.8
D × A	-14.6	8.7	-14.6	4.4	-24.7	7.8	-24.7	3.9

ANOVA based on top performing RILs for grain yield and parental lines highlighted a significant effect of the genotype group and its interaction with the environment, when using grain yield as the response variable (Appendix Table 16). Accordingly, the group ranking varied greatly with the environment, except for PS lines belonging to the GS target genetic base that resulted top performing in all the environments (Table 19). Genotype mean data across environments confirmed the superiority of this genotype group relative to the others,

while GS lines from the non-target genetic base and parental lines emerged as the worst-performing groups, with range values mostly reflecting this pattern (Table 19). ANOVA models for the same lines with onset of flowering and farmer acceptability score as the response variables both highlighted a significant effect of the genotype group (Appendix Table 17; Appendix Table 18). Interestingly, GS lines resulted as more late flowering compared with PS ones within each GS set, as well as lines belonging to the GS target genetic base relative to those from the non-target genetic base. Parental lines and PS RILs from the GS non-target genetic base emerged as the groups featuring the significantly earliest flowering (Table 19). Farmer acceptability data followed a similar pattern, with PS lines displaying significantly higher values compared with GS ones within each GS set, as well as lines from the GS target relative to the non-target genetic base. RILs selected by GS from the non-target genetic base displayed the lowest farmer preference (Table 19). ANOVA based on parental lines and RILs selected for protein yield highlighted a significant effect of both genotype group and its interaction with the environment when protein yield was used as the response variable (Appendix Table 19). PS lines from the GS target genetic base emerged as the best group, despite with no significant difference from PS lines of the non-target genetic base in the single environments, and from GS lines from the target genetic base in Lodi and Perugia 2023 (Table 20). Genotype mean data across environments confirmed the superiority of PS lines from the GS target genetic base, but with no significant difference from GS lines belonging to the same material set, while PS proved significantly superior to GS for the non-target genetic base (Table 20). Grain yield mean genotype data of the same lines reflected exactly the pattern just described for protein yield (Appendix Table 20; Table 20), and a similar trend was detected also for protein content, with PS and GS being equivalent for the GS target genetic base, while PS resulted significantly better in the other material set (Appendix Table 21; Table 20).

Table 19. Grain yield means (t/ha) within or across environments and mean genotype range across environments for genotype groups, of which four consisted in the combination of selection type (PS or GS) and GS set (target and non-target) levels, and the fifth in parental lines. Genotype group means across environments were displayed also for flowering time (days since March 1) and farmer score (1-9). Different letters indicate significant differences at alpha = 0.05 between genotype group means based on ANOVA (Appendix Table 16, 17, and 18) and Duncan's test results.

	Grain yield					Flowering time	Farmer score
Material	**Lodi 2022**	**Lodi 2023**	**Perugia 2023**	**Mean**	**Range**	**Mean**	**Mean**
PS-target	3.19 a	5.65 a	4.45 a	4.43 a	3.83-5.07	59.3 b	5.8 b
GS-target	2.79 b	5.29 a	4.02 a	4.03 b	3.02-4.59	60.8 a	5.7 a
PS-non-target	3.12 a	4.98 a	3.31 b	3.80 b	3.43-4.23	56.8 d	5.5 d
GS-non-target	2.81 b	3.89 b	2.97 b	3.23 c	2.49-4.45	58.4 c	5.1 c
Parents	3.12 a	4.15 b	3.03 b	3.43 c	2.84-4.16	57.0 d	5.4 d

Table 20. Protein yield means within or across environments and mean genotype range across environments for genotype groups, of which four consisted in the combination of selection type (PS or GS) and GS set (target and non-target) levels, and the fifth in parental lines. Genotype group means across environments were displayed also for grain yield and protein content. Different letters indicate significant differences at alpha = 0.05 between genotype group means based on ANOVA (Appendix Table 19, 20, and 21) and Duncan's test results.

	Protein yield (t/ha)					Grain yield (t/ha)	Protein content (%)
Material	**Lodi 2022**	**Lodi 2023**	**Perugia 2023**	**Mean**	**Range**	**Mean**	**Mean**
PS-target	0.82 a	1.37 a	1.04 a	1.06 a	0.96-1.17	4.29 a	24.98 a
GS-target	0.74 b	1.26 a	0.95 a	0.98 ab	0.74-1.16	3.96 ab	25.04 a
PS-non-target	0.81 a	1.15 ab	0.83 ab	0.93 bc	0.84.1.05	3.80 bc	24.69 b
GS-non-target	0.77 ab	1.02 b	0.64 b	0.81 d	0.69-1.11	3.33 d	24.39 c
Parents	0.81 a	0.98 b	0.73 b	0.84 cd	0.67-1.04	3.43 cd	24.52 bc

5.4. Discussion

The choice of using two different sets of environments for GS training and PS evaluations was motivated by the attempt of sticking as much as possible to a real-life scenario, in which the options would be either to employ an existing GS model or to perform field trials to get data for PS. However, the difference in heritability between the two sets of environments likely played an important role in determining the superiority of PS compared with GS top-performing lines observed in the first analysis for grain and protein yield both across and within most RIL populations. Indeed, mean broad-sense heritability for grain yield was equal to 0.71 in PS environments (Annicchiarico et al., 2021), while amounting to 0.52 in GS training environments (Crosta et al., 2022). Mean protein yield broad-sense heritability amounted to 0.54 in GS training environments, whereas it was not possible to compute it in PS environments, since protein content was measured on pooled replicate material of each genotype. Anyway, the strict relationship observed between grain and protein yield in our

data and in the work by Crosta et al. (2022), suggests that the difference in mean heritability between PS and GS environments for protein yield was likely close to that detected for grain yield. Another reason behind PS advantage over GS might be the fact that one of PS environments, namely Lodi 2019-20, featured early winter sowing likewise the evaluation experiments of the top-performing lines, which were sown during early or late winter, while GS training environments were autumn-sown. This means that GS models likely accounted for a strong impact of cold tolerance especially on grain yield determination, but this aspect was probably not so important in the test environments of the best lines and in Lodi 2019-20, since winter sowing may have implied delayed plant emergence allowing for cold stress avoidance. Interestingly, the advantage of PS lines over GS ones was someway inferior for protein yield compared with grain yield, especially for the GS target genetic base (Table 19 and 20). This might be due to protein content featuring a higher within-trial broad-sense heritability and a lower influence of G × E compared with grain yield (Crosta et al., 2022), possibly mitigating the advantage of PS over GS training environments for protein yield. The significant farmer preference for genotypes selected by PS compared with GS for grain yield within each GS set was also probably related to the higher grain yield heritability and similarity to the test experiments of PS relative to GS training environments. In addition, the significantly earlier flowering of PS lines relative to GS ones within each GS set was likely due to higher terminal drought characterizing PS with respect to GS training environments, as supported by a lower rainfall amount during late winter and spring, especially in Lodi 2019-20 (Appendix Table 11). Based on these considerations, a scenario according to which grain yield in the PS and evaluation environments was primarily limited by terminal drought, whereas winter cold stress represented the main constraint in GS training experiments, can be hypothesized. Indeed, only in Lodi 2022, namely the evaluation environment featuring the lowest rainfall and so probably the highest terminal drought level (Appendix Table 11), the top-performing lines identified for grain yield by PS resulted significantly superior to those detected by GS within each GS set for this trait. As expected for the GS target genetic base, GS tended to become increasingly advantageous in terms of achieved genetic gains per year for grain and protein yield compared with PS when the latter was based on an increasing number of years, while the trend was the opposite for the non-target genetic base. In this context, the superior genetic gain per year achieved for protein yield by GS relative to both PS scenarios in A × G was quite surprising, especially considering that an opposite trend characterized grain yield selections. This might be related to an elevated GS predictive ability for protein content in this population, as suggested by the results of previous analyses

(paragraph *4.3.1.*). The significantly higher mean displayed by the GS target compared with the non-target genetic base for all traits and selections in the analysis excluding the parental lines, was confirmed by the analysis that included these genotypes, highlighting a significant superiority of the GS target genetic base within each selection type for target traits or their components across environments. This result was largely expected for GS, but not for PS, for which it was likely motivated by the superior breeding value of the GS target genetic base for the traits of interest that was at the base of its choice as GS model training set. In conclusion, GS resulted largely convenient in terms of achieved genetic gains per year for grain and protein yield compared with PS for material from the same genetic base used for GS model training, while the opposite tended to be true for materials not included in the GS training set, with possible exceptions for protein yield in specific germplasm sets.

6. Conclusions

Although GS concept was introduced in the early 2000's by Meuwissen et al. (2001), its large-scale application was possible only in more recent years thanks to the decrease of genotyping costs brought about by Next Generation Sequencing advancements. While several studies have been conducted about GS application to breeding of major cereals, little research work has been performed on minor crops, such as legumes. In this context, the current work was aimed at investigating the potential of GS for the prediction of grain yield, protein content, and their combination in pea in different scenarios.

The two GWAS performed for each of grain yield and protein content highlighted overall many significant markers spread across the genome, as well as many markers approaching significance in different genomic regions, suggesting a polygenic control of both traits. For this reason, although a finer gene mapping might be of interest to identify major-effect genes, GS appears as the most valid option to account for all the genomic regions influencing these traits. The higher GS predictive ability detected for protein content both in the intra- and inter-population prediction scenarios compared with grain and protein yield (paragraphs *2.3.2.*, *3.3.2.*, and *4.3.1.*) suggests a lower genetic complexity of the former trait, as supported by the higher within-trial broad-sense heritability (paragraphs *2.3.2.* and *3.3.2.*) and lower size of G × E relative to genetic effects (paragraph *2.3.1.*). Protein yield, which was regarded as the trait featuring the highest farmer interest, appeared mainly determined by grain yield both in breeding (paragraph *2.3.1.*) and germplasm material (paragraph *3.3.1.*), stressing the importance of improving this component to enhance protein production per unit of area. Moreover, the lack of substantial genetic correlation between grain yield and protein content detected in both narrow (paragraph *2.3.1.*) and wide genetic bases (paragraph *3.3.1.*) and in different environments encourages the simultaneous improvement of these traits. The mean predictive ability of GS models in the inter-environment, intra-population scenario resulted satisfactory for all the target traits in the narrow genetic base represented by three connected RIL populations, despite with some differences between the single populations employed for validation (paragraph *4.3.1.*). These encouraging results were confirmed in a wider genetic base, namely the worldwide germplasm collection, although in this case the validation was performed on data from the same environment used for model training (paragraph *3.3.2.*). On average, the GS inter-population prediction scenario implied a considerable predictive ability drop compared with the intra-population scenario, especially for grain and protein yield, although its size varied largely depending on the trait and the combination of training and

validation sets and environments (paragraphs *2.3.2.*, *3.3.2.*, and *4.3.1.*). Indeed, a remarkable predictive ability variation between different validation sets emerged for all traits in this GS scenario both for partially related or unrelated training and validation materials, as represented by sets of RIL populations with one common parent (paragraphs *2.3.2.* and *4.3.1.*), or a worldwide germplasm collection and three connected RIL populations (paragraph *3.3.2.*). A higher genetic diversity in the training and validation set seems to contribute to better GS performance especially for grain yield prediction (paragraphs *3.3.2.*, and *4.3.1.*), while predictive ability for protein content seems mainly influenced by the genetic diversity in a more restricted set of top-effect markers, at least in breeding material (paragraph *4.3.1.*). A better understanding of the mechanisms underlying these differences in GS model performance would have a great practical importance, possibly allowing the preliminary assessment of GS predictive potential for a specific germplasm set. This evaluation can also be conducted by estimating GS predictive ability on small subsets of lines from each material set and may enable an effective application of GS on specific materials not included in the training population, considering the high predictive ability values observed even for grain and protein yield of some RIL populations. Moreover, it cannot be excluded that, especially for GS models trained on the germplasm collection, increasing marker density may contribute to reduce the differences in predictive ability between populations, allowing to track most of the relevant QTLs for each validation population. GS showed a considerably higher predicted efficiency for protein yield improvement both in the intra- and inter-population scenarios compared with PS based on two-year testing (paragraph *2.3.3.*), which likely represents the minimum evaluation time, considering that climate change is increasing the relative importance of genotype × year and genotype × location × year variance components relative to genotype × location one (Annicchiarico, 2020; Crosta et al., 2022). While our results in terms of achieved genetic gains tended to confirm those of predicted efficiency for RIL populations belonging to the GS target genetic base for grain and protein yield, the scenario was the opposite for the non-target genetic base, despite the occurrence of relevant variation between traits and populations in both material sets (paragraph *5.3.1.*). These results suggest that using the same environments for GS training and PS evaluations may favour an unbiased comparison of the achieved genetic gains by reducing the influence of environmental confounding factors, which in our case implied a much higher grain yield heritability in PS compared with GS environments. In addition, adopting a similar sowing time for both selection and evaluation environments should further decrease the background noise. Finally, it is useful to remind that the use of a limited number

of GS training environments and the application of GS models to predict target traits in genetic bases differing from the training one, respond to the need to limit the phenotyping investment for a cost-efficient application of GS. In this regard, it is fundamental to emphasize once again the importance of conducting in parallel the development of the GS training set by methods allowing for accelerated generation advancement (growth chamber or greenhouse) and that of the target selection population to exploit GS reduced cycle length relative to PS.

Overall, the main conclusions that can be drawn from this work are:

- Protein content is an easier breeding target than grain yield, because of lower G × E size relative to genetic effects and higher within-trial broad-sense heritability
- Grain yield and protein content do not show sizeable inverse genetic correlation in most conditions and materials, thereby facilitating their simultaneous improvement
- Grain yield has a much higher impact on protein yield than protein content
- The expected polygenic control of grain yield and protein content was confirmed
- GS can be a valid method for the identification of superior genotypes for grain yield, protein content, and protein yield in an intra-population scenario, showing greater achieved genetic gains per unit time than PS
- The application of GS models on a different genetic base from that employed for training can be a valid option for protein content improvement, while its value for grain and protein yield depends on the specific genetic base

7. References

- Abbo, S., and Gopher, A. (2017). Near eastern plant domestication: a history of thought. *Trends Plant Sci.* 22, 491–511. doi: 10.1016/j.tplants.2017.03.010
- Al Bari, M. A., Zheng, P., Viera, I., Worral, H., Szwiec, S., Ma, Y., et al. (2021). Harnessing genetic diversity in the USDA pea germplasm collection through genomic prediction. *Front. Genet.* 12, 707754. doi: 10.3389/fgene.2021.707754
- Alberta Pulse Growers. (n.d.). *Field Peas*. Retrieved December 1, 2023, from https://albertapulse.com/growing-peas/
- Alemu, A., Brantestam, A. K., and Chawade, A. (2022). Unraveling the genetic basis of key agronomic traits of wrinkled vining pea (*Pisum sativum* L.) for sustainable production. *Front. Plant Sci.* 13, 844450. doi: 10.3389/fpls.2022.844450
- Alessandri, A., De Felice, M., Zeng, N., Mariotti, A., Pan, Y., Cherchi, A., et al. (2014). Robust assessment of the expansion and retreat of Mediterranean climate in the 21st century. *Sci. Rep.* 4, 7211. doi:10.1038/srep07211
- Ali, M., and Sarker, A. (2013). More benefit from less land: a new rice-pea-rice cropping pattern for resource-poor farmers of Bangladesh. *J. Agric. Sci. Technol.* 3, 204–210
- Ambrose, M. J. (2004). A novel allele at the afila (Af) locus and new alleles at the tendril-less (Tl) locus. *Pisum Genetics* 36, 1-2
- Ambrose, M. (2008). Garden pea. In Prohens, J., and Nuez, F. (Ed.), *Vegetables II: Fabaceae, Liliaceae, Solanaceae, and Umbelliferae* (pp. 3-26). Springer New York. doi: 10.1007/978-0-387-74110-9
- Anderson, V. L., Lardy, G. P., and Ilse, B. R. (2007). Field pea grain for beef cattle. *Prof. Anim. Sci.* 23, 1-7. doi: 10.1532/S1080-7446(15)30931-1
- Andridge, R. R., and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* 78, 40–64. doi: 10.1111/j.1751-5823.2010.00103.x
- Angus, J. F., Kirkegaard, J. A., Hunt, J. R., Ryan, M. H., Ohlander, L., and Peoples, M. B. (2015). Break crops and rotations for wheat. *Crop Pasture Sci.* 66, 523–552. doi: 10.1071/CP14252
- Annicchiarico, P. (2005). Scelta varietale in pisello e favino rispetto all'ambiente e all'utilizzo. *Inf. Agrar.* 61, 47–52
- Annicchiarico, P. (2008). Adaptation of cool-season grain legume species across climatically-contrasting environments of southern *Europe. Agron. J.* 100, 1647–1654. doi: 10.2134/agronj2008.0085
- Annicchiarico, P. (2017). Feed legumes for truly sustainable crop animal systems. *Ital. J. Agron.* 12, 151–160. doi: 10.4081/ija.2017.880
- Annicchiarico, P., and Filippi L. (2007). A field pea ideotype for organic systems of Northern Italy. *J. Crop. Improv.* 20, 193- 203. doi: 10.1300/J411v20n01_11
- Annicchiarico, P., and Iannucci, A. (2007). Winter survival of pea, faba bean and white lupin cultivars across contrasting Italian locations and sowing times, and implications for selection. *J. Agric. Sci.* 145, 611–622. doi: 10.1017/S0021859607007289
- Annicchiarico, P., and Iannucci, A. (2008). Adaptation strategy, germplasm type and adaptive traits for field pea improvement in Italy based on variety responses across climatically contrasting environments. *Field Crops Res.* 108, 133-142. doi: 10.1016/j.fcr.2008.04.004
- Annicchiarico, P., Nazzicari, N., Pecetti, L., Romani, M., Ferrari, B., Wei, Y., et al. (2017a). GBS-based genomic selection for pea grain yield under severe terminal drought. *Plant Genome* 10, plantgenome2016.07.0072. doi: 10.3835/plantgenome2016.07.0072
- Annicchiarico, P., Romani, M., Cabassi, G., and Ferrari, B. (2017b). Diversity in a pea (*Pisum sativum*) world collection for key agronomic traits in a rain-fed environment of Southern Europe. *Euphytica* 213, 245. doi: 10.1007/s10681-017-2033-y
- Annicchiarico, P., Nazzicari, N., Wei, Y., Pecetti, L., and Brummer, E. C. (2017c). Genotyping-by-sequencing and its exploitation for forage and cool-season grain legume breeding. *Front. Plant Sci.* 8, 679. doi: 10.3389/fpls.2017.00679

- Annicchiarico, P., Nazzicari, N., Pecetti, L., Romani, M., and Russi, L. (2019). Pea genomic selection for Italian environments. *BMC Genom.* 20, 603. doi: 10.1186/s12864-019-5920-x
- Annicchiarico, P., Nazzicari, N., Laouar, M., Thami-Alami, I., Romani, M., and Pecetti, L. (2020). Development and proof-of-concept application of genome-enabled selection for pea grain yield under severe terminal drought. *Int. J. Mol. Sci.* 21, 2414. doi: 10.3390/ijms21072414
- Annicchiarico, P., Nazzicari, N., Notario, T., Monterrubio Martin, C., Romani, M., Ferrari, B., et al. (2021). Pea breeding for intercropping with cereals: variation for competitive ability and associated traits, and assessment of phenotypic and genomic selection strategies. *Front. Plant Sci.* 12, 731949. doi: 10.3389/fpls.2021.731949
- Aubert, G., Kreplak, J., Leveugle, M., Duborjal, H., Klein, A., Boucherot, K., et al. (2023). SNP discovery by exome capture and resequencing in a pea genetic resource collection. *PeerJ* 3, e100. doi: 10.24072/pcjournal.332
- Balázs, B., Kelemen, E., Centofanti, T., Vasconcelos, M. W., and Iannetta, P. P. M. (2021). Integrated policy analysis to identify transformation paths to more sustainable legume-based food and feed value-chains in Europe. *Agroecol. Sust. Food* 45, 931-953. doi: 10.1080/21683565.2021.1884165
- Baldoni, R., and Giardini, L. (1981). Coltivazioni erbacee (1st ed.). Pàtron Editore
- Baldoni, R., and Giardini, L. (2001). Coltivazioni erbacee (3rd ed.). Pàtron Editore
- Barbieri, P., Pellerin, S., Seufert, V., Smith, L., Ramankutty, N., and Nesme, T. (2021). Global option space for organic agriculture is delimited by nitrogen availability. *Nat. Food*, 2, 363-372. doi: 10.1038/s43016-021-00276-y
- Bărbieru, A. (2021). Correlations between yield and several traits in a set of winter pea cultivars. *Rom. Agric. Res.* 38, 2021–2045
- Bastianelli, D., Grosjean, F., Peyronnet, C., Duparque, M., and Regnier, J. M. (1998). Feeding value of pea (*Pisum sativum*, L.) 1. Chemical composition of different categories of pea. *Animal science* 67, 609-619. doi: 10.1017/S1357729800033051
- Batista, G. E., and Monard, M. C. (2002). A study of K-nearest neighbour as an imputation method. *His* 87, 251-260
- Bénézit, M., Biarnès, V., and Jeuffroy, M. H. (2017). Impact of climate and diseases on pea yields: what perspectives with climate change? *OCL* 24, D103. doi: 1051/ocl/2016055
- Blixt, S. (1963). A presentation of the Lamprechtian *Pisum*-material. Report from *Weibullsholm Plant Breed. Inst. Sweden.*
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/ btu170
- Brenes, A., Trevino, J., Centeno, C., and Yuste, P. (1989). Influence of peas (*Pisum sativum*) as a dietary ingredient and flavomycin supplementation on the performance and intestinal microflora of broiler chicks. *Br. Poult. Sci.* 30, 81-89. doi: 10.1080/00071668908417127
- Brito, F. V., Neto, J. B., Sargolzaei, M., Cobuci, J. A., and Schenkel, F. S. (2011). Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. *BMC Genet.* 12, 1-10. doi: 10.1186/1471-2156-12-80
- Brito, D. S., Quinhones, C. G., Neri-Silva, R., Heinemann, B., Schertl, P., Cavalcanti, J. H. F., et al. (2022). The role of the electron-transfer flavoprotein: ubiquinone oxidoreductase following carbohydrate starvation in *Arabidopsis* cell cultures. *Plant Cell Rep.* 1-16. doi: 10.1007/s00299-021-02822-1
- Brondizio, E. S., Settele, J., Díaz, S., and Ngo, H. T. (2019). *Global Assessment Report on Biodiversity and Ecosystem Services.* (Report No. 978-3-947851-20-1). Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). https://ipbes.net/global-assessment. doi: 10.5281/zenodo.3831673
- Bues, A., Preissel, S., Reckling, M., Zander, P., Kuhlman, T., Topp, K., et al. (2013). *The environmental role of protein crops in the new common agricultural policy* (Report No. 2012-067). European Union. https://library.wur.nl/WebQuery/wurpubs/fulltext/262633
- Burstin, J., Marget, P., Huart, M., Moessner, A., Mangin, B., Duchene, C., et al. (2007). Developmental genes have pleiotropic effects on plant morphology and source capacity, eventually impacting on seed protein content and productivity in pea. *Plant Physiol.* 144, 768-781. doi: 10.1104/pp.107.096966

- Burstin, J., Gallardo, K., Mir, R. R., Varshney, R. K., and Duc, G. (2011). Improving protein content and nutrition quality. In Pratap, A., and Kumar, J. (Ed.), *Biology and Breeding of Food Legumes* (pp. 314-328). CAB International
- Burstin, J., Salloignon, P., Chabert-Martinello, M., Magnin-Robert, J. B., Siol, M., Jacquin, F., et al. (2015). Genetic diversity and trait genomic prediction in a pea diversity panel. *BMC Genom.* 16, 105. doi: 10.1186/s12864-015-1266-1
- Burstin, J., Kreplak, J., Macas, J., and Lichtenzveig, J. (2020). *Pisum sativum* (pea). *Trends Genet.* 36, 312-313. doi: 10.1016/j.tig.2019.12.009
- Carranca, C., De Varennes, A., and Rolston, D. (1999). Biological nitrogen fixation by fababean, pea and chickpea, under field conditions, estimated by the 15N isotope dilution technique. *Eur. J. Agron.* 10, 49-56. doi: 10.1016/S1161-0301(98)00049-5
- Carrouée, B., Crépon, K., and Peyronnet, C. (2003). Les protéagineux: intérêt dans les systèmes de production fourragers français et européens. *Fourrages* 174, 163-182.
- Cesar Australia. (n.d.). *Pea weevil.* Retrieved November 22, 2023, from https://cesaraustralia.com/pestnotes/beetles/pea-weevil/
- Christopher, S. F., and Lal, R. (2007). Nitrogen management affects carbon sequestration in North American cropland soils. *CRC Crit. Rev. Plant Sci.* 26, 45-64. doi: 10.1080/07352680601174830
- Clément, T., Joya, R., Bresson, C., Clément, C., Serra Caracciolo, A., Simons, J., et al. (2018). *Market developments and policy evaluation aspects of the plant protein sector in the EU.* (Report No. KF-03-18-447-EN-N). European Commission. https://data.europa.eu/doi/10.2762/022741
- Corre-Hellou, G., and Crozat, Y. (2005). N_2 fixation and N supply in organic pea (*Pisum sativum* L.) cropping systems as affected by weeds and peaweevil (*Sitona lineatus* L.). *Eur. J. Agron.* 22, 449-458. doi: 10.1016/j.eja.2004.05.005
- Cousin, R., Messager, A., and Vingère, A. (1985). Breeding for yield in combining peas. In Hebblethwaite, P. D., Heath, M. C., and Dawkins, T. C. K. (Ed.), *The Pea Crop* (pp. 115–129). London: Butterworths.
- Cousin, R. (1997). Peas (*Pisum sativum* L.). *Field Crops Res.* 53, 111-130. doi: 10.1016/S0378-4290(97)00026-9
- Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PloS One* 11, e0156744. doi: 10.1371/journal.pone.0156744
- Craig, J., Lloyd, J. R., Tomlinson, K., Barber, L., Edwards, A., Wang, T. L., et al. (1998). Mutations in the gene encoding starch synthase II profoundly alter amylopectin structure in pea embryos. *Plant Cell* 10, 413–426. doi: 10.1105/tpc.10.3.413
- Craig, J., Barratt, P., Tatge, H., Dejardin, A., Handley, L., Gardner, C. D., et al. (1999). Mutations at the *rug4* locus alter carbon and nitrogen metabolism of pea plants through an effect on sucrose synthase. *Plant J.* 17, 353–362. doi: 10.1046/j.1365-313X.1999.00382.x
- Crews, T. E., and Peoples, M. B. (2004). Legume versus fertilizer sources of nitrogen: ecological tradeoffs and human needs. *Agr. Ecosyst. Environ.* 102, 279-297. doi: 10.1016/j.agee.2003.09.018
- Crosta, M., Nazzicari, N., Ferrari, B., Pecetti, L., Russi, L., Romani, M., et al. (2022). Pea grain protein content Across Italian environments: genetic relationship with grain yield, and opportunities for genome-enabled selection for protein yield. *Front. Plant Sci.* 12, 718713. doi: 10.3389/fpls.2021.718713
- Crosta, M., Romani, M., Nazzicari, N., Ferrari, B., and Annicchiarico, P. (2023). Genomic prediction and allele mining of agronomic and morphophysiological traits in pea germplasm collections. *Front. Plant Sci.* 14, 1320506. doi: 10.3389/fpls.2023.1320506
- Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., and Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* 200, 1341-1348. doi: 10.1534/genetics.115.178038
- Dahl, W. J., Foster, L. M., and Tyler, R. T. (2012). Review of the health benefits of peas (*Pisum sativum* L.). *Br. J. Nutr.* 108, S3-S10. doi:10.1017/S0007114512000852
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). 1000 genomes project analysis group. The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/ btr330

- Davies, D. R. (1977). Restructuring the pea plant, *Sci. Prog.* 64, 201-214
- Davis, P. H. (1970). *Pisum* L. In: Davis, P.H. (Ed.), *Flora of Turkey. Vol. 3* (pp. 370–372). Edinburgh Univ. Press.
- De Visser, C. L. M., Schreuder, R., and Stoddard, F. (2014). The EU's dependency on soya bean import for the animal feed industry and potential for EU produced alternatives. *OCL* 21, D407. doi: 10.1051/ocl/2014021
- DeLacy, I. H., Basford, K. E., Cooper, M., Bull, I. K., and McLaren, C. G. (1996). Analysis of multi-environment trials: an historical perspective. In Cooper, M., and Hammer, G. L. (Ed.), *Plant Adaptation and Crop Improvement* (pp. 39–124). CAB International
- Doré, T., Meynard, J. M., and Sebillotte, M. (1998). The role of grain number, nitrogen nutrition and stem number in limiting pea crop (*Pisum sativum*) yields under agricultural conditions. *Eur. J. Agron.* 8, 29–37. doi: 10.1016/S1161-0301(97)00006-3
- Doudna, J. A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science* 346, 1258096. doi: 10.1126/science.1258096
- Dröge-Laser, W., Snoek, B. L., Snel, B., and Weiste, C. (2018). The Arabidopsis bZIP transcription factor family - an update. *Curr. Opin. Plant Biol.* 45, 36-49. doi: 10.1016/j.pbi.2018.05.001
- Duc, G., Agrama, H., Bao, S., Berger, J., Bourion, V., De Ron, A. M., et al. (2015). Breeding annual grain legumes for sustainable agriculture: new methods to approach complex traits and target new cultivar ideotypes. *Crit. Rev. Plant Sci.* 34, 381–411. doi: 10.1080/07352689.2014.898469
- Elgert, L. (2013). *Shifting the debate about "responsible soy" production in Paraguay. A critical analysis of five claims about environmental, economic, and social sustainability.* (Report No. LDPI Working Paper 23). The Land Deal Politics Initiative. https://d1wqtxts1xzle7.cloudfront.net/31343336/LDPIWP23elgert-libre.pdf?1392441579=&response-content-disposition=inline%3B+filename%3DShifting_the_debate_about_responsible_so.pdf&Expires=1656604468&Signature=HDJbTVr0q0UfmKUG~9pFS0ZKy30eH2oYzHNpMMRmITAcCZz7BxVPIG61yCkOwkPawSY662WYZ7evBK7uVxMKzKDAgyOr1hHWEsu1N8xNVJlyXLFPykPcIS8vwzOzbImsRtrKuucmwiySJdXMf9Br2aszJn3wN-pqje6jptICHbrv~JDERoHk2QzYpDF9U-8xwU2hOUNuErKfbhbDYgloLN-SyOMmMbkf~fX7lrjysD4ok86M6pyPw8xzDANvYmN19IrmtMNa2An9l8smPDw-lgxIh210sFZRUw1WqDL39YFW4DOwwZP3FB7myDPkaCW~Iiq~lwUuZRlMK9TNmHIt0A__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, e19379. doi: 10.1371/journal.pone.0019379
- European Commission. (1993). Implementation of memorandum of understanding on oilseeds. https://aei.pitt.edu/101406/1/P%2D34.pdf
- European Commission. (2018). *Report from the commission to the council and the European Parliament: on the development of plant proteins in the European Union.* https://ec.europa.eu/info/sites/default/files/progress_report_romania_com2021_370_fina.pdf
- European Commission. (n.d.). *Key reforms in the new CAP.* Retrieved October 30, 2023, from https://ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/new-cap-2023-27/key-reforms-new-cap_en#agreenerpolicy
- European Feed Manufacturers' Federation (FEFAC). (n.d.). Retrieved October 30, 2023, from *EU Protein Plan.* Fefac. https://fefac.eu/priorities/markets-trade/eu-protein-plan/#:~:text=The%20main%20purpose%20of%20the%20EU%20Protein%20Plan,on%20development%20of%20plant%20proteins%20in%20the%20EU%29
- Falconer, D. S. (1989). Introduction to quantitative genetics (3rd ed.). Harlow: Longman. doi: 10.1017/S0016672300028573

- Fang, Y., and L. Xiong. (2015). General mechanisms of drought response and their application in drought resistance improvement in plants. *Cell. Mol. Life Sci.* 72, 673–689. doi: 10.1007/s00018-014-1767-0
- Ferrari, M., Marcon, E., and Menta, A. (2006). Fitopatologia, entomologia agraria e biologia applicata. Edagricole.
- Ferrari, B., Romani, M., Aubert, G., Boucherot, K., Burstin, J., Pecetti, L., et al. (2016). Association of SNP markers with agronomic and quality traits of field pea in Italy. *Czech J. Genet. Plant* 52, 83–93. doi: 10.17221/22/2016-CJGPB
- Food and Agriculture Organization of the United Nations (FAO). (n.d.a). *FAOSTAT*. Retrieved October 21, 2023 from http://www.fao.org/faostat/en/#compare
- Food and Agriculture Organization of the United Nations (FAO). (n.d.b). *FAOSTAT*. Retrieved October 21, 2023 from http://www.fao.org/faostat/en/#data/QCL
- Food and Agriculture Organization of the United Nations (FAO). (n.d.c). *FAOSTAT*. Retrieved October 21, 2023 from https://www.fao.org/faostat/en/#data/QCL
- Food and Agriculture Organization of the United Nations (FAO). (n.d.d). *FAOSTAT*. Retrieved October 21, 2023 from http://www.fao.org/faostat/en/#data/QC.FAO
- Gali, K. K., Liu, Y., Sindhu, A., Diapari, M., Shunmugam, A. S., Arganosa, G., et al. (2018). Construction of high-density linkage maps for mapping quantitative trait loci for multiple traits in field pea (*Pisum sativum* L.). *BMC Plant Biol.* 18, 172. doi: 10.1186/s12870-018-1368-4
- Gali, K. K., Sackville, A., Tafesse, E. G., Lachagari, V. B., McPhee, K., Hybl, M., et al. (2019). Genome-wide association mapping for agronomic and seed quality traits of field pea (*Pisum sativum* L.). *Front. Plant Sci.* 10, 1538. doi: 10.3389/fpls.2019.01538
- Gatel, F., and Grosjean, F. (1990). Composition and nutritive value of peas for pigs: A review of European results. *Livest. Prod. Sci.* 26, 155-175. doi: 10.1016/0301-6226(90)90077-J
- Gemede, H. F., and Ratta, N. (2014). Antinutritional factors in plant foods: potential health benefits and adverse effects. *Int. J. Food Sci. Nutr.* 3, 284-289. doi: 10.11648/j.ijnfs.20140304.18
- George, E. I., and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc*. 88, 881-889. doi: 10.2307/2290777
- Ghosh, I. (2022, October 24). *The Domino Effects of Tropical Deforestation.* Visual Capitalist. https://www.visualcapitalist.com/sp/the-domino-effects-of-tropical-deforestation/
- Gopinath, K. A, Saha, S., Mina, B. L., Pande, H., Kumar, N., and Srivastva, A. K. (2009). Yield potential of garden pea (*Pisum sativum* L.) varieties, and soil properties under organic and integrated management systems. *Arch. Agron. Soil Sci.* 55, 157-167. doi: 10.1080/03650340802382207
- Guilioni, L., Wery, J., and Lecoeur, J. (2003). High temperature and water deficit may reduce seed number in field pea purely by decreasing plant growth rate. *Funct. Plant Biol.* 30, 1151–1164. doi: 10.1071/FP03105
- Gust, A. A., Willmann, R., Desaki, Y., Grabherr, H. M., and Nürnberger, T. (2012). Plant LysM proteins: modules mediating symbiosis and immunity. *Trends Plant Sci.* 17, 495-502. doi: 10.1016/j.tplants.2012.04.003
- Ha, K. V., Marschner, P., Bünemann, E. K., and Smernik, R. J. (2007). Chemical changes and phosphorus release during decomposition of pea residues in soil. *Soil Biol. Biochem.* 39, 2696-2699. doi: 10.1016/j.soilbio.2007.05.017
- Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinform.* 12, 186. doi: 10.1186/1471-2105-12-186
- Han, G., Qiao, Z., Li, Y., Yang, Z., Wang, C., Zhang, Y., et al. (2022). RING zinc finger proteins in plant abiotic stress tolerance. *Front. Plant Sci.* 13, 877011. doi: 10.3389/fpls.2022.877011
- Hardwick, R. C., Andrews, D. J., Hole, C. C., and Salter, P. J. (1979). Variability in number of pods and yield in commercial crops of vining peas (*Pisum sativum* L.). *J. Agric. Sci*. 92, 675-681. doi: 10.1017/S0021859600053910
- Harland, S. C. (1948). Inheritance of immunity to mildew in Peruvian forms of *Pisum sativum*. *Hered.* 2, 263-269

- Häusling, M. (2011). *Report: The EU protein deficit: What solution for a long-standing problem?* (Report No. 2010/2111(INI)). European Parliament. Committee on Agriculture and Rural Development. http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&reference=A7-2011-0026&language=EN
- Hay, F. J. (2019, April 3). *Soybeans for biodiesel production.* Farm energy. https://farm-energy.extension.org/soybeans-for-biodiesel-production/
- Hedley, C. L., and Ambrose, M. J. (1981). Designing 'leafless' plants for improving the dried pea crop. *Adv. Agron.* 34, 225-277. doi: 10.1016/S0065-2113(08)60888-3
- 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
- Heffner, E. L., Lorenz, A. J., Jannink, J. L., and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci.* 50, 1681–1690. doi: 10.2135/cropsci2009.11.0662
- Henseler, M., Piot-Lepetit, I., Ferrari, E., Mellado, A. G., Banse, M., Grethe, H., et al. (2013). On the asynchronous approvals of GM crops: potential market impacts of a trade disruption of EU soy imports. *Food Policy* 41, 166-176. doi: 10.1016/j.foodpol.2013.05.005
- Hradilova, I., Trněný, O., Valkova, M., Cechová, M., Janská, A., Prokešová, L., et al. (2017). A combined comparative transcriptomic, metabolomic, and anatomical analyses of two key domestication traits: pod dehiscence and seed dormancy in pea (*Pisum* sp.). *Front. Plant Sci*. 8, 542. doi: 10.3389/fpls.2017.00542
- Huang, M., Liu, X., Zhou, Y., Summers, R. M., and Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience* 8, 154. doi: 10.1093/gigascience/giy154
- Iannetta, P. P. M., Hawes, C., Begg, G. S., Maaß, H., Ntatsi, G., Savvas, D., et al. (2021). A multifunctional solution for wicked problems: value-chain wide facilitation of legumes cultivated at bioregional scales is necessary to address the climate-biodiversity-nutrition nexus. *Front. Sustain. Food Syst.* 5, 692137. doi: 10.3389/fsufs.2021.692137
- Irzykowska, L., and Wolko, B. (2004). Interval mapping of QTLs controlling yield-related traits and seed protein content in *Pisum sativum*. *J. Appl. Genet.* 45, 297-306.
- Jaiswal, S. K., and Dakora, F. D. (2019). Widespread distribution of highly adapted *Bradyrhizobium* species nodulating diverse legumes in Africa. *Front. Microbiol.* 10, 310. doi: 10.3389/fmicb.2019.00310
- Janss, L., and Ramstein, G. P. (2023). *Statistical Models for Genomic Prediction in Animals and Plants*. Brightspace. https://brightspace.au.dk/d2l/le/lessons/103685/topics/1481775
- Jensen, E.S. (1986). The influence of rate and time of nitrate supply on nitrogen fixation and yield in pea (*Pisum sativum* L.). *Fertil. Res.* 10, 193–202. doi: 10.1007/BF01049349
- Jensen, E. S., and Hauggaard-Nielsen, H. (2003). How can increased use of biological N_2 fixation in agriculture benefit the environment? *Plant Soil* 252, 177-186. doi: 10.1023/A:1024189029226
- Jensen, E. S., Peoples, M. B., Boddey, R. M., Gresshoff, P. M., Hauggaard-Nielsen, H., Alves, B. J. R., et al. (2012). Legumes for mitigation of climate change and the provision of feedstock for biofuels and biorefineries. A review. *Agron. Sustain. Dev.*, 32, 329-364. doi: 10.1007/s13593-011-0056-7
- Jha, A. B., Tar'an, B., Diapari, M., and Warkentin, T. D. (2015). SNP variation within genes associated with amylose, total starch and crude protein concentration in field pea. *Euphytica* 206, 459–471. doi: 10.1007/s10681-015-1510-4
- Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521
- Karjalainen, R., and Kortet, S. (1987). Environmental and genetic variation in protein content of peas under northern growing conditions and breeding implications. *Agric. Food Sci.* 59, 1–9. doi: 10.23986/afsci.72238
- Karkanis, A., Ntatsi, G., Kontopoulou, C. K., Pristeri, A., Bilalis, D., and Savvas, D. (2016). Field pea in European cropping systems: adaptability, biological nitrogen fixation and cultivation practices. *Not. Bot. Horti Agrobot. Cluj Napoca* 44, 325-336. doi: 10.15835/NBHA44210618

- Khorasani, G. R., Okine, E. K., Corbett, R. R., and Kennelly, J. J. (2001). Nutritive value of peas for lactating dairy cattle. *Can. J. Anim. Sci*. 81, 541-551. doi: 10.4141/A01-019
- Klein, A., Houtin, H., Rond, C., Marget, P., Jacquin, F., Boucherot, K., et al. (2014). QTLs analysis of frost damage in pea suggests different mechanisms involved in frost tolerance. *Theor. Appl. Genet.* 127, 1319–1330. doi: 10.1007/ s00122-014-2299-6
- Klein, A., Houtin, H., Rond-Coissieux, C., Naudet-Huart, M., Touratier, M., Marget, P., et al. (2020). Meta-analysis of QTL reveals the genetic control of yield-related traits and seed protein content in pea. *Sci. Rep*. 10, 15925. doi: 10.1038/s41598-020-72548-9
- Knüpffer, H., and Van Hintum, T. J. L. (1995). The barley core collection: an international effort. In Hodgkin, T., Brown, A. H. D., Van Hintum, T. J. L., and Morales, E. A. V. (Ed.), *Core collections of plant genetic resources* (pp. 171-178). Wiley & Sons.
- Koirala, S. (2018, April 7). *Characteristics and economic importance of family Papilionaceae (Leguminosae).* Overall Science. https://overallscience.com/characteristics-and-economic-importance-of-family-papilionaceae-leguminosae/
- Köpke, U., and Nemecek, T. (2010). Ecological services of faba bean. *Field Crops Res.*, 115(3), 217-233. DOI: 10.1016/j.fcr.2009.10.012
- Krajewski, P., Bocianowski, J., Gawłowska, M., Kaczmarek, Z., Pniewski, T., Święcicki, W., et al. (2011). QTL for yield components and protein content: a multienvironment study of two pea (*Pisum sativum* L.) populations. *Euphytica* 183, 323-336. doi: 10.1007/s10681-011-0472-4
- Kreplak, J., Madoui, M. A., Cápal, P., Novák, P., Labadie, K., Aubert, G., et al. (2019). A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* 51, 1411-1422. doi: 10.1038/s41588-019-0480-1
- Książkiewicz, M., Rychel, S., Nelson, M. N., Wyrwa, K., Naganowska, B., and Wolko, B. (2016). Expansion of the phosphatidylethanolamine binding protein family in legumes: a case study of *Lupinus angustifolius* L. FLOWERING LOCUS T homologs, LanFTc1 and LanFTc2. *Bmc Genom.* 17, 1-21. doi: 10.1186/s12864-016-3150-z
- Lake, L., Guilioni, L., French, B., and Sadras, V. O. (2021). Field pea. In Sadras, V. O. and Calderini, D. F. (Ed.), *Crop Physiology Case Histories for Major Crops* (pp. 320-341). Academic Press.
- Lamprecht, H. 1974. Monographie der gattung Pisum, Steiermarkische, Landesdruckerei, Graz.
- Lanza, M., Bella, M., Priolo, A., and Fasone, V. (2003). Peas (*Pisum sativum* L.) as an alternative protein source in lamb diets: growth performances, and carcass and meat quality. *Small Rumin. Res.* 47, 63- 68.
- Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* 34, 591–602. doi: 10.1002/gepi.20516
- Lejeune-Hénaut, I., Hanocq, E., Béthencourt, L., Fontaine, V., Delbreil, B., Morin, J., et al. (2008). The flowering locus Hr colocalizes with a major QTL affecting winter frost tolerance in *Pisum sativum* L. *Theor. Appl. Genet.* 116, 1105–1116. doi: 10.1007/s00122-008-0739-x
- Lemontey, C. (1999). Influence du génotype maternel sur les divisions cellulaires dans l'embryon: conséquences pour le potentiel de croissance de la graine de pois. [PhD thesis, Institut National Agronomique de Paris Grignon]. Theses.fr. https://www.theses.fr/1999INAP0011
- Lemontey, C., Mousset-Déclas, C., Munier-Jolain, N., and Boutin, J.P. (2000). Maternal genotype influences pea seed size by controlling both mitotic activity during early embryogenesis and final endoreduplication level/cotyledon cell size in mature seed. *J. Exp. Bot.* 51, 167–175. doi: 10.1093/jexbot/51.343.167
- Lhuillier-Soundele, A., Munier-Jolain, N. G., and Ney, B. (1999). Influence of nitrogen availability on seed nitrogen accumulation in pea. *Crop Sci.* 39, 1741–1748. doi: 10.2135/cropsci1999.3961741x
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

- Li, G., Liu, R., Xu, R., Varshney, R. K., Ding, H., Li, M., et al. (2023). Development of an Agrobacterium-mediated CRISPR/Cas9 system in pea (*Pisum sativum* L.). *Crop J.* 11, 132-139. doi: 10.1016/j.cj.2022.04.011
- LMC International. (2009). *Evaluation of measures applied under the common agricultural policy to the protein crop sector*. https://ec.europa.eu/info/sites/default/files/food-farming-fisheries/key_policies/documents/ext-eval-protein-crops-synth-sum_2009_en.pdf
- Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., et al. (2011). Genomic selection in plant breeding. Knowledge and prospects. *Adv. Agron.* 110, 77–123. doi: 10.1016/B978-0-12-385531-2.00002-5
- Lüscher, A., Mueller-Harvey, I., Soussana, J. F., Rees, R. M., and Peyraud, J. L. (2014). Potential of legume-based grassland–livestock systems in Europe: a review. *Grass Forage Sci.*, 69, 206-228. doi: 10.1111/gfs.12124
- Maia, S. M. F., Ogle, S. M., Cerri, C. E. P., and Cerri, C. C. (2010). Soil organic carbon stock change due to land use activity along the agricultural frontier of the southwestern Amazon, Brazil, between 1970 and 2002. *Glob. Chang. Biol.*, 16, 2775-2788. doi: 10.1111/j.1365-2486.2009.02105.x
- Maqbool, A., Shafiq, S., and Lake, L. (2010). Radiant frost tolerance in pulse crops - A review. *Euphytica* 170, 1-12. doi: 10.1007/s10681-009-0031-4
- Marroni, F., Pinosio, S., Zaina, G., Fogolari, F., Felice, N., Cattonaro, F., et al. (2011). Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genet. Genomes* 7, 1011–1023. doi: 10.1007/s11295-011-0391-5
- Marx, G. A. (1977). Classification, genetics and breeding. In Sutcliffe, J. F., and Pate, J. S. (Ed.), *The physiology of the garden pea* (pp. 21-43). Academic Press.
- Matthews, P., and Arthur, E. (1985). Genetic and environmental components of variation in protein content of peas. In Hebblethwaite, P. D., Heath, M. C., and Dawkins, T. C. K. (Ed.), *The Pea Crop* (pp. 369-381). London: Butterworths.
- McCallum, M. H., Kirkegaard, J. A., Green, T. W., Cresswell, H. P., Davies, S. L., Angus, J. F., et al. (2004). Improved subsoil macroporosity following perennial pastures. *Aust. J. Exp. Agr.* 44, 299-307. doi: 10.1071/EA03076
- 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, 1819–1829. doi: 10.1093/genetics/157.4.1819
- Meyer, D. W., and Badaruddin, M. (2001). Frost tolerance of ten seedling legume species at four growth stages. *Crop Sci.* 41, 1838- 1842. doi: 10.2135/cropsci2001.1838
- Mikić, A., Mihailović, V., Ćupina, B., Kosev, V., Warkentin, T., McPhee, K., et al. (2011). Genetic background and agronomic value of leaf types in pea (*Pisum sativum*). *Ratar. Povrt.* 48, 275–284. doi: 10.5937/ratpov1102275M
- Minderhoud, K. (2010). *Round table for responsible soy association: breaking ground for responsible soy: an institutional response to agricultural expansion and intensification in Argentina* [Master thesis, Utrecht University]. Studenttheses.uu. https://studenttheses.uu.nl/bitstream/handle/20.500.12932/4591/PDFscriptieCAMFinal.pdf?sequence=1&isAllowed=y
- Munier-Jolain, N., and Ney, B. (1998). Seed growth rate in grain legumes. II. Seed growth rate depends on cotyledon cell number. *J. Exp. Bot.* 49, 1971–1976. doi: 10.1093/jxb/49.329.1971
- Munier-Jolain, N., Biarnes, V., and Chaillet, I. (2010). *Physiology of the pea crop* (1st ed.). CRC Press, Boca Raton.
- Murray, G. A., Eser, D., Gusta, L. V., and Eteve, G. (1988). Winterhardiness in pea, lentil, faba bean and chickpea. In Summerfield, R. J. (Ed.), *World crops: Cool season food legumes: a global perspective of the problems and prospects for crop improvement in pea, lentil, faba bean and chickpea* (pp. 831-843). Springer.
- Murray, K. D., and Borevitz, J. O. (2018). Axe: rapid, competitive sequence read demultiplexing using a trie. *Bioinformatics* 34, 3924–3925. doi: 10.1093/ bioinformatics/bty432
- National Aeronautics and Space Administration (NASA). (n.d.). Global climate change. Retrieved November 3, 2023 from https://climate.nasa.gov/faq/19/what-is-the-greenhouse-effect/
- Nazzicari, N., and Biscarini, F. (2017). GROAN: Genomic regression workbench (version 1.0.0). https://cran.r-project.org/package=GROAN
- Nei, M. (1972). Genetic distances between populations. *Am. Nat.* 106, 283–292. doi: 10.1086/282771

- Neill, C., Coe, M. T., Riskin, S. H., Krusche, A. V., Elsenbeer, H., Macedo, M. N., et al. (2013). Watershed responses to Amazon soya bean cropland expansion and intensification. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 368, 20120425. doi: 10.1098/rstb.2012.0425
- Nemecek, T., Richthofen, J. S., Dubois, G., Casta, P., Charles. R., and Pahl. H. (2008). Environmental impacts of introducing grain legumes into European crop rotations. *Eur. J. Agron.* 28, 380- 393. doi: 10.1016/j.eja.2007.11.004
- Nikolopoulou, D., Grigorakis, K., Stasini, M., Alexis, M. N., and Iliadis, K. (2007). Differences in chemical composition of field pea (*Pisum sativum*) cultivars: effects of cultivation area and year. *Food Chem.* 103, 847-852. doi: 10.1016/j.foodchem.2006.09.035
- Oliveira, M., Castro, C., Coutinho, J., and Trindade, H. (2021). Grain legume-based cropping systems can mitigate greenhouse gas emissions from cereal under Mediterranean conditions. *Agric., Ecosyst. Environ.* 313, 107406. doi: 10.1016/j.agee.2021.107406
- Onofri, A. (2019, May 10). *Dealing with correlation in designed field experiments: part II*. Stat for biology. https://www.statforbiology.com/2019/stat_general_correlationindependence2/
- Pachauri, R. K., and Meyer, L. A. (2014). *Climate change 2014: mitigation of climate change. Contribution of working group III to the fifth assessment report of the Intergovernmental Panel on Climate Change.* (Report No. 978-92-9169-143-2). Intergovernmental Panel on Climate Change (IPCC). https://www.ipcc.ch/report/ar5/wg3/
- Park, T., and Casella, G. (2008). The bayesian lasso. *J. Am. Stat. Assoc.* 103, 681-686. doi: 10.1198/016214508000000337
- Pavan, S., Delvento, C., Nazzicari, N., Ferrari, B., D'Agostino, N., Taranto, F., et al. (2022). Merging genotyping-by-sequencing data from two ex situ collections provides insights on the pea evolutionary history. *Hortic. Res.* 9, uhab062. doi: 10.1093/hr/ uhab062
- Pecetti, L., Marcotrigiano, A. R., Russi, L., Romani, M., and Annicchiarico, P. (2019). Adaptation of field pea varieties to organic farming across different environments of Italy. *Crop and Pasture Sci*. 70, 327-333. doi: 10.1071/CP18216
- Piepho, H. P. (2018). Allowing for the structure of a designed experiment when estimating and testing trait correlations. *J. Agric. Sci*. 156, 59-70. doi: 10.1017/S0021859618000059
- Pecetti, L., Annicchiarico, P., Crosta, M., Notario, T., Ferrari, B., and Nazzicari, N. (2023). White lupin drought tolerance: genetic variation, trait genetic architecture, and genome-enabled prediction. *Int. J. Mol. Sci*. 24, 2351. doi: 10.3390/ijms24032351
- Peters, S. O., Sinecen, M., Kizilkaya, K., and Thomas, M. G. (2020). Genomic prediction with different heritability, QTL, and SNP panel scenarios using artificial neural network. *IEEE Access* 8, 147995-148006. doi: 10.1109/ACCESS.2020.3015814
- Piotrowska, A., and Wilczewski, E. (2012). Effects of catch crops cultivated for green manure and mineral nitrogen fertilization on soil enzyme activities and chemical. *Geoderma* 189-190, 72-80. doi: 10.1016/j.geoderma.2012.04.018
- Poggio, S. L., Satorre, E. H., Dethiou, S., and Gonzalo, G. M. (2005). Pod and seed numbers as a function of photothermal quotient during the seed set period of field pea (*Pisum sativum*) crops. *Eur. J. Agron.* 22, 55–69. doi: 10.1016/j.eja.2003.12.003
- Polhill, R. M., and Van der Maesen, L. J. G. (1985). Taxonomy of grain legumes. In Summerfield, R. J. (Ed.), *Grain legume crops* (pp. 3-36). Sheridan House Inc.
- Poore, J., and Nemecek, T. (2018). Reducing food's environmental impacts through producers and consumers. *Science* 360, 987-992. doi: 10.1126/science.aaq0216
- Puritz, J. B., Hollenbeck, C. M., and Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* 2, e431. doi: 10.7717/peerj.431
- Ranalli, P. (1995). Improvement of pulse crops in Europe. *Eur. J. Agron.* 4, 151–166. doi: /10.1016/S1161-0301(14)80042-7
- Raveneau, M. P., Coste, F., Moreau-Valancogne, P., Lejeune-Hénaut, I., and Durr, C. (2011). Pea and bean germination and seedling responses to temperature and water potential. *Seed Sci. Res.* 21, 205- 213. doi: 10.1017/S0960258511000067

- Reckling, M., Döring, T., Stein-Bachinger, K., Bloch, R., and Bachinger, J. (2015). Yield stability of grain legumes in an organically managed monitoring experiment. *Asp. Appl. Biol.* 128, 57-62. doi: 10.13140/RG.2.1.1122.0966
- Ritchie, H. (2019, November 6). *Food production is responsible for one-quarter of the world's greenhouse gas emissions.* Our World in Data. https://ourworldindata.org/food-ghg-emissions
- Ritchie, H. (2020, September 18). *Sector by sector: where do global greenhouse gas emissions come from?* Our World in Data. https://ourworldindata.org/ghg-emissions-by-sector
- Rochester, I. J., Peoples, M. B., Hulugalle, N. R., Gault, R., and Constable, G. A. (2001). Using legumes to enhance nitrogen fertility and improve soil condition in cotton cropping systems. *Field Crops Res.* 70, 27-41. doi: 10.1016/S0378-4290(00)00151-9
- Rubiales, D., Pérez-de-Luque, A., Cubero, J. I., and Sillero, J. C. (2003). Crenate broomrape (*Orobanche crenata*) infection in field pea cultivars. *Crop Prot.* 22, 865-872. doi: 10.1016/S0261-2194(03)00070-X
- Sablowski, R. W., and Meyerowitz, E. M. (1998). A homolog of NO APICAL MERISTEM is an immediate target of the floral homeotic genes APETALA3/PISTILLATA. *Cell* 92, 93-103. doi: 10.1016/S0092-8674(00)80902-2
- Sadras, V.O. (2007). Evolutionary aspects of the trade-off between seed size and number in crops. *Field Crop Res.* 100, 125–138. doi: 10.1016/j.fcr.2006.07.004
- Sadras, V.O., Lake, L., Leonforte, A., McMurray, L.S., and Paull, J.G. (2013). Screening field pea for adaptation to water and heat stress: associations between yield, crop growth rate and seed abortion. *Field Crop Res.* 150, 63–73. doi: 10.1016/j.fcr.2013.05.023
- Sadras, V.O., Lake, L., Kaur, S., and Rosewarne, G. (2019). Phenotypic and genetic analysis of pod wall ratio, phenology and yield components in field pea. *Field Crop Res.* 241, 107551. doi: 10.1016/j.fcr.2019.06.008
- Sagan, M., Ney, B., and Duc, G. (1993). Plant symbiotic mutants as a tool to analyse nitrogen nutrition and yield relationship in field-grown peas (*Pisum sativum* L.). *Plant Soil* 153, 33-45. doi: 10.1007/BF00010542
- Salon, C., Munier-Jolain, N. G, Duc, G., Voisin, A. S., Grandgirard, D., Larmure, A., et al. (2001). Grain legume seed filling in relation to nitrogen acquisition: a review and prospects with particular reference to pea. *Agronomie* 21, 539–552. doi: 10.1051/agro:2001143
- Sandaña, P., and Calderini, D.F. (2012). Comparative assessment of the critical period for grain yield determination of narrow-leafed lupin and pea. *Eur. J. Agron.* 40, 94–101. doi: 10.1016/j.eja.2012.02.009
- Schiltz, S., Munier-Jolain, N., Jeudy, C., Burstin, J., and Salon, C. (2005). Dynamics of exogenous nitrogen partitioning and nitrogen remobilization from vegetative organs in pea revealed by ^{15}N in vivo labeling throughout seed filling. *Plant Physiol.* 137, 1463–1473. doi: 10.1104/pp.104.056713
- Schneider, A. V. C. (2002). Overview of the market and consumption of pulses in Europe. *Br. J. Nutr.* 88, S243-50. doi: 10.1079/BJN2002713
- Searchinger, T., Waite, R., Hanson, C., Ranganathan, J., Dumas, P., Matthews, E., et al. (2019). *Creating a sustainable food future: a menu of solutions to feed nearly 10 billion people by 2050. Final report.* World Resources Institute. https://agritrop.cirad.fr/593176/1/WRR_Food_Full_Report_0.pdf
- Seed Savers Exchange. (n.d.). Catalog. Retrieved January 11, 2024, from https://exchange.seedsavers.org/page/catalog/search/variety?plant-type=PEA&avail=
- Sepngang, B. K., Muel, F., Smadja, T., Stauss, W., Stute, I., Simmen, M., et al. (2020). *Report on legume markets in the EU.* (Report No. D3.1). Fachhochschule Südwestfalen Fachbereich Agrarwirtschaft. https://www.legumehub.eu/wp-content/uploads/2021/06/d31-report-on-legume-markets-in-the-eu.pdf
- Seymour, M., Kirkegaard, J. A., Peoples, M. B., White, P. F., and French, R. J. (2012). Break-crop benefits to wheat in Western Australia—insights from over three decades of research. *Crop Pasture Sci.* 63, 1–16. doi: 10.1071/CP11320
- Siczek, A., Lipiec, J., Wielbo, J., Szarlip, P., and Kidaj, D. (2013). Pea growth and symbiotic activity response to Nod factors (lipo-chitooligosaccharides) and soil compaction. *Appl. Soil Ecol.* 72, 181-186. doi: 10.1016/j.apsoil.2013.06.012

- Siddique, K. H. M., Erskine, W., Hobson, K., Knights, E. J., Leonforte, A., Khan, T. N., et al. (2013). Cool-season grain legume improvement in Australia—use of genetic resources. *Crop Pasture Sci.* 64, 347–360. doi: 10.1071/CP13071
- Sinclair, T. R., and De Witt, C.T. (1975). Photosynthate and nitrogen requirements for seed production by various crops. *Science* 189, 565–567. doi: 10.1126/science.189.4202.565
- Sinclair, T. R., and De Witt, C.T. (1976) Analysis of the carbon and nitrogen limitations of soybean yield. *Agron. J.* 68, 319–324. doi: 10.2134/agronj1976.00021962006800020021x
- Singh, M., Upadhyaya, H. D., and Bisht, I. S. (2013). *Genetic and genomic resources of grain legume improvement* (1st ed.). Elsevier.
- Singh, N., Wu, S., Raupp, W. J., Sehgal, S., Arora, S., Tiwari, V., et al. (2019). Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. *Sci. Rep.* 9, 650. doi: 10.1038/s41598-018-37269-0
- Siol, M., Jacquin, F., Chabert-Martinello, M., Smýkal, P., Le Paslier, M. C., Aubert, G., et al. (2017). Patterns of genetic structure and linkage disequilibrium in a large collection of pea germplasm. *G3-Genes Genom. Genet.* 7, 2461-2471. doi: 10.1534/g3.117.043471
- Smartt, J. (1990). *Grain legumes: evolution and genetic resources* (1st ed.). Cambridge University Press.
- Smýkal, P., Coyne, C. J., Ambrose, M. J., Maxted, N., Schaefer, H., Blair, M. W., et al. (2015). Legume crops phylogeny and genetic diversity for science and breeding. *Crit. Rev. Plant Sci.* 34, 43–104. doi: 10.1080/07352689.2014.897904
- Snoad, B. (1974). A preliminary assessment of 'leafless peas'. *Euphytica* 23, 257-265. doi: 10.1007/BF00035866
- Soetedjo, P., Martin, L., and Janes, A. (1998, July). Canopy architecture, light utilization and productivity of intercrops of field pea and canola. 9th Australian Agronomy Conference 1998. http://www.regional.org.au/au/asa/1998/5/137soetedjo.htm
- Sosulski, F., McLean, L., and Austenson, H. (1974). Management for yield and protein of field peas in Saskatchewan. *Can. J. Plant Sci.* 54, 247–251. doi: 10.4141/cjps74-039
- Soto-Navarro, S. A., Encinias, A. M., Bauer, M. L., Lardy, G. P., and Caton, J. S. (2012). Feeding value of field pea as a protein source in forage-based diets fed to beef cattle. *J. Anim. Sci*. 90, 585-591. doi: 10.2527/jas.2011-4098
- Souer, E., van Houwelingen, A., Kloos, D., Mol, J., and Koes, R. (1996). The no apical meristem gene of Petunia is required for pattern formation in embryos and flowers and is expressed at meristem and primordia boundaries. *Cell* 85, 159-170. doi: 10.1016/S0092-8674(00)81093-4
- Sprent, J. I., Stephens, J. H., and Rupela, O. P. (1988). Environmental effects on nitrogen fixation. In Sumerfield, R. J. (Ed.), *World Crops: Cool Season Food Legumes* (pp. 801–810). Kluwer Academic Publishers.
- Stekhoven, D. J., and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597
- Stelling, D. (1997). Dry peas (*Pisum sativum* L.) grown in mixtures with faba beans (*Vicia faba* L.)—a rewarding cultivation alternative. *J. Agron. Crop Sci.* 179, 65–74. doi: 10.1111/j.1439-037X.1997.tb00500.x
- Stoddard, F. L. (2013). *Agronomic case studies in Legume Futures.* (Report No. 1.2). Legume hub European Union. https://www.legumehub.eu/wp-content/uploads/2021/06/Legume-Futures-Report-1.2.pdf
- Stoddard, F. L., Balko, C., Erskine, W., Khan, H. R., Link, W., and Sarker, A. (2006). Screening techniques and sources of resistance to abiotic stresses in cool-season food legumes. *Euphytica*, 147, 167-186. doi: 10.1007/s10681-006-4723-8
- Strandén, I., Kantanen, J., Russo, I. R. M., Orozco-terWengel, P., Bruford, M. W., and Climgen Consortium. (2019). Genomic selection strategies for breeding adaptation and production in dairy cattle under climate change. *Heredity* 123, 307-317. doi: 10.1038/s41437-019-0207-1
- Sultana, N., Islam, S., Juhasz, A., and Ma, W. (2021). Wheat leaf senescence and its regulatory gene network. *Crop J.* 9, 703-717. doi: 10.1016/j.cj.2021.01.004
- Tafesse, E. G., Warkentin, T. D., and Bueckert, R. A. (2019). Canopy architecture and leaf type as traits of heat resistance in pea. *Field Crop Res.* 241, 107561. doi: 10.1016/j.fcr.2019.107561

- Tan, Y., Hu, F., Chai, Q., Li, G., Coulter, J.A., Zhao, C., et al. (2020). Expanding row ratio with lowered nitrogen fertilization improves system productivity of maize/pea strip intercropping. *Eur. J. Agron.* 113, 125986. doi: 10.1016/j.eja.2019.125986
- Tar'an, B., Warkentin, T., Somers, D. J., Miranda, D., Vandenberg, A., Blade, S., et al. (2004). Identification of quantitative trait loci for grain yield, seed protein concentration and maturity in field pea (*Pisum sativum* L.). *Euphytica* 136, 297-306. doi: 10.1023/B:EUPH.0000032721.03075.a0
- Taranto, F., Nicolia, A., Pavan, S., De Vita, P., and D'Agostino, N. (2018). Biotechnological and digital revolution for climate-smart plant breeding. *Agronomy* 8, 277. doi: 10.3390/agronomy8120277
- Tayeh, N., Klein, A., Le Paslier, M. C., Jacquin, F., Houtin, H., Rond, C., et al. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Front. Plant Sci.* 6, 941. doi: 10.3389/fpls.2015.00941
- Terres Inovia. (2017). Guide de culture du pois. Terres Inovia. https://www.terresinovia.fr/-/telecharger-le-guide-pois
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., B: Stat. Methodol.* 58, 267-288. doi: 10.1111/j.2517-6161.1996.tb02080.x
- Turner, S. R, Barrat, D. H. P., and Casey, R. (1990). The effect of different alleles at the *r* locus on the synthesis of seed storage proteins in *Pisum sativum. Plant Mol. Biol.* 14, 793–803. doi: 10.1007/BF00016512
- Turner, N. C., Wright, G. C., and Siddique, K. H. M. (2001). Adaptation of grain legumes (pulses) to water-limited environments. *Adv. Agron.* 71, 194– 233. doi:10.1016/S0065-2113(01)71015-2
- United Nations. (2015, November 12). https://unfccc.int/sites/default/files/english_paris_agreement.pdf
- Urbatzka, P., Graß, R., Haase, T., Schüler, C., Trautz, D., and Heß, J. (2011). Grain yield and quality characteristics of different genotypes of winter pea in comparison to spring pea for organic farming in pure and mixed stands. *Org. Agric.* 1, 187-202. doi: 10.1007/s13165-011-0015-2
- Uzun, A., Bilgili, U., Sincik, M., Filya, I., and Acikgoz, E. (2005). Yield and quality of forage type pea lines of contrasting leaf types. *Eur. J. Agron.* 22, 85–94. doi: 10.1016/j.eja.2004.01.001
- Vaidyanathan, L., Sylvester-Bradley, R., Bloom, T., and Murray, A. (1987). Effects of previous cropping and applied nitrogen on grain nitrogen content in winter wheat. *Asp. Appl. Biol.* 15, 227–237.
- Vanderschuren, H., Chatukuta, P., Weigel, D., and Mehta, D. (2023). A new chance for genome editing in Europe. *Nat. Biotechnol*. 1-3. doi: 10.1038/s41587-023-01969-4
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *JDS* 91, 4414-4423. doi: 10.3168/jds.2007-0980
- Vocanson, A., and Jeuffroy, M. H. (2007). Agronomic performance of different pea cultivars under various sowing periods and contrasting soil structures. *Agron. J.* 3, 748–759. doi: 10.2134/agronj2005.0301
- Voisin, A. S., Salon, C., Munier-Jolain, N. G., and Ney, B. (2002). Effect of mineral nitrogen on nitrogen nutrition and biomass partitioning between the shoot and roots of pea (*Pisum sativum* L.). *Plant Soil* 242, 251–262. doi: 10.1023/A:1016214223900
- Voisin, A. S., Salon, C., Jeudy, C., and Warembourg, F. R. (2003a). Root and nodule growth in *Pisum sativum* L. in relation to photosynthesis: analysis using ^{13}C labelling. *Ann Bot.* 92, 1–7. doi: 10.1093/aob/mcg174
- Voisin, A. S., Salon, C., Jeudy, C., and Warembourg, F. R. (2003b) Symbiotic N_2 fixation in relation to C economy of *Pisum sativum* L. as function of plant phenology. *J. Exp. Bot.* 54, 2733–2744. doi: 10.1093/jxb/erg290
- Volpelli, L. A., Comellini, M., Masoero, F., Moschini, M., Lo Fiego, D. P., and Scipioni, R. (2009). Pea (*Pisum sativum*) in dairy cow diet: effect on milk production and quality. *Ital. J. Anim. Sci.* 8, 245-257. doi: 10.4081/ijas.2009.245
- Wang, T. L., and Hedley, C. L. (1985). Genetic and developmental analysis of the seed. In Casey, R. and Davies, D. R. (Ed.), *Peas: Genetics, Molecular Biology and Biotechnology* (pp. 83-120). CAB International.
- Wang, H., Misztal, I., Aguilar, I., Legarra, A., and Muir, W. M. (2012). Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res.* 94, 73-83. doi: 10.1017/S0016672312000274

- Wang, M., and Xu, S. (2019). Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity* 123, 287-306. doi: 10.1038/s41437-019-0205-3
- Wang, J., and Zhang, Z. (2021). GAPIT version 3: boosting power and accuracy for genomic association and prediction. *GPB* 19, 629–640. doi: 10.1016/j.gpb.2021.08.005
- Wearn, O. R., Reuman, D. C., and Ewers, R. M. (2012). Extinction debt and windows of conservation opportunity in the Brazilian Amazon. *Science* 337, 228-232. doi: 10.1126/science.1219013
- Weber, H., Borisjuk, L., and Wobus, U. (1996). Controlling seed development and seed size in *Vicia faba*: a role for seed coat-associated invertases and carbohydrate state. *Plant J.* 10, 823–834. doi: 10.1046/j.1365-313X.1996.10050823.x
- Weeden, N. F. (2007). Genetic changes accompanying the domestication of *Pisum sativum*: is there a common genetic basis to the 'domestication syndrome' for legumes? *Ann. Bot.* 100, 1017–1025. doi: 10.1093/aob/mcm122
- Wientjes, Y. C., Calus, M. P., Goddard, M. E., and Hayes, B. J. (2015). Impact of QTL properties on the accuracy of multi-breed genomic prediction. *Genet. Sel. Evol.* 47, 1-16. doi: 10.1186/s12711-015-0124-6
- World Trade Organization (WTO). (1947). General Agreement on Tariffs and Trade. https://www.wto.org/english/docs_e/legal_e/gatt47_e.pdf
- Yadav, S. S., McNeil, D. L., Redden, R., and Patil, S. A. (2010). *Climate change and management of cool season grain legume crops* (1st ed.). Springer.
- Yang, T., Liu, R., Luo, Y., Hu, S., Wang, D., Wang, C., et al. (2022). Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. *Nat. Genet.* 54, 1553-1563. doi: 10.1038/s41588-022-01172-2
- Yao, J., Zhao, D., Chen, X., Zhang, Y., and Wang, J. (2018). Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (*Triticum aestivum* L.). *Crop J.* 6, 353-365. doi: 10.1016/j.cj.2018.05.003
- Yendle, P. W., and MacFie, H. J. (1989). Discriminant principal components analysis. *J. Chemom.* 3, 589–600. doi: 10.1002/cem.1180030407
- Yin, T., Pimentel, E. C. G., Borstel, U. K. V., and König, S. (2014). Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature× humidity-dependent covariate. *J. Dairy Sci*. 97, 2444-2454. doi: 10.3168/jds.2013-7143
- Zander, P., Amjath-Babu, T. S., Preissel, S., Reckling, M., Bues, A., Schläfke, N., et al. (2016). Grain legume decline and potential recovery in European agriculture: a review. *Agron. Sustain. Dev.*, 36, 1-20. doi: 10.1007/s13593-016-0365-y
- Zohary, D., and Hopf, M. (1973). Domestication of pulses in the Old World: legumes were companions of wheat and barley when agriculture began in the Near East. *Science* 182, 887–894. doi: 10.1126/science.182.4115.887

8. Appendix

Figure 1. Quantile-Quantile plots of expected vs. observed association scores of 18,674 SNPs for two pea traits and 276 pea lines belonging to three connected RIL populations. The red line represents equality between the expected and observed quantiles and the grey area the associated 95% confidence interval. The two upper plots refer to mean trait data across three environments, while the lower two to data from two single environments.

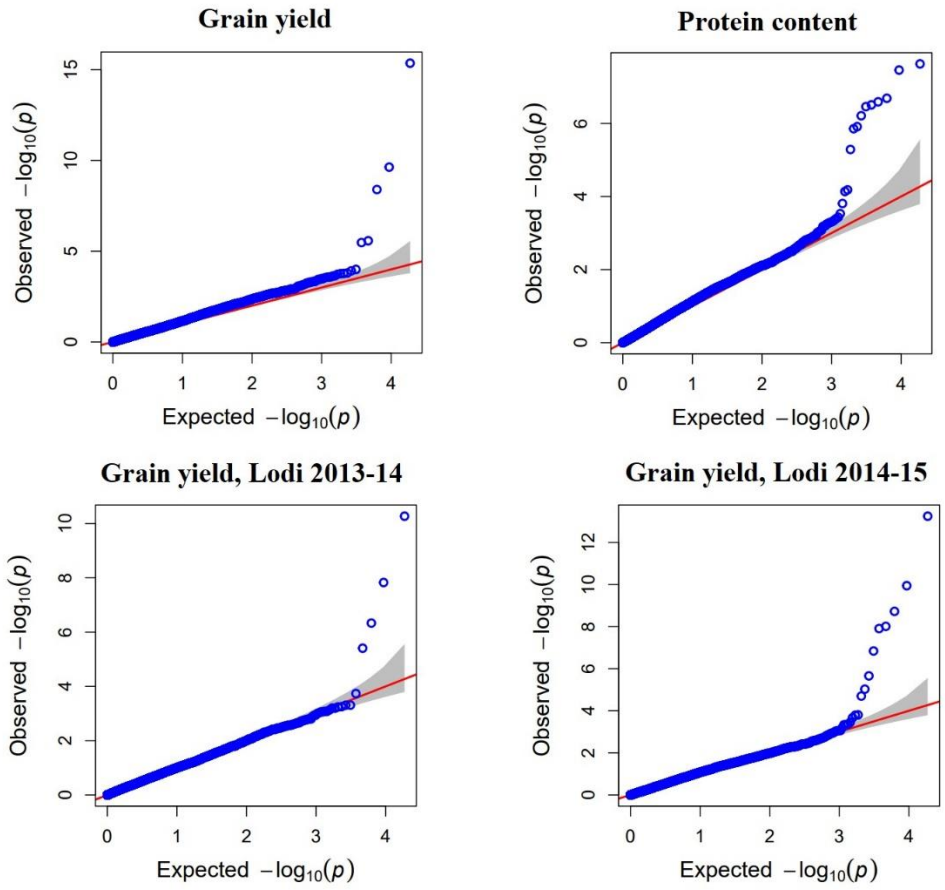


Table 1. Trait mean value in three test environments of 306 pea inbred lines belonging to three connected RIL populations. Row means followed by different letters differ at $p < 0.05$.

Trait	Lodi 2013-2014	Lodi 2014-2015	Perugia 2013-2014	SE
Grain yield (t/ha)	6.31 a	4.59 b	2.90 c	0.35
Protein content (%)	25.32 a	23.22 c	24.26 b	0.15
Protein yield (t/ha)	1.60 a	1.07 b	0.70 c	0.09

Table 2. Climate, soil, and long-term climate characteristics of three pea test environments according to FAO (2006) Guidelines for soil description, 4th. Rome: Food and Agricultural Organization.

Item	Lodi 2013-14	Lodi 2014-15	Perugia 2013-14	Lodi long-term	Perugia long-term
Crop management system	Organic	Conventional	Organic	-	-
Rainfall, Jan.-Mar. (mm)	343	198	280	161	177
Rainfall, Apr.-May (mm)	122	147	179	154	142
Absolute minimum daily T (°C)	–5.7	–11.6	–3.6	–7.7	–5.0
Mean of max. daily T, May (°C)	23.2	23.9	23.4	21.8	23.0
Soil texture	Silt-loam	Sandy-loam	Silty-clay-loam	-	-
Soil pH	7.9	6.3	7.6	-	-

Table 3. Components of variance relative to genotype (S^2_G), genotype × environment interaction (S^2_{GE}), RIL population (S^2_R), genotype within RIL population ($S^2_{G(R)}$), RIL population × environment interaction (S^2_{RE}), and genotype within RIL population × environment interaction ($S^2_{G(R)E}$) for three traits in three test environments of 306 pea lines belonging to three connected RIL populations. All variance components were significantly different from zero at $p < 0.01$.

Trait	Without RIL population			With RIL population			
	S^2_G	S^2_{GE}	S^2_G / S^2_{GE}	S^2_R	$S^2_{G(R)}$	S^2_{RE}	$S^2_{G(R)E}$
Grain yield (t/ha)	0.575	1.435	0.401	0.080	0.520	1.121	0.693
Protein content (%)	0.724	0.302	2.393	0.131	0.637	0.199	0.167
Protein yield (t/ha)	0.036	0.085	0.422	0.003	0.034	0.068	0.040

Table 4. Mean value of parental lines of three connected RIL populations and cultivar Spacial for three pea traits in three test environments. Row means followed by different letters differ at $p < 0.05$.

Trait	Environment	Mean value			
		Attika	**Kaspa**	**Isard**	**Spacial**
Yield (t/ha)	Lodi 2013-2014	4.97 b	7.16 a	6.36 ab	7.82 a
	Lodi 2014-2015	1.34 c	2.14 c	6.39 a	3.66 b
	Perugia 2013-2014	2.13 c	3.34 ab	2.59 bc	3.70 a
Protein content (%)	Lodi 2013-2014	23.68 c	26.82 a	24.30 bc	24.87 b
	Lodi 2014-2015	22.71 b	23.54 a	21.90 c	21.68 c
	Perugia 2013-2014	22.89 c	26.08 a	23.04 c	24.16 b
Protein yield (t/ha)	Lodi 2013-2014	1.18 c	1.92 a	1.54 b	1.95 a
	Lodi 2014-2015	0.30 c	0.50 bc	1.40 a	0.80 b
	Perugia 2013-2014	0.49 b	0.87 a	0.60 b	0.90 a

Figure 2. Plots of linkage disequilibrium (r^2) decay with physical distance for pea chromosomes. r^2 was estimated on pairwise combinations of 18,674 SNPs within a 100 kb window for 276 pea lines belonging to three connected RIL populations.

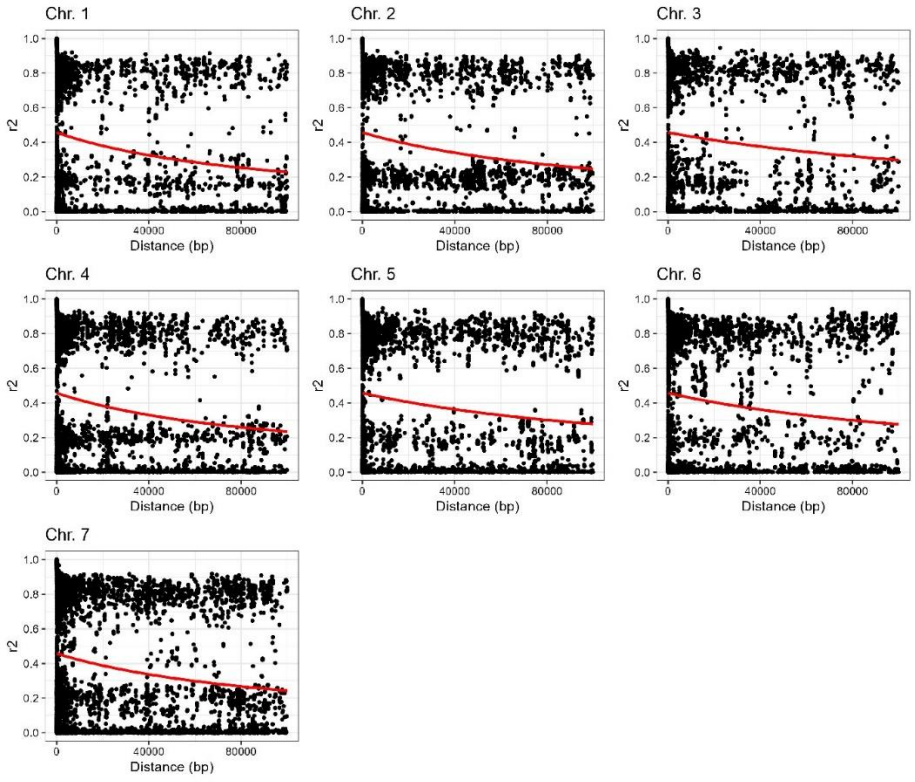


Table 5. Significant markers detected by a GWAS based on 18,674 SNPs and 276 pea lines belonging to three connected RIL populations for two traits averaged across three test environments and one trait in two single environments, with the relative estimated effect.

SNP	Trait	Effect
chr1LG6_167158777	Grain yield; grain yield, Lodi 2014-15	0.29
chr3LG5_110943532	Grain yield	0.36
chr6LG2_112957117	Grain yield	0.30
chr5LG3_555982018	Grain yield; grain yield, Lodi 2014-15	0.33
chr6LG2_72049750	Grain yield	0.26
chr5LG3_548298174	Protein content	0.33
scaffold00731_102978	Protein content	0.10
chr1LG6_14126991	Protein content	0.30
chr3LG5_122898115	Protein content	0.36
chr5LG3_238119406	Protein content	0.36
chr4LG4_7076997	Protein content	0.33
chr5LG3_112688292	Protein content	0.37
chr1LG6_55881393	Protein content	0.14
chr2LG1_23689490	Protein content	0.34
chr3LG5_13152476	Protein content	0.30
chr6LG2_75499720	Grain yield, Lodi 2013-14	0.33
chr2LG1_293191007	Grain yield, Lodi 2013-14	0.35
chr6LG2_252910779	Grain yield, Lodi 2013-14	0.28
chr3LG5_302412301	Grain yield, Lodi 2013-14	0.07
chr3LG5_213572246	Grain yield, Lodi 2014-15	0.36
chr6LG2_78688716	Grain yield, Lodi 2014-15	0.30
chr5LG3_207358795	Grain yield, Lodi 2014-15	0.28
chr1LG6_195757998	Grain yield, Lodi 2014-15	0.35
chr2LG1_383235559	Grain yield, Lodi 2014-15	0.36
chr3LG5_109454037	Grain yield, Lodi 2014-15	0.36

Table 6. List of genes potentially associated to the significant SNPs detected by a GWAS based on 18,674 SNPs and 276 lines belonging to three connected RIL populations for two pea traits averaged across three test environments. Candidate genes were identified by scanning a 100 kb region in both directions from each significant SNP and are reported with their annotated function (https://urgi.versailles.inra.fr/).

Significant_SNP	Trait	Candidate_Gene	Function
chr1LG6_167158777	Grain yield	Psat1g096760	Phosphatidylethanolamine-binding protein
chr1LG6_167158777	Grain yield	Psat1g096800	TatD related DNase
chr1LG6_167158777	Grain yield	Psat1g096840	Major intrinsic protein
chr1LG6_167158777	Grain yield	Psat1g096880	GDA1/CD39 (nucleoside phosphatase) family
chr1LG6_167158777	Grain yield	Psat1g096920	Unknown gene
chr3LG5_110943532	Grain yield	Psat3g051800	Utp11 protein
chr3LG5_110943532	Grain yield	Psat3g051840	Zinc finger + C3HC4 RING-type
chr3LG5_110943532	Grain yield	Psat3g051880	RING-variant domain
chr3LG5_110943532	Grain yield	Psat3g051920	TPR repeat region circular profile
chr3LG5_110943532	Grain yield	Psat3g051960	PfkB family carbohydrate kinase
chr3LG5_110943532	Grain yield	Psat3g052000	Unknown gene
chr6LG2_112957117	Grain yield	Psat6g081160	Unknown gene

chr6LG2_112957117	Grain yield	Psat6g081200	Zinc-binding in reverse transcriptase
chr6LG2_112957117	Grain yield	Psat6g081240	Glutathione S-transferase + N-terminal domain
chr6LG2_112957117	Grain yield	Psat6g081280	Unknown gene
chr5LG3_555982018	Grain yield	Psat5g289640	Electron transfer flavoprotein-ubiquinone oxidoreductase + 4Fe-4S
chr5LG3_555982018	Grain yield	Psat5g289680	Regulation of cellular nucleobase + nucleoside + nucleotide and nucleic acid metabolic process
chr5LG3_555982018	Grain yield	Psat5g289720	Metallo-beta-lactamase superfamily
chr5LG3_555982018	Grain yield	Psat5g289760	BZIP transcription factor
chr6LG2_72049750	Grain yield	Psat6g064240	WD domain + G-beta repeat
chr6LG2_72049750	Grain yield	Psat6g064280	Protein of unknown function (DUF861)
chr6LG2_72049750	Grain yield	Psat6g064320	Beta-ketoacyl synthase + C-terminal domain
chr6LG2_72049750	Grain yield	Psat6g064360	Beta-ketoacyl synthase + N-terminal domain
chr6LG2_72049750	Grain yield	Psat6g064400	Zinc-binding dehydrogenase
chr6LG2_72049750	Grain yield	Psat6g064440	TPR repeat region circular profile
chr6LG2_72049750	Grain yield	Psat6g064480	Unknown gene
chr6LG2_72049750	Grain yield	Psat6g064520	ABC transporter
chr5LG3_548298174	Protein content	Psat5g282440	UAA transporter family
chr5LG3_548298174	Protein content	Psat5g282480	Homeobox' domain profile
chr5LG3_548298174	Protein content	Psat5g282520	Unknown gene
chr5LG3_548298174	Protein content	Psat5g282600	3 +4-dihydroxy-2-butanone 4-phosphate synthase
chr1LG6_14126991	Protein content	Psat1g010880	Transferase activity + transferring phosphorus-containing groups
chr1LG6_14126991	Protein content	Psat1g010920	Cellular nucleobase + nucleoside + nucleotide and nucleic acid metabolic process
chr1LG6_14126991	Protein content	Psat1g010960	Serine/cysteine peptidase + trypsin-like
chr1LG6_14126991	Protein content	Psat1g011000	Serine/cysteine peptidase + trypsin-like
chr3LG5_122898115	Protein content	Psat3g058360	Unknown gene
chr3LG5_122898115	Protein content	Psat3g058400	AP2 domain
chr3LG5_122898115	Protein content	Psat3g058440	Ring finger domain
chr3LG5_122898115	Protein content	Psat3g058480	Autophagy protein Apg9
chr3LG5_122898115	Protein content	Psat3g058520	Unknown gene
chr3LG5_122898115	Protein content	Psat3g058560	Myc-type + basic helix-loop-helix (bHLH) domain profile
chr3LG5_122898115	Protein content	Psat3g058600	HR-like lesion-inducing
chr5LG3_238119406	Protein content	Psat5g132320	LysM domain
chr5LG3_238119406	Protein content	Psat5g132360	Transcription factor Tfb4
chr5LG3_238119406	Protein content	Psat5g132400	Transcription factor Tfb4
chr5LG3_238119406	Protein content	Psat5g132440	Histone chaperone domain CHZ
chr4LG4_7076997	Protein content	Psat4g006600	Galactosyltransferase
chr4LG4_7076997	Protein content	Psat4g006640	Unknown gene
chr4LG4_7076997	Protein content	Psat4g006680	Methyltransferase TYW3
chr5LG3_112688292	Protein content	Psat5g062600	NYN domain
chr5LG3_112688292	Protein content	Psat5g062640	EamA-like transporter family
chr5LG3_112688292	Protein content	Psat5g062680	Unknown gene
chr5LG3_112688292	Protein content	Psat5g062720	Triose-phosphate Transporter family
chr5LG3_112688292	Protein content	Psat5g062760	ABC transporter
chr5LG3_112688292	Protein content	Psat5g062800	Cyanobacterial and plant NDH-1 subunit O
chr1LG6_55881393	Protein content	NA	NA
chr2LG1_23689490	Protein content	Psat2g022120	FAM91 N-terminus
chr2LG1_23689490	Protein content	Psat2g022160	Leucine rich repeat N-terminal domain
chr2LG1_23689490	Protein content	Psat2g022240	Clathrin adaptor complex small chain
chr2LG1_23689490	Protein content	Psat2g022280	Intracellular membrane-bounded organelle

chr2LG1_23689490	Protein content	Psat2g022320	Ethylene insensitive 3
chr3LG5_13152476	Protein content	Psat3g004240	BTB/POZ domain
chr3LG5_13152476	Protein content	Psat3g004280	Unknown gene
chr3LG5_13152476	Protein content	Psat3g004320	BTB/POZ domain
chr3LG5_13152476	Protein content	Psat3g004360	BTB And C-terminal Kelch
chr3LG5_13152476	Protein content	Psat3g004400	BTB/POZ domain
chr3LG5_13152476	Protein content	Psat3g004440	BTB/POZ domain

Table 7. Information regarding a worldwide pea germplasm collection including 220 landraces from 19 regional pools and 11 modern cultivars.

Accession name	Germplasm pool	Geographic area	Germplasm type	Donor institution
IG116297	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG51989	India	India	Landrace/old cultivar	ICARDA
IG112140	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG50673	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG123136	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG50311	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG128863	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG51529	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG51994	India	India	Landrace/old cultivar	ICARDA
IG49633	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG52455	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG52596	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG52459	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG52036	China	China	Landrace/old cultivar	ICARDA
IG52442	Western Asia	Western Asia	Landrace/old cultivar	ICARDA
IG49610	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG123006	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG123242	China	China	Landrace/old cultivar	ICARDA
IG50756	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG50358	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG50250	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG125543	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG52383	China	China	Landrace/old cultivar	ICARDA
IG134619	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG134828	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG123313	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG50235	India	India	Landrace/old cultivar	ICARDA
IG115331	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG123237	China	China	Landrace/old cultivar	ICARDA
IG52535	Western Asia	Western Asia	Landrace/old cultivar	ICARDA
IG50559	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG52456	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG52426	Western Asia	Western Asia	Landrace/old cultivar	ICARDA
IG123244	China	China	Landrace/old cultivar	ICARDA
IG125600	Central Asia	Central Asia	Landrace/old cultivar	ICARDA

IG116232	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG51948	India	India	Landrace/old cultivar	ICARDA
IG134080	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG49176	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG123248	China	China	Landrace/old cultivar	ICARDA
IG49544	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG51576	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG123080	Western Asia	Western Asia	Landrace/old cultivar	ICARDA
IG51927	India	India	Landrace/old cultivar	ICARDA
IG134772	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG51687	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG115341	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG50570	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG114914	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG50584	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG123245	China	China	Landrace/old cultivar	ICARDA
IG52586	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG134788	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG125597	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG122966	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG134718	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG52496	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG115145	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG123118	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG52040	China	China	Landrace/old cultivar	ICARDA
IG123041	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG49181	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG52417	Western Asia	Western Asia	Landrace/old cultivar	ICARDA
IG51957	India	India	Landrace/old cultivar	ICARDA
IG114899	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG134746	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG115114	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG125324	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG50303	India	India	Landrace/old cultivar	ICARDA
IG50641	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG128856	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG125471	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG134109	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG49189	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG124857	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG51520	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG134862	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG115100	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG125326	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG114977	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG124664	China	China	Landrace/old cultivar	ICARDA
IG52050	India	India	Landrace/old cultivar	ICARDA

IG122974	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG122996	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG51976	India	India	Landrace/old cultivar	ICARDA
IG115266	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG50362	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG134744	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG123312	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG134094	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG134841	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG50669	China	China	Landrace/old cultivar	ICARDA
IG115228	Nepal	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG128973	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG125415	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG125550	China	China	Landrace/old cultivar	ICARDA
IG125336	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG125472	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG123073	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG134823	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG123240	China	China	Landrace/old cultivar	ICARDA
IG125589	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG134707	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG129002	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG52534	Western Asia	Western Asia	Landrace/old cultivar	ICARDA
IG123311	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG123050	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG134060	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG134609	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG123227	China	China	Landrace/old cultivar	ICARDA
IG51513	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG52521	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG50935	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG50357	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG123211	China	China	Landrace/old cultivar	ICARDA
IG123028	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG51891	India	India	Landrace/old cultivar	ICARDA
IG51562	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG123034	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG128913	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG123029	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG134770	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG128934	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG123004	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
IG134857	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG134782	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG123280	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG123021	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG51991	India	India	Landrace/old cultivar	ICARDA

IG51551	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG128887	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG134621	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG52401	Western Asia	Western Asia	Landrace/old cultivar	ICARDA
IG52081	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG51688	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG125378	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG128983	Russia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG51536	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG123288	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG51516	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG50248	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG52005	India	India	Landrace/old cultivar	ICARDA
IG124843	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG52017	India	India	Landrace/old cultivar	ICARDA
IG123102	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG123281	Central Asia	Central Asia	Landrace/old cultivar	ICARDA
IG50592	China	China	Landrace/old cultivar	ICARDA
IG52092	Ethiopia	Ethiopia	Landrace/old cultivar	ICARDA
IG51993	India	India	Landrace/old cultivar	ICARDA
IG134649	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG49224	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG125469	Georgia	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG52367	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG134870	Balkans	Central Europe	Landrace/old cultivar	ICARDA
IG116455	Turkey	Western Asia	Landrace/old cultivar	ICARDA
IG49327	Greece	Southern Europe	Landrace/old cultivar	ICARDA
IG134750	Ukraine	Ukraine, Georgia, Russia	Landrace/old cultivar	ICARDA
IG49203	Afghanistan	Afghanistan, Nepal	Landrace/old cultivar	ICARDA
IG52595	Maghreb	Maghreb	Landrace/old cultivar	ICARDA
MG 100948	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 101126	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 106069	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 106871	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 110243	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 110416	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 110417	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 110418	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 111850	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 111988	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
MG 115084	Italy	Southern Europe	Landrace/old cultivar	CNR-IBBR, Bari
ZP0064	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP0076	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP0126	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP0181	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP0202	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP0213	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid

ZP0535	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP0798	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP0799	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP1261	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP1264	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP1282	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP1294	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
ZP1300	Spain	Southern Europe	Landrace/old cultivar	ITACyL Valladolid
Witham Wonder	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Emerald Gem	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Englishsabel	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Fillbasket	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
English Wonder	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Kentish Invicta	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Alderman	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Mummy Pea	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Magnum Bonum	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Raina Victoria	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Knights Marrow	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Knights Dwarf White	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
English Maple	United Kingdom	Western Europe	Landrace/old cultivar	JIC Norwich
Gloire de Correze	France	Western Europe	Landrace/old cultivar	INRA Dijon
Serpette D'Auvergne	France	Western Europe	Landrace/old cultivar	INRA Dijon
Picar	France	Western Europe	Landrace/old cultivar	INRA Dijon
Quarante-deux de Sarcelles	France	Western Europe	Landrace/old cultivar	INRA Dijon
Piver	France	Western Europe	Landrace/old cultivar	INRA Dijon
Triomphe de Roissy	France	Western Europe	Landrace/old cultivar	INRA Dijon
Chemin Long	France	Western Europe	Landrace/old cultivar	INRA Dijon
Serpette de Vitry	France	Western Europe	Landrace/old cultivar	INRA Dijon
Gris de Bourgogne	France	Western Europe	Landrace/old cultivar	INRA Dijon
Haute Loire	France	Western Europe	Landrace/old cultivar	INRA Dijon
Champagne	France	Western Europe	Landrace/old cultivar	INRA Dijon
Cote D'Or	France	Western Europe	Landrace/old cultivar	INRA Dijon
Serpette de Paris	France	Western Europe	Landrace/old cultivar	INRA Dijon
CL 19cvs1	Central Europe	Central Europe	Landrace/old cultivar	CRI, Praha
Hrach Z Pardubic	Central Europe	Central Europe	Landrace/old cultivar	CRI, Praha
Kocovska 108	Central Europe	Central Europe	Landrace/old cultivar	CRI, Praha
Pulawska Slodka Nr 2	Central Europe	Central Europe	Landrace/old cultivar	IPK Gatersleben
Kapucin Belokvety	Central Europe	Central Europe	Landrace/old cultivar	CRI, Praha
Landrace Orava	Central Europe	Central Europe	Landrace/old cultivar	CRI, Praha
PIS 278	Central Europe	Central Europe	Landrace/old cultivar	IPK Gatersleben
PIS 657	Central Europe	Central Europe	Landrace/old cultivar	IPK Gatersleben
PIS 845	Central Europe	Central Europe	Landrace/old cultivar	IPK Gatersleben
PIS 2856	Central Europe	Central Europe	Landrace/old cultivar	IPK Gatersleben
Attika	Improved Variety	Improved Variety	Improved Variety	-
Genial	Improved Variety	Improved Variety	Improved Variety	-
Messire	Improved Variety	Improved Variety	Improved Variety	-

Santana	Improved Variety	Improved Variety	Improved Variety	-
Spirale	Improved Variety	Improved Variety	Improved Variety	-
Cartuce	Improved Variety	Improved Variety	Improved Variety	-
Dove	Improved Variety	Improved Variety	Improved Variety	-
Enduro	Improved Variety	Improved Variety	Improved Variety	-
Isard	Improved Variety	Improved Variety	Improved Variety	-
Viriato	Improved Variety	Improved Variety	Improved Variety	-
Cigarron	Improved Variety	Improved Variety	Improved Variety	-

Figure 3. Quantile-Quantile plots of expected vs. observed association scores of 41,114 SNPs for two pea traits and 223 accessions from a worldwide germplasm collection. The red line represents equality between the expected and observed quantiles and the grey area the associated 95% confidence interval.

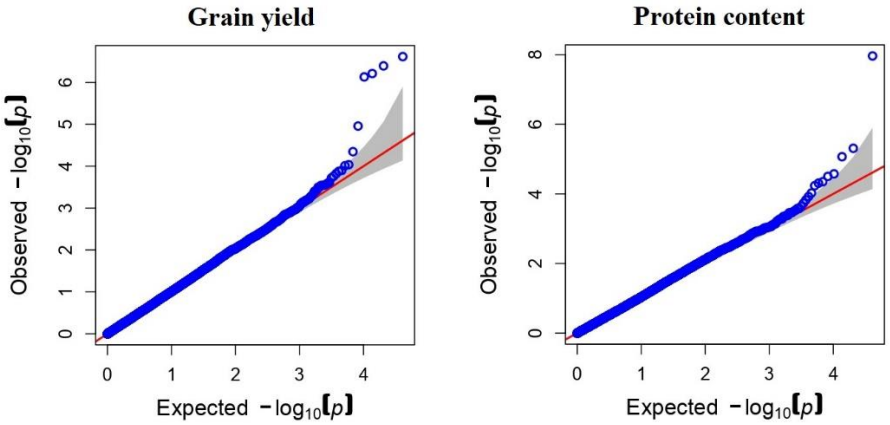


Figure 4. Plots of linkage disequilibrium (r^2) decay with physical distance for pea chromosomes. r^2 was estimated on pairwise combinations of 41,114 SNPs within a 100 kb window for 212 landraces from 19 regional pools and 11 modern cultivars from a worldwide germplasm collection.

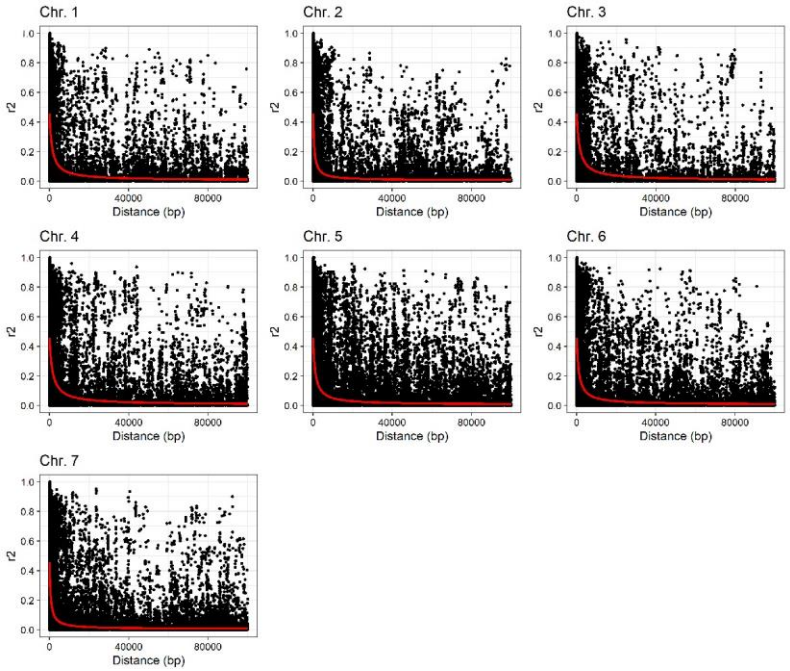


Table 8. Significant markers detected by a GWAS based on 41,114 SNPs and performed on 212 landraces from 19 regional pools and 11 modern cultivars from a worldwide germplasm collection for two pea traits. The SNPs are presented from the most to the least significant for each trait with the relative estimated effect.

SNP	Trait	Effect
chr6LG2_72901872	Grain yield	0.32
chr1LG6_47025851	Grain yield	0.22
chr7LG7_183744462	Grain yield	-0.27
chr4LG4_186752146	Grain yield	-0.34
chr5LG3_492526140	Protein content	0.95

Table 9. List of genes potentially associated to the significant SNPs detected by a GWAS based on 41,114 SNPs and performed on 212 landraces from 19 regional pools and 11 modern cultivars from a worldwide germplasm collection for two pea traits. Candidate genes were identified by scanning a region long as the average chromosome distance at which LD dropped to 0.05 in both directions from each significant SNP and are reported with their annotated function (https://urgi.versailles.inra.fr/).

SNP	Trait	Gene	Function
chr6LG2_72901872	Grain yield	Psat6g064800	Helix-loop-helix DNA-binding domain
chr1LG6_47025851	Grain yield	Psat1g031400	Protein kinase domain
chr4LG4_186752146	Grain yield	Psat4g098400	RNA recognition motif. (a.k.a. RRM + RBD + or RNP domain)
chr7LG7_183744462	Grain yield	Psat7g111400	No apical meristem (NAM) protein
chr5LG3_492526140	Protein content	Psat5g246720	Rhomboid family

Table 10. Lines selected from each of six RIL populations, which were represented by the first two letters of genotype names, for two pea traits by PS and GS.

Grain yield		Protein yield	
PS	GS	PS	GS
AG_L162	AG_L35	AG_L162	AG_L35
AG_L48	AG_L194	AG_L176	AG_L33
AG_L176	AG_L81	AI_L126	AI_L117
AI_L210	AI_L104	AI_L210	AI_L104
AI_L264	AI_L117	CI_L121	CI_L29
AI_L126	AI_L212	CI_L208	CI_L88
CI_L208	CI_L88	DA_L47	DA_L147
CI_L85	CI_L73	DA_L50	DA_L220
CI_L88	CI_L29	KA_L122	KA_L105
DA_L46	DA_L115	KA_L203	KA_L258
DA_L47	DA_L2	KI_L198	KI_L61
DA_L50	DA_L147	KI_L61	KI_L166
KA_L122	KA_L258		
KA_L134	KA_L175		
KA_L203	KA_L105		
KI_L198	KI_L61		
KI_L61	KI_L166		
KI_L140	KI_L110		

Table 11. Climatic variables characterizing GS training, PS, and test experiments of top-performing lines for grain and protein yield according to PS and GS, including rainfall in the period from January to May, and minimum daily temperature (Min. daily T) and number of frost days (No. frost days), namely days with a measured minimum temperature below -0.5°C, during the whole cropping cycle.

Item	Sowing date	Rainfall (mm)	Min. daily T (°C)	No. frost days
GS training experiments				
Lodi 2013-14	07/11/2013	465	–5.7	35
Perugia 2013-14	25/11/2013	459	–3.6	9
Lodi 2014-15	22/10/2014	345	–11.6	34
PS experiments				
Lodi 2018-19	25/10/2018	308	–12.0	54
Lodi 2019-20	10/12/2019	192	–10.9	34
Test experiments				
Lodi 2022	01/02/2022	119	–3.3	13
Lodi 2023	07/02/2023	209	–4.0	7
Perugia 2022-23	01/12/2022	320	–4.1	12

Table 12. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and grain yield of three lines selected for this trait by either GS or PS from each of six RIL populations (Sel = selection type; Env = environment; RIL = RIL population; Rep = replicate; Geno = genotype; ":" indicates an interaction between factors, while "/"nesting of the second factor within the first one).

Grain yield (t/ha)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Sel	1	19.29	19.29	20.59	9.55E-06
GS_Set	1	41.43	41.43	44.23	2.49E-10
Env	2	216.53	108.26	115.57	1.42E-34
Sel:GS_Set	1	0.65	0.65	0.69	0.407
Sel:Env	2	2.32	1.16	1.24	0.293
GS_Set:Env	2	19.24	9.62	10.27	5.55E-05
GS_Set/RIL	4	14.33	3.58	3.82	0.005
Env/Rep	6	61.11	10.18	10.87	1.56E-10
Sel:RIL	4	3.04	0.76	0.81	0.518
Env:RIL	8	23.19	2.90	3.09	0.003
Sel:GS_Set:Env	2	3.01	1.50	1.60	0.204
Sel:GS_Set/Geno	24	44.17	1.84	1.96	0.006
Sel:Env:RIL	8	7.81	0.98	1.04	0.406
Env:Sel:GS_Set/Geno	48	85.14	1.77	1.89	0.001
Residuals	210	196.73	0.94		

Table 13. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and protein yield of two lines selected for this trait by either GS or PS from each of six RIL populations (Sel = selection type; Env = environment; RIL = RIL population; Rep = replicate; Geno = genotype; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Protein yield (t/ha)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Sel	1	0.55	0.55	8.18	0.005
GS_set	1	1.24	1.24	18.50	3.20E-05
Env	2	6.72	3.36	50.23	4.07E-17
Sel:GS_set	1	0.02	0.02	0.35	0.556
Sel:Env	2	0.05	0.03	0.39	0.677
GS_set:Env	2	0.79	0.40	5.92	0.003
GS_set/RIL	4	0.94	0.23	3.51	0.009
Env/Rep	6	2.36	0.39	5.87	1.78E-05
Sel:RIL	4	0.44	0.11	1.66	0.164
Env:RIL	8	0.97	0.12	1.81	0.080
Sel:GS_set:env	2	0.05	0.02	0.35	0.703
Sel:GS_set:geno	12	1.23	0.10	1.53	0.120
Sel:Env:RIL	8	0.57	0.07	1.06	0.397
Sel:GS_set:Env:geno	24	3.44	0.14	2.14	0.003
Residuals	138	9.24	0.07		

Table 14. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and grain yield of two lines selected for protein yield by either GS or PS from each of six RIL populations (Sel = selection type; Env = environment; RIL = RIL population; Rep = replicate; Geno = genotype; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Grain yield (t/ha)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Sel	1	8.64	8.64	7.87	0.0057
GS_set	1	16.93	16.93	15.44	0.0001
Env	2	146.13	73.06	66.62	5.63E-21
Sel:GS_set	1	0.26	0.26	0.23	0.6300
Sel:Env	2	0.93	0.47	0.42	0.6550
GS_set:Env	2	11.31	5.65	5.16	0.0069
GS_set/RIL	4	12.47	3.12	2.84	0.0265
Env/Rep	6	38.64	6.44	5.87	1.76E-05
Sel:RIL	4	6.49	1.62	1.48	0.2118
Env:RIL	8	15.87	1.98	1.81	0.0803
Sel:GS_set:env	2	1.62	0.81	0.74	0.4803
Sel:GS_set:geno	12	20.32	1.69	1.54	0.1155
Sel:Env:RIL	8	9.43	1.18	1.07	0.3843
Sel:GS_set:Env:geno	24	58.48	2.44	2.22	0.0022
Residuals	138	151.35	1.10		

Table 15. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and protein content of two lines selected for protein yield by either GS or PS from each of six RIL populations (Sel = selection type; Env = environment; RIL = RIL population; Rep = replicate; Geno = genotype; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Protein content (%)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Sel	1	0.88	0.88	2.92	0.090
GS_set	1	12.47	12.47	41.24	2.02E-09
Env	2	216.64	108.32	358.35	2.27E-55
Sel:GS_set	1	2.39	2.39	7.91	0.006
Sel:Env	2	3.88	1.94	6.42	0.002
GS_set:Env	2	11.11	5.56	18.38	8.36E-08
GS_set/RIL	4	26.16	6.54	21.64	7.11E-14
Env/Rep	6	16.83	2.81	9.28	1.53E-08
Sel:RIL	4	3.67	0.92	3.03	0.020
Env:RIL	8	18.82	2.35	7.78	1.37E-08
Sel:GS_set:env	2	8.14	4.07	13.47	4.53E-06
Sel:GS_set:geno	12	23.59	1.97	6.50	4.18E-09
Sel:Env:RIL	8	11.62	1.45	4.81	3.11E-05
Sel:GS_set:Env:geno	24	27.42	1.14	3.78	3.80E-07
Residuals	138	41.71	0.30		

Table 16. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and grain yield of three lines selected for this trait by either GS or PS from each of six RIL populations and the relative parental lines (Group = genotype group, represented by PS and GS top-performing genotypes whose RIL populations of origin were included or not in the GS training set, and the relative parental lines; Env = environment; Geno = genotype; Rep = replicate; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Grain yield (t/ha)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Group	4	70.33	17.58	17.09	2.21E-12
Env	2	208.23	111.13	108.02	2.13E-34
Group:Env	8	32.61	4.08	3.96	0.0002
Group/Geno	37	71.45	1.93	1.88	0.0027
Env/Rep	6	72.61	12.10	11.76	1.37E-11
Env:Group/Geno	74	126.24	1.71	1.66	0.0022
Residuals	246	253.09	1.03		

Table 17. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and onset of flowering data of three lines selected for grain yield by either GS or PS from each of six RIL populations and the relative parental lines (Group = genotype group, represented by PS and GS top-performing genotypes whose RIL populations of origin were included or not in the GS training set, and the relative parental lines; Env = environment; Geno = genotype; Rep = replicate; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Flowering time (days since March 1)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Group	4	829.80	207.45	152.01	2.88E-65
Env	2	23619.28	11809.64	8653.70	1.07E-228
Group:Env	8	55.53	6.94	5.09	7.20E-06
Group/Geno	37	3242.25	87.63	64.21	6.28E-106
Env/Rep	6	51.62	8.60	6.30	3.50E-06
Env:Group/Geno	74	505.64	6.83	5.01	7.37E-22
Residuals	246	335.71	1.36		

Table 18. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and farmer acceptability score data of three lines selected for grain yield by either GS or PS from each of six RIL populations and the relative parental lines (Group = genotype group, represented by PS and GS top-performing genotypes whose RIL populations of origin were included or not in the GS training set and the relative parental lines; Env = environment; Geno = genotype; Rep = replicate; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Farmer score (1-9)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Group	4	28.07	7.02	10.24	1.10E-07
Env	2	237.96	118.98	173.60	9.59E-48
Group:Env	8	6.65	0.83	1.21	2.92E-01
Group/Geno	37	146.67	3.96	5.78	5.88E-18
Env/Rep	6	51.94	8.66	12.63	2.05E-12
Env:Group/Geno	74	111.42	1.51	2.20	3.54E-06
Residuals	246	168.60	0.69		

Table 19. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and protein yield data of two lines selected for this trait by either GS or PS from each of six RIL populations and the relative parental lines (Group = genotype group, represented by PS and GS top-performing genotypes whose RIL populations of origin were included or not in the GS training set and the relative parental lines; Env = environment; Geno = genotype; Rep = replicate; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Protein yield (t/ha)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Group	4	2.31	0.58	7.90	7.14E-06
Env	2	6.92	3.46	47.38	3.74E-17
Group:Env	8	1.32	0.16	2.26	0.026
Group/Geno	25	3.30	0.13	1.81	0.015
Env/Rep	6	2.78	0.46	6.35	4.66E-06
Env:Group/Geno	50	5.63	0.11	1.54	0.022
Residuals	174	12.70	0.07		

Table 20. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and grain yield data of two lines selected for protein yield by either GS or PS from each of six RIL populations and the relative parental lines (Group = genotype group, represented by PS and GS top-performing genotypes whose RIL populations of origin were included or not in the GS training set and the relative parental lines; Env = environment; Geno = genotype; Rep = replicate; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Grain yield (t/ha)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Group	4	33.13	8.28	6.91	3.50E-05
Env	2	153.36	76.68	63.93	1.53E-21
Group:Env	8	20.41	2.55	2.13	0.036
Group/Geno	25	49.19	1.97	1.64	0.035
Env/Rep	6	49.15	8.19	6.83	1.60E-06
Env:Group/Geno	50	93.89	1.88	1.57	1.83E-02
Residuals	174	208.71	1.20		

Table 21. Degrees of freedom (Df), type III sum of squares (SS), mean squares (MS), and F and p value (P) from an ANOVA model based on the factors listed in the 1st column and protein content data of two lines selected for protein yield by either GS or PS from each of six RIL populations and the relative parental lines (Group = genotype group, represented by PS and GS top-performing genotypes whose RIL populations of origin were included or not in the GS training set and the relative parental lines; Env = environment; Geno = genotype; Rep = replicate; ":" indicates an interaction between factors, while "/" the nesting of the second factor within the first one).

Protein content (%)					
Factor	**Df**	**SS**	**MS**	**F value**	**P**
Group	4	18.36	4.59	12.77	3.89E-09
Env	2	279.89	139.95	389.41	5.67E-65
Group:Env	8	23.66	2.96	8.23	2.00E-09
Group/Geno	25	71.55	2.86	7.96	4.03E-18
Env/Rep	6	15.24	2.54	7.07	9.47E-07
Env:Group/Geno	50	77.23	1.54	4.30	4.63E-13
Residuals	174	62.53	0.36		