Genotyping and phenotyping the common pea and its wild relatives

Kirstie Fiona HETHERINGTON

A thesis submitted for the degree of Doctor of Philosophy

at the

University of East Anglia
Earlham Institute

March, 2020

Abstract

Genotyping and phenotyping the common pea and its wild relatives

by Kirstie Fiona HETHERINGTON

In 1868, three men, Gregor Mendel, Charles Darwin and Friedrich Miescher made significant contributions in genetic inheritance, plant domestication and DNA extraction respectively. 150 years later, this thesis aims to better understand pea domestication through genotyping and phenotyping the common pea and its wild relatives. Peas (*Pisum sativum*) are a cool season legume important to food security due to their ability to fix nitrogen and produce nutritious food and animal fodder. A core collection of 350 accessions of wild, landrace and cultivated material was developed from the John Innes *Pisum* Collection. To characterise these accessions, image analysis, a modern phenotyping method was used. Current tools require user expertise, are not cross platform, are not applicable to certain plants or phenotypes. Here, MktStall, a novel multi-organ image analysis is presented, which requires no computational expertise. Pea is a large (4.5Gb) and highly repetitive genome. Here, the first publicly accessible pea genome reference is announced. In combination with a genotyping by sequencing (GBS) approach of this core collection a genome-wide association study (GWAS) was performed using on seed weight, plant height, leaflet margin, seed shape and pod shape. The results in this thesis show statistically significant differences in plant height in cultivars and leaflet length, perimeter and area in landraces in addition to identifying statistically significant loci for leaflet teeth, seed perimeter and seed eccentricity. Furthermore, potential candidate genes have be identified with roles in carbohydrate metabolism known to cause seed wrinkledness and POWERDRESS known to increase leaf area. The combination of novel contributions results in new tools, genomic resources and additional knowledge of pea domestication which can be used in marker assisted selection and improved breeding practices for an important crop for food security.

Contents

List of Figures

List of Tables

List of Abbreviations

ATFC	Australian Temperate Field Crop
BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
BNF	Biological Nitrogen Fixation
BP	Biological Process
BUSCO	Benchmarking Universal Single-Copy Orthologs
CSV	Comma-Separated Values
DAPC	Discriminant Analysis of Principal Components
DNA	Deoxyribonucleic Acid
EFD	Elliptical Fourier Descriptors
FDR	False Discovery Rate
FTP	File Transfer Protocol
G2P	Genotype-to-Phenotype
GAPIT	Genome Association and Prediction Integrated Tool
GBS	Genotyping by Sequencing
GEBV	Genomic Estimated Breeding Values
GH	Glasshouse
GLM	Generalised linear model
GO	Gene Ontology
GUI	Graphical User Interface
GWAS	Genome-Wide Association Study
$\mathbf{H_0}$	Null Hypothesis
HSV	Hue Saturation Value
IA	Image Analysis
INRA	Institut National de la Recherche Agronomique
JIC	John Innes Centre
JZ	Ji Zhou
K-S test	Kolmogorov-Smirnov test
LJG	Laura-Jayne Gardiner
LF	Leighton Folkes
LY	Luis Yanes
LD	Linkage Disequilibrium
LDA	Linear Discriminate Analysis
MAF	Minor Allele Frequency
MAS	Marker-Assisted Selection
MF	Molecular Function
MLM	mixed linear model
NCBI	National Center for Biotechnology Information

NGS	Next-Generation Sequencing
OCR	Optical Character Recognition
OLC	Overlap-Layout-Consensus
PCR	Polymerase Chain Reaction
PGS	Pea Genome Sequencing
QTL	Quantative Trait Loci
PE	Paired-end
Q-Q	Quantile-Quantile plots
RAD-seq	Restriction site Associated DNA sequencing
RBIP	Retrotransposon-Based Insertion Polymorphism
RGB	Red, Green and Blue
RNS	Root Nodule Symbiosis
SBS	Sequencing by Synthesis
SE	Single-end
SNF	Symbiotic Nitrogen Fixation
SNP	Single Nucleotide Polymorphism
SMSDs	Simple and Morphological Shape Descriptors
TAIR	The Arabidopsis Information Portal
USDA	United States Department of Agriculture
VCF	Variant Call Format
W2RAP	WGS (Wheat) Robust Assembly Pipeline
WGS	Whole Genome Shotgun

Acknowledgements

The author would like to thank Prof Neil Hall for the supervision of this project.

The author acknowledges exchanges with the remainder of the supervisory panel: Dr Sarah Ayling, Dr Rob Davey, Prof Anthony Hall, Mr Mike Ambrose, Prof Clare Domoney and Prof Enrico Coen, and members of their laboratories. The author also expresses gratitude to Dr Jose de Vega for useful discussion.

The author would like to thank Dr Laura-Jayne Gardiner for the provision of a script and the contributors of the tool "MktStall": Dr Leighton Folkes, Dr Ji Zhou and Mr Luis Yanes. Particular appreciation goes to Dr Leighton Folkes for the continual support and encouragement provided.

This project was funded by the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership grant number BB/J014524/1 and in part by BBSRC, Core Strategic Programme Grant BB/CSP17270/1 at the Earlham Institute.

Chapter 1

Introduction

1.1 Mendel

Gregor Johann Mendel, was an Austrian monk considered to be the founding father of genetics. He investigated dozens of organisms, however, it was his 8-year study in peas that revolutionised the face of science as we know it today. His seminal work entitled, "Experiments in Plant Hybridization" was read out in 1865 (Mendel, 1996) and it was this work that subsequently led to the Principles of Genetics, or Mendel's postulates once rediscovered (Miko, 2008).

Pea as a model

Peas were a particularly good model for Mendel's research. Mendel chose to work with peas because after studying many different types of legume, he found pea to meet the following expectations the best out of all the legumes (Mendel, 1996). Firstly, Mendel required a plant with observable phenotypes and although he was well-resourced with a team of eight other scientists, phenotyping is laborious manual process and he required the phenotypes to be easily distinguishable by eye. Secondly, Mendel required plants that could be easily grown. Peas have a short life cycle and grow equally well in field or in pots (Mendel, 1996).

Mendel's Legacy

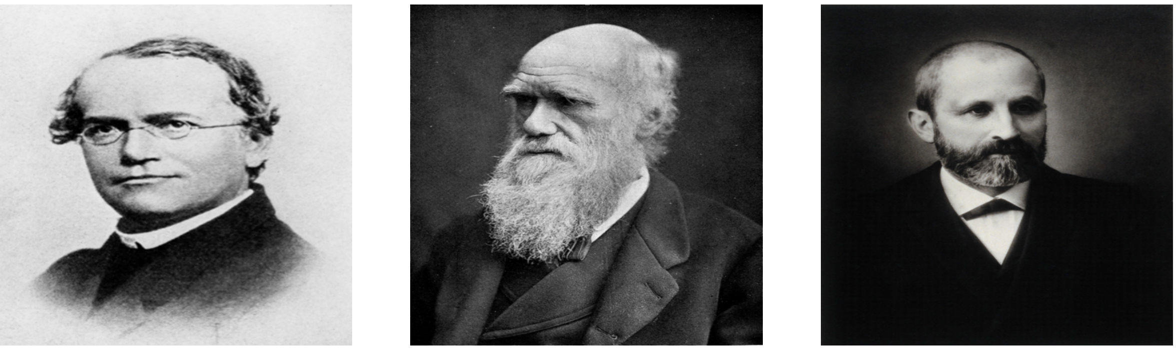
Mendel's work was largely ignored at the time, however, as time goes by, history has been much more favourable to him. Mendel's main legacy was his documentation of robust experimental design to produce rigorous results with strong statistical power.

He performed many repeats to ensure the statistical robustness of his results and by observing all members of a generation he also ensured a good sample size. Mendel points out how important sample size is within scientific studies as with low sample size there can be large variation (Mendel, 1996). Mendel always used the most vigorous plant for phenotyping because the sub-viable ones lead to errors and he discarded sub-optimal plants that would produce outliers or erroneous data. Mendel was careful to avoid any discrepancies in his data through strong experimental design. Mendel chose to work with phenotypes that were easily qualified and distinct (Mendel, 1996). This means that the results were easily qualified and this removes a lot of uncertainty because having eight scientists visually assess less easily described phenotypes may have produced differing results based on subjectivity. He ensured the full development of seeds by keeping the pods on the plant until ripened and dry which is vital for assessing seed shape and colour (Mendel, 1996).

This combination of diligence to statistical power and experimental design, combined with his ability to deduce the Laws of Inheritance from his work, convey that Mendel was a shrewd scientist and provided an exemplary precedent of implementing robust research methods in biological research. His work has since been validated by generations of geneticists and this may demonstrate that the aforementioned unwelcome reception of his seminal work was unwarranted.

150 years since 1868

In order to understand the significance of the outlined work, one must consider that the year of 1868 was a crucial year for science, although unrecognised at the time.



(A) Gregor Mendel
(1822-1884)
Laws of Inheritance

(B) Charles Darwin
(1809-1882)
Natural Selection

(C) Friedrich Miescher
(1844-1895)
Isolating Nucleic Acid

FIGURE 1.1: **Mendel, Darwin and Miescher.** Three key scientists whose scientific contribution form the framework for the investigations outlined in this thesis. Images obtained from the Wellcome Collection and shared under the Creative Commons Attribution 4.0 International Licence, `https://creativecommons.org/licenses/by/4.0/` therefore appear with permissions. Mendel appears with cropping (*Wellcome Collection - Portrait of Gregor Johann Mendel, Garrison.*). Darwin (*Wellcome Collection - Portrait of Charles Darwin.*) and Miescher (*Wellcome Collection - Johann Friedrich Miescher. Photograph*) remain unchanged.

In 1868, three years after the dissemination of his seminal work, read out in Berlin, Mendel was discouraged by how his work lacked impact and continued to focus his efforts in his work as a priest and subsequently becomes elected to abbot (Dunn, 2003).

Charles Darwin is most famous for his work in putting forward the hypothesis of evolution through natural selection (Darwin, 1859). However, at around about the same time, Darwin wrote his book, "The Variation of Animals and Plants under Domestication" (Darwin, 1868) completely unaware of Mendel's contribution. In this book he puts forward a new theory to describe heredity, **pangenesis**. Pangenesis has long been considered to be incorrect (Galton, 1871). The key feature here is that Mendel did not continue to disseminate his work and therefore, Darwin and his wide circle of scientific supporters did not hear of his principles. If Darwin been aware of Mendel's principles and supported them, our knowledge of inheritance could have been expedited, especially given Darwin's contribution to the theory of Natural Selection through Evolution. Instead, Britain must wait for William Bateson, former director of the John Innes Centre, to revive Mendel's ideas in 1900 at a Royal Horticultural Society meeting (Bateson, 2004), alongside other key scientists doing similarly in other countries like Hugo de Vries, Carl Correns and Erich von Tschermak (Olby, 1987).

Towards the end of 1868, Friedrich Miescher begins work in the laboratory, working on looking at the nucleus of white blood cells found in the pus of bandages obtained from the local hospital. Little did he know that this work would allow him to be the first to isolate nucleic acid. He carefully managed to wash the white blood cells off of the bandages and was able to extract what he called "nuclein" (Miescher-Rüsch, 1871). This was the first documented DNA extraction, which is no mean feat considering centrifuges were not available in laboratory settings at the time (Elzen, 1986). It is only later that Miescher begins to understand the consequences of his discovery and begins to consider that "nuclein" might be the material that carries heredity (Dahm, 2008).

In summary, 1868 was a ground-breaking year for science, unbeknownst to scientists of the time. The impact of the work of these three scientists was not fully comprehended at the time but their work has formed a strong foundation for modern day science through the understanding of the heritability of traits.

1.2 Nitrogen Fixation

Legumes, such as pea, are capable of fixing nitrogen. Nitrogen is an important organic element found in many vital molecules; it is a key part of amino acids which are used to make protein. Nitrogen is the most abundant gas found in air which comprises of 78% nitrogen (Miller, 1954). In order to understand the important role of nitrogen in legumes, it is apt to first explore the role of nitrogen in the world around us through the nitrogen cycle.

The Nitrogen Cycle

The nitrogen cycle is the recycling of nitrogen into different chemical forms (Canfield, Glazer, and Falkowski, 2010). Despite being abundant in air, it cannot be used by plants in its common diatomic form, N_2, due to the triple covalent bond which requires a considerable amount of energy to break (Canfield, Glazer, and Falkowski, 2010). The energy required to break this triple covalent bond is 941 $kJmol^{-1}$ (*Chemistry LibreTexts* 2017). Therefore, for plants, the nitrogen source must come from the soil in the form of nitrates so it can be assimilated by the plant (Venkateshwaran and Ané, 2011).

The nitrogen cycle is composed of several parts. The **nitrification stage**, the passing of the nitrogen compounds up the food chain to the next trophic level and the **decomposition** of these nitrogen compounds from excretion products and when these organisms die.

There are several ways in which nitrogen can be fixed in the nitrification stage - by natural sources such as lightning, biological nitrogen fixation by legumes and decomposition by **denitrifying bacteria** or by artificial sources such as the Haber-Bosch process, a reaction that makes fertiliser using vast amounts of energy which often comes from burning fossil fuels (Wagner, 2011).

The importance of Biological Nitrogen Fixation

Biological nitrogen fixation (BNF) - sometimes referred to as Symbiotic Nitrogen fixation (SNF) - is an important process for reducing the need for fertiliser (Abi-Ghanem et al., 2013) due to the conversion of N_2 (Venkateshwaran and Ané, 2011). Here, the link between the nitrogen cycle, the carbon cycle, fossil fuels and food shortage within the context of biological nitrogen fixation is explored.

One key benefit of using BNF is that it is a way to reduce fertiliser usage and produce food simultaneously (Gresshoff et al., 2015). Fertilisers have large consequences on the environment, firstly, because the Haber-Bosch process used to create fertilisers is often powered by fossil fuels (Menge, Wolf, and Funk, 2015), and, secondly, excessive use of fertilisers can leach out of the soil and into the water supply which can lead to eutrophication (Menge, Wolf, and Funk, 2015).

Furthermore, with the increase in intensive agriculture, the soil becomes nitrogen deficient and soil quality reduces (Tilman et al., 2002; Abawi and Widmer, 2000). Legumes can help prevent this, in a natural, cost effective and sustainable manner (Venkateshwaran and Ané, 2011). Using legumes, in crop rotation or through intercropping strategies can reduce the cost of fertilising the soil (Venkateshwaran and Ané, 2011), provide a low impact pest and disease management strategy (Abawi and Widmer, 2000) and reduce weed growth when grown as a companion legume in intercropping strategies (Liebman and Dyck, 1993).

Nitrogen is fixed using the following chemical equation (Venkateshwaran and Ané, 2011):

$$\mathrm{N_2} + 8\,\mathrm{e^-} + 16\,\mathrm{MgATP} + 8\,\mathrm{H^+} \xrightarrow{\text{nitrogenase}} 2\,\mathrm{NH_3} + \mathrm{H_2} + 16\,\mathrm{MgADP} + 16\,\mathrm{P}_i$$

This equation is catalysed by the enzyme, nitrogenase. Legumes have species specificity for their symbiotic bacteria to ensure the host only uptakes the correct bacteria and in pea, the bacterial symbiont is *Rhizobium leguminosarum biovar viceae* (Wagner, 2011). *Rhizobium leguminosarum biovar viceae* have basic nitrogenase assembly machinery (Venkateshwaran and Ané, 2011) and have 8 *nif* genes responsible for the assembly and synthesis of this key enzyme (Venkateshwaran and Ané, 2011).

Rhizobia are endosymbionts and cause the plant to form nodules so the *Rhizobium* can colonize intracellularly. *Rhizobium* converts dinitrogen (N_2) to $NH_4{}^+$ which is a more assimilable form. In return, the host reduces its natural plant defence against its symbiotic bacteria and produces Leghaemoglobin to ensure optimal nitrogen fixing conditions (Venkateshwaran and Ané, 2011; Abi-Ghanem et al., 2013). This is because *Rhizobium* is sensitive to oxygen and is aerobic but the nitrogenase enzyme does not function well in high oxygen concentrations. Leghaemoglobin is made by the plant to ensure these conditions are met so the nitrogenase enzyme can function correctly (Appleby, 1984; Venkateshwaran and Ané, 2011). It works in a similar way to human haemoglobin, allowing sufficient oxygen for the bacteria to fix nitrogen but not too much to inhibit the nitrogenase enzyme (Wagner, 2011).

The key mechanisms of species specificity, reducing its natural defence, and producing of leghaemoglobin, suggest legumes have evolutionarily adapted to host nitrogen-fixing bacteria (Venkateshwaran and Ané, 2011).

Nodule organogenesis

The formation of root nodules on the plant root is called **nodulation organogenesis** and this occurs either by Nod factor dependent strategies or Nod factor independent strategies (Venkateshwaran and Ané, 2011).

The Nod factor dependent strategy, relies on **isoflavonoids** which are a secondary metabolite present in pea used as a chemoattractant for nitrogen-fixing bacteria to colonize the nodule (Wagner, 2011; Leigh, Signer, and Walker, 1985; Venkateshwaran and Ané, 2011).

The Nod factor independent strategy is thought to be triggered by **cytokinins**. An example would be the use of Rhizobial exopolysaccharides (EPS) to help infect the plant by suppressing plant defence responses (Venkateshwaran and Ané, 2011).

Therefore, these two different strategies for nodulation form the part of the evolutionarily adaption to nitrogen fixation in legumes.

1.3 *Pisum sativum*

Peas (*Pisum sativum*) are cool season legumes which are economically considered the second most common crop family (Smýkal et al., 2012). Peas provide a nutritious food source in the human diet and plays a vital role in animal fodder. They are a rich and diverse source of nutrients, particularly starch and protein, as well as vitamins and minerals, in addition to other metabolites such as isoflavonoids (Smýkal et al., 2012) - well known for their anti-cancer properties (Perabo et al., 2008).

Peas are thought to have originally have been domesticated around the Mediterranean and Middle East when early humans were first involved in selective breeding of the peas in 9-10,000 BC, and are therefore considered one of the oldest domesticated crops (Zohary and Hopf, 1973). However, Henry, Brooks, and Piperno (2010) investigated the plaque found in the teeth of Neanderthals and deduced that either pea or chickpea was consumed in their diet, so pea has potential to have been foraged far earlier than this date. Peas are thought to have originated in the Fertile Crescent (Brown et al., 2009; Smỳkal et al., 2015), which is defined as the area of valleys and hills around the rivers Jordan, Tigris and Euphrates (Brown et al., 2009) and this would include countries such as Israel, Palestine, Jordan, Lebanon, Syria, Iraq and Turkey (Zair, Maxted, and Amri, 2018). Wild varieties are thought to have originated from across Asia, Northern Africa and Southern Europe (Smýkal et al., 2012). Several species comprise the *Pisum* genus: *Pisum sativum* which is the common domesticated pea, *Pisum abysinnicum* which is an independently domesticated cultivar from Yemen and Ethiopia (Smýkal et al., 2012; Jing et al., 2010; Maxted and Ambrose, 2001; Westphal, 1974), *Pisum fulvum* which is a wild species from Israel, Jordan, Syria and Lebanon (Jing et al., 2010; Smỳkal et al., 2011) and the wild species *Pisum elatius* from the Mediterrean Basin (Jing et al., 2010; Zohary, Hopf, and Weiss, 2012; Smartt, 1984).

The pea in agriculture

The United Nations Food and Agricultural Organisation have a statistics database, the Food and Agricultural Organisation Corporate Statistical Database (FAOSTAT), and this holds agricultural data for more than 250 countries from 1961. The author has mined this database to compare pea with other agriculturally relevant legumes - soybean and lentil - and with other agriculturally relevant crops: wheat, rice and oats. The author has compared these for yield (defined as production per hectare for the area that is cultivated) (*Food and Agricultural Organisation Statistical Databases of the United Nations*), for area (which is defined as the area cultivated in hectares, in other words, total sown area) (*Food and Agricultural Organisation Statistical Databases of the United Nations*) and for production quantity in tonnes (which is defined as the harvested production) (*Food and Agricultural Organisation Statistical Databases of the United Nations*). Here, these factors are compared for these agriculturally important crops for the World + (Total) region over 56 years of accessible data (1961-2017).

Figure 1.2 displays yield over a 56-year period, rice and wheat have the greatest yield (the greatest production per hectare of area) in comparison to soybean, oats, peas and lentils. Figure 1.2 shows that rice and wheat are high-yielding crops, producing more plant material per area sown. For all other crops except pea, the yield trend lines are closely correlated to the actual yield and follow a slight increase. The yield trendline for pea is a large increase over the 56-year period with the actual yield line highly fluctuating above and below this trendline. This may indicate that pea is highly susceptible to the environmental variables caused by climate change. Also, for all crops, irrespective of change, all crops have some form of increase in yield over the 56-year period. This may be due to the improvement in our understanding of agronomy and agriculture: farmers maybe taking better care of soil quality (Rasmussen, 1999), plant health may have improved with the increased use of pesticides and improved understanding of host-plant resistance (Widawsky et al., 1998) or more productive harvesting techniques may have developed over time (Ruysschaert et al., 2004).

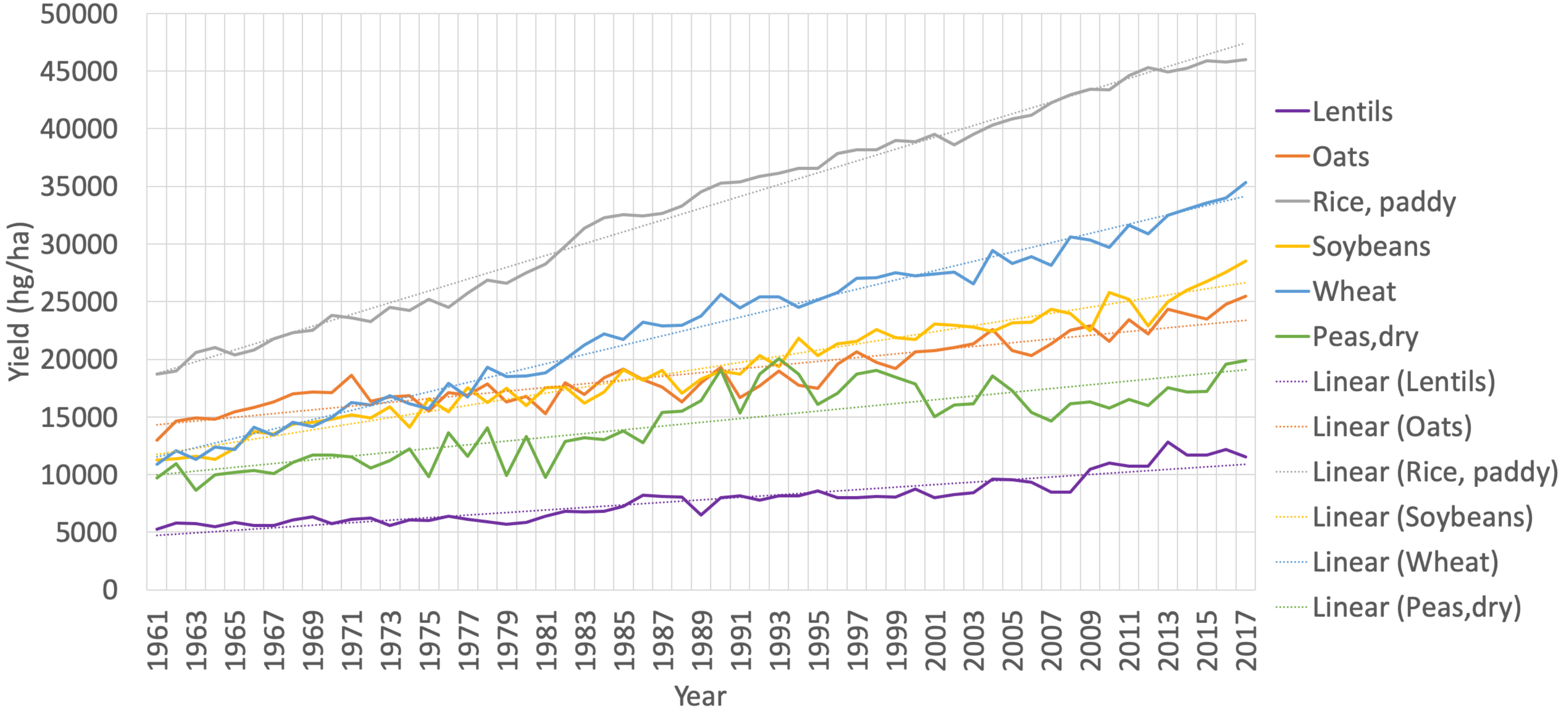


FIGURE 1.2: **Yield for three grain legumes and for three grain cereals (wheat, rice and oat) over five decades (1961-2017).** Legumes plotted: pea (green), soybean (yellow) and lentil (purple). Cereals plotted: wheat (blue), rice (grey) and oats (orange). Trendlines for each are shown as dotted lines. Data obtained from *Food and Agricultural Organisation Statistical Databases of the United Nations*. Original in colour.

Figure 1.3 shows that rice, wheat and soybean are the most produced crops with an increase in production over time. Oats, peas and lentils are produced in much lower quantities. For the area harvested, Figure 1.4 shows that wheat, rice and soybean have a greater area of land dedicated to cultivating these crops. Again, oats, lentils and peas have far less area dedicated to these crops world-wide. Interestingly, in both graphs, oats have an unusual rate of decline, starting in 1961 as higher than soybean in production quantity and area harvested but as less area is dedicated to oats in 1971, the production quantity of oats decreased in 1972. As the years go on, as less area is dedicated to oats this correlates with decreased production quantity, and since then oat is the only crop of the six where production has steadily declined.

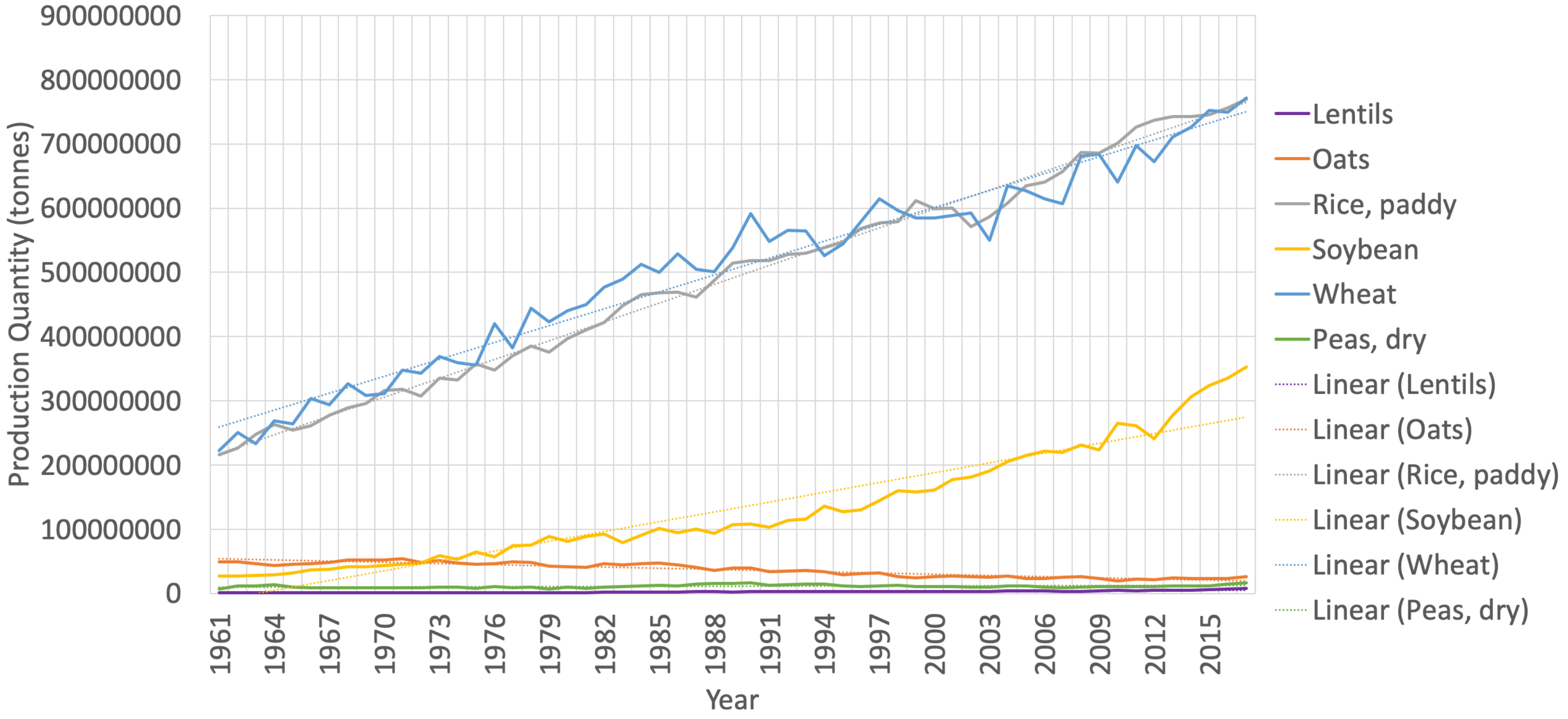


FIGURE 1.3: **Production Quantity for grain legumes and for three grain cereals (wheat, rice and oat) over five decades (1961-2017).** Legumes plotted: pea (green), soybean (yellow) and lentil (purple). Cereals plotted: wheat (blue), rice (grey) and oats (orange). Data obtained from *Food and Agricultural Organisation Statistical Databases of the United Nations*. Original in colour.

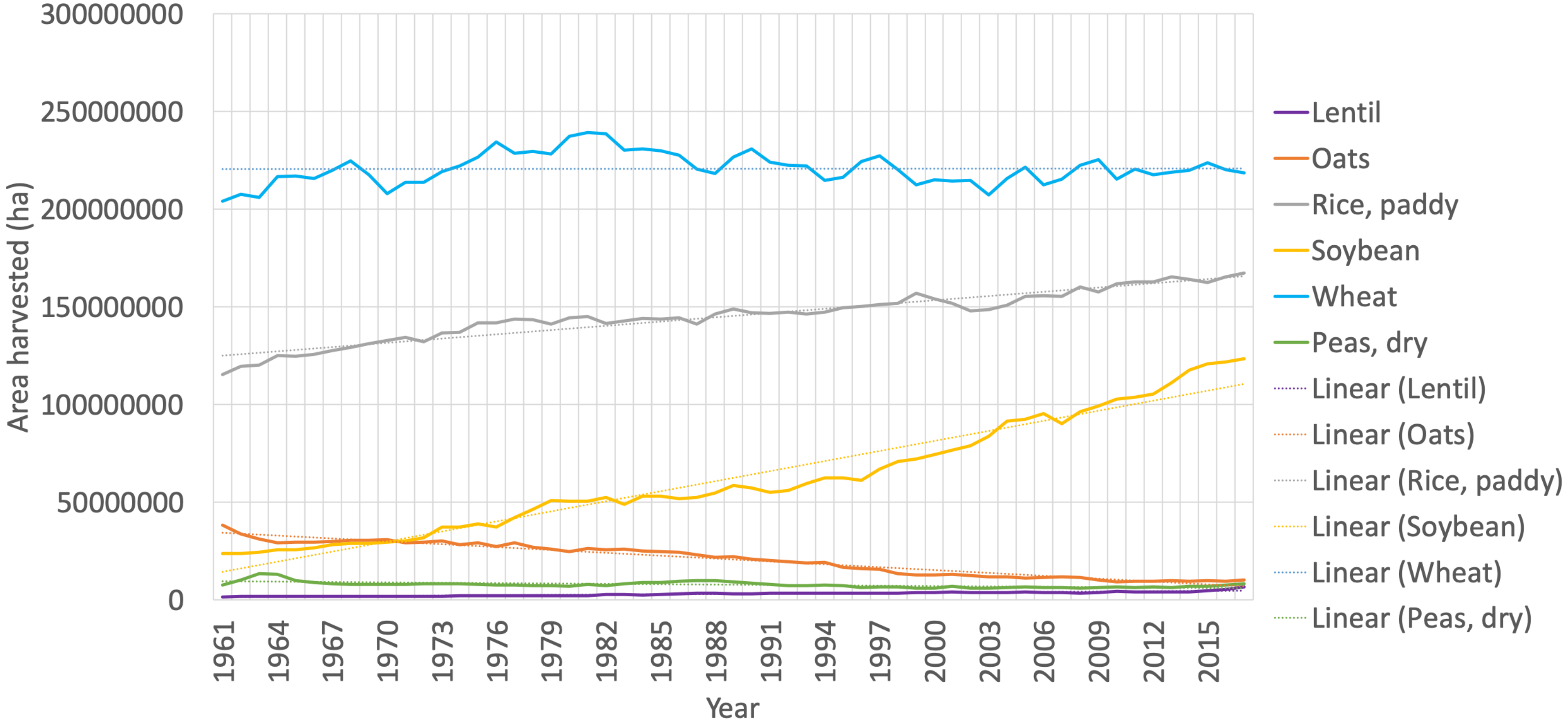


FIGURE 1.4: **Area Harvested for grain legumes and for three grain cereals (wheat, rice and oat) over five decades(1961-2017).** Legumes plotted: pea (green), soybean (yellow) and lentil (purple). Cereals plotted: wheat (blue), rice (grey) and oats (orange). Data obtained from *Food and Agricultural Organisation Statistical Databases of the United Nations*. Original in colour.

Nutritional Value

For these 6 crops, the nutritional values can be mined from the USDA National Nutrient database (*USDA National Nutrient Database for Standard Reference*). Pea is higher in protein composition in comparison to the other crops except lentil and it has the highest levels of Vitamin K (15.9 µg), Potassium (852 mg) and Riboflavin (0.244 mg) (Table 1.1). It is important to note, whilst grain cereals also have high protein values, not all protein is of the same nutritional quality as not all can be digested (Qaisrani et al., 2014). Proteins can be considered to be made up of essential and non-essential amino acids, the former cannot be made by the body and therefore must be derived from the diet (Tessari, Lante, and Mosca, 2016).

For instance, cereal proteins have a tendency to be lysine poor (Shewry, 2007) whereas legume proteins are lysine rich (Iqbal et al., 2006). Lysine is an essential amino acid that must be consumed in the diet and cannot be synthesised in the body. One study compared 4 grain legumes: chickpea, lentil, cowpea and green pea and found them to be rich in essential amino acids such as lysine, leucine and arginine (Iqbal et al., 2006). Often milling of cereals can reduce amounts of lysine as well as other nutrients such as vitamins and minerals (Oghbaei and Prakash, 2016). Brown rice has 2.13mg of Zinc per 100g (Table 1.1), however the milling of brown rice into white rice can reduce this amount (Oghbaei and Prakash, 2016).

The grain cereals all appear to have more Magnesium (Mg 116-177mg) compared to the grain legumes (Mg 47-65mg) and the grain cereals have less Potassium (K 250-429 mg) compared to grain legumes (K 620-852mg).

Legumes are vital to human and animal health, yet are still under consumed (Foyer et al., 2016). Recent initiatives by the United Nations Food and Agricultural Organisation have recognised this, and resulted with an International Year of Pulses in 2016 to promote the nutritional value of pulses (grain legumes) and make steps towards more sustainable agriculture (Foyer et al., 2016).

Legumes are good for our health because they are high in protein, fibre, vitamins, minerals and other phytochemicals but are also low Glycaemic Index (GI) (Kouris-Blazos and

TABLE 1.1: **Nutritional composition of 3 grain legumes and 3 grain cereals.** Data obtained from *USDA National Nutrient Database for Standard Reference*. USDA Codes: Peas, green, split, mature seeds, raw (172428); Lentils, raw (16069); Soybeans, green, raw (11450); Oats (20038); Rice, brown, long-grain, raw (20036); Wheat, flour, whole-grain (20080).

		Pea	Lentils	Soybean	Oats	Rice	Wheat
Nutrient	Unit	/100 g	/100 g	/100 g	/100 g	/100 g	/100 g
Proximates							
Water	g	8.69	8.26	67.5	8.22	11.8	10.74
Energy	kcal	364	352	147	389	367	340
Protein	g	23.12	24.63	12.95	16.89	7.54	13.21
Total lipid (fat)	g	3.89	1.06	6.8	6.9	3.2	2.5
Carbohydrate by difference	g	61.63	63.35	11.05	66.27	76.25	71.97
Fiber total dietary	g	22.2	10.7	4.2	10.6	3.6	10.7
Sugars total	g	3.14	2.03	-	-	0.66	0.41
Minerals							
Calcium, Ca	mg	46	35	197	54	9	34
Iron, Fe	mg	4.73	6.51	3.55	4.72	1.29	3.6
Magnesium, Mg	mg	63	47	65	177	116	137
Phosphorus, P	mg	334	281	194	523	311	357
Potassium, K	mg	852	677	620	429	250	363
Sodium, Na	mg	5	6	15	2	5	2
Zinc, Zn	mg	3.49	3.27	0.99	3.97	2.13	2.6
Vitamins							
Vitamin C, total ascorbic acid	mg	1.8	4.5	29	0	0	0
Thiamin	mg	0.719	0.873	0.435	0.763	0.541	0.502
Riboflavin	mg	0.244	0.211	0.175	0.139	0.095	0.165
Niacin	mg	3.608	2.605	1.65	0.961	6.494	4.957
Vitamin B-6	mg	0.14	0.54	0.065	0.119	0.477	0.407
Folate DFE	µg	15	479	165	56	23	44
Vitamin B-12	µg	0	0	0	0	0	0
Vitamin A, RAE	µg	7	2	9	0	0	0
Vitamin A, IU	IU	149	39	180	0	0	9
Vitamin E (alpha-tocopherol)	mg	0.12	0.49	-	-	0.6	0.71
Vitamin D	IU	0	0	0	0	0	0
Vitamin K (phylloquinone)	µg	15.9	5	-	-	0.6	1.9

Belski, 2016). Evidence has shown that those whose diets are void of legumes have a higher incidence of all-cause mortality (Chang et al., 2012). Furthermore, daily consumption of legumes reduces mortality in old people by as much as 8% (Darmadi-Blackberry et al., 2004) to 10% (Trichopoulou et al., 1995). Kouris-Blazos and Belski (2016) found that eating legumes reduces the incidence of cancer (including bowel cancer). Hashemi et al. (2014) looked at 11 studies and found daily eating of legumes for a month causes less fasting blood glucose and insulin. This is particularly important because legumes as part of a low GI diet, increase the control of glucose, reduce risk of cardiovascular disease and cancer through regulation of blood pressure, cholesterol and inflammation; but also act as a prebiotic which changes the flora of the bowel and affect gut hormones as well as changes in appetite (Kouris-Blazos and Belski, 2016; Sievenpiper et al., 2009).

Pea farming and food security

Peas are very relevant for food security because of the nitrogen fixation process, that was explored in Section 1.2. Peas require little fertiliser input so allow for less intensive farming and make the soil nitrogen-rich for subsequent crop, which make peas a good target for crop rotation.

With the human population expected to exceed 9 billion in 2050 (United Nations and Social Affairs, 2017), producing enough sustainably sourced food is becoming an increasingly pressing issue. More arable land is needed to help produce food for a growing population. Issues such as climate change, droughts, flooding and pests and diseases will affect the amount of food produced and cause yield instability. Producing more food of better nutritional quality from the same amount of land is a challenging task, but one that must be tackled (BBSRC, 2015).

The intensive use of fertiliser can have a negative effect on the environment (Jensen and Hauggaard-Nielsen, 2003). Peas are important as a break crop meaning that this will have impact on pests and diseases and in maintaining biodiversity. Intercropping is a strategy that can be employed, it is the growth of a cereal with a **companion legume** to reduce weed growth and increase soil fertility. The opposite of this is known as a **monoculture**.

Gene Banks

Gene banks have an important role in food security. Gene banks conserve plant genetic diversity, carefully characterise the accessions, distribute seed and maintain stocks of seeds in their possession (McCouch et al., 2012). One of the most famous examples of research within a gene bank would be the International Rice Research Institute's gene bank, where thousands of rice accessions were genotyped by high-density SNP array (McCouch et al., 2012). In addition to this, many other plant gene banks have been sequenced (Keilwagen et al., 2014; Carvalho et al., 2013; Wen et al., 2011).

Phenotyping a gene bank is an expensive and labour-intensive task, yet thanks to the reduction in cost of genotyping, it can be used to help to predict phenotypes in other closely related accessions, through identification of genetic similarity and population structure (Treuren and Hintum, 2014; Kilian and Graner, 2012). A combination of the advances in high-throughput genotyping and mid-throughput phenotyping mean entire collections can be characterised.

Due to the large volume of accessions to maintain and curate, gene banks must regenerate seed stock before it falls to a low threshold. Constricted financial budgets force gene banks to limit the collection to biologically important accessions whilst still striking a balance with a variety of accessions (McCouch et al., 2012). With so many accessions under their care, it is vital for correct documentation and accurate labelling of seed. Most gene banks have limited access to genetic information and only hold genetic markers on a subset of accessions (McCouch et al., 2012; Finkers et al., 2014). This project can be of great help to gene banks such as the JIC collection by providing a large number of markers on a subset of peas.

Genetic diversity, population structure and domestication of the pea

The genetic diversity of pea is an accumulation of all the genetic mutations that have occurred over its history as a result of a response to selection pressures in its environment (Jing et al., 2010). Wild pea relatives developed into local landraces when artificially selected and bred for beneficial traits by early breeders (Figure 1.5). Modern day breeders have taken landraces to form modern cultivars by selective breeding (Figure 1.5). This is domestication, a process primarily influenced by humans via artificial selection, domesticating wild species into landrace and cultivated material (Jing et al., 2010).

[Redacted]

FIGURE 1.5: **Proposed effect of domestication on pea genetic diversity.** Figure redacted.

The domestication of the pea is an important event in the pea's history. Domestication can have a negative impact on food security and to help reduce the adverse effects of domestication, it is important to understand more about genetic diversity (Hyten et al., 2006). Domestication can lead to a lack of variation and lack of allelic evenness and richness, which is defined as **genetic erosion**. The function of gene banks is to conserve plant genetics. Modern breeding practices such as excessive use of fertilizers or urbanisation of local areas can cause a loss of landraces since they are not adapted to cope with this environment. Furthermore, climate change has adversely affected some varieties, and a growing human population has changed food preferences as well as changed demand for local products. These factors have caused genetic erosion and ultimately have a negative effect on local and global food security (Wouw et al., 2010).

The change in diversity comes in different phases, and have different bottlenecks (Figure 1.5). The **domestication bottleneck** is an intensive artificial selection on agriculturally desirable traits whilst the **dispersal bottleneck** is dependent on how the crops have been dispersed (Wouw et al., 2010). These bottlenecks can cause a reduction in diversity, but this is eventually alleviated by the gene flow between domestic and wild relatives. It is not uncommon for landraces can gain new diversity by **introgression** - a gene entering another species gene pool through hybridisation and backcrossing (Harrison and Larson, 2014). The **modernisation bottleneck** is a result of scientific and modern breeding practices (Wouw et al., 2010).

Diversity Structure Analysis of the pea from Retrotransposon based insertion polymorphism (RBIP) markers (Jing et al., 2010) has divided the peas within the JIC collection into different groups (Figure 1.6 reproduced from Jing et al. (2010)). Groups 1 and 2 are predominantly cultivated varieties where group 1 consists of predominantly landraces and group 2 of vining and combine cultivars. Next comes an unclassified group called Unknown that is mainly "weedy" material and then Group 3, which is more exotic, wild material. In some cases, group Unknown is referred to as Group 3 and Group 3 is referred to as Group 4, in order to acknowledge the presence of the unknown group. These groups have been subdivided into a total of 14 subgroups; seven subgroups in group 1, two subgroups in group 2, and seven subgroups in group 3 (Jing et al., 2010).

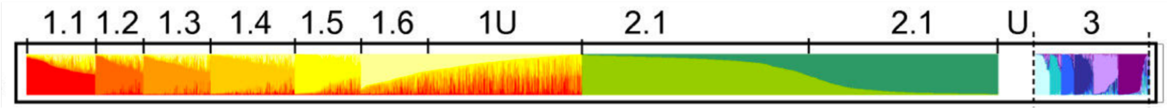


FIGURE 1.6: **Structure analysis from RBIP markers on the entire John Innes Centre collection categorising these into 3 main groups.** Figure obtained from Jing et al. (2010) and was cropped. This is an Open Access article under Creative Commons Attribution License (`http://creativecommons.org/licenses/by/2.0`). Original in colour.

From a figure produced by Jing et al. (2010) (see annotated Figure 1.7), the wild varieties are found in subgroups 2 and 6 of Group 3 (alternatively called group 4) and the majority of landrace and cultivated material are found in groups 1 and 2. For the pea, the cultivated peas are *Pisum sativum sativum*, but also include *Pisum sativum abysinnicum* which is an independently domestivated cultivar (Smýkal et al., 2012). The wild relatives include *Pisum sativum elatius*, *Pisum sativum humile* and *Pisum fulvum*. *P. fulvum* originated from around the Middle East. *P. abyssinicum* originated from around Yemen, Ethiopia. *P.elatius* is distributed across the Mediterranean (see annotated Figure 1.7 from Jing et al. (2010)).

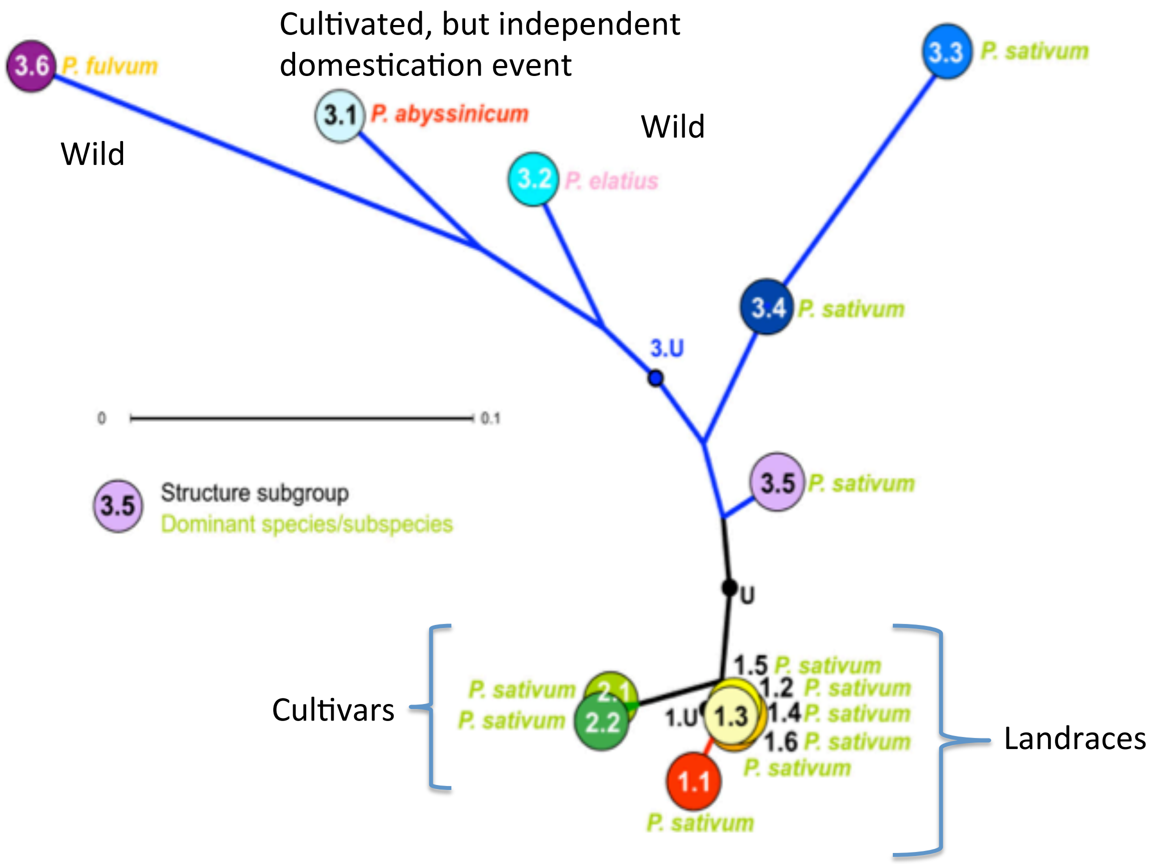


FIGURE 1.7: **The different subgroups of pea based on genetic distance.** It shows how they have evolved and domesticated. Figure was adapted and taken from Jing et al. (2010). This is an Open Access article under Creative Commons Attribution License (`http://creativecommons.org/licenses/by/2.0)` Original in colour.

The pea genome

The pea genome is large and complex, with a haploid genome size of 4.45Gb (Doležel and Greilhuber, 2010) arranged into 7 chromosomes (Neumann et al., 2002). The genome is thought to be highly repetitive (75-97%) with a GC content of 37.7% (Salinas et al., 1988). Furthermore, there is currently no publicly available genome reference.

This makes pea genomic analysis particularly challenging because, with no reference to compare to, all sequencing requires greater coverage to have more confidence for variant calling and can limit bioinformatic analysis to the tools which are able to assemble reads *de novo* (Pop, 2009). Furthermore, sequencing a draft reference is costly and particularly challenging since the genome is large, requiring greater computational memory resources than a smaller genome would (Zhang et al., 2011). Additionally, the challenges are exacerbated by the repetitiveness of the genome (Murray, Peters, and Thompson, 1981) since it becomes difficult to assemble sequencing reads because it becomes difficult to accurately place the repeats (Phillippy, Schatz, and Pop, 2008). All of the issues above, must be carefully considered.

The Pea Reference - Current Status

Currently, there is an International Consortium set up for Pea Genome Sequencing (PGS), consisting of scientists from around the world spear-headed by Institut National de la Recherche Agronomique (INRA) in France. The pea reference genome, based on the cultivated variety Cameor, is currently complete but is not being released until the pseudomolecules are finished (Jonathan Kreplak, *pers.comms.* 2017).

This reference genome assembly was formed using a combination of Illumina reads, PacBio reads and physical maps. Using 146x of Paired End reads, 129x Mate Pair, 13x PacBio reads and the BioNano to reduce chimerism, the assembly metrics are as follows: 19,582 scaffolds; 553,390 L50; 1986 N50; 12.08% N and a max size of 2,846,597 (Jonathan Kreplak, *pers.comms.* 2017).

Since the reference is not publicly available, it would be prudent to assemble one based on a cultivar of interest. This will enable the alignment of reads and lead to more confidence in the results.

Pea Genetic Maps

In genome mapping, there are two types: genetic maps and physical maps (Brown, 2002). **Genetic maps** are based on genetic distances (measured in **centimorgans**) from linkage analysis. This is created by obtaining genetic markers from a mapping population. Using Mendel's laws of Co-segregation, the proximity of these markers indicated the likelihood of being inherited together - called **Linkage Disequilibrium**. The first genetic map was made in the *Drosophila* (Sturtevant, 1913). Conversely, **physical maps** are based on physical distances (measured in base pairs) (Brown, 2002).

These gene maps document genetic loci and distances between them. In pea, there are many linkage maps that have been made. One example by Ellis et al. (1992), used a RIL cross to make this map, and highlighted the importance of performing many different crosses to provide more useful data. Additionally, Weeden, Brauner, and Przyborowski (2002) created a genetic map with additional markers for pod dehiscence - which will be an important map for studying domestication and one transcriptomic study generated a genetic map from the SNPs called from their study (Sindhu et al., 2014). These maps can be used in the orientation of markers.

The pea transcriptome

There have been several transcriptome assemblies produced, with the basic statistics of key assemblies in the community outlined in Table 1.2.

The majority of these assemblies have been produced using 454 sequencing. The exceptions to these are, Unigene v1 which used Expressed Sequence Tags (ESTs) (*Pisum sativum unigene v1*), Unigene v2 which used 454 and ESTs (*Pisum sativum unigene v2*), CSFL RefTrans v1 (*P. sativum CSFL RefTrans V1*) which used a combination of 454 and Illumina sequencing and CSFL RefTrans v2 (*P. sativum CSFL RefTrans V2*) which used a combination of 454 and Illumina sequencing with ESTs.

Franssen et al. (2011) produced a first and second pass transcriptomic assembly which resulted in what the authors of the paper consider to be near complete transcript coverage. The difficulties of assembling a *de novo* transcriptome assembly of a large and repetitive genome have impacted the work in this study by producing many redundant contigs. Redundant

TABLE 1.2: **Comparison of some key versions of the pea transcriptome**

Analysis Name	Sequencing Method	Number of reads	Assembly tool	Number of Contigs	Number of Singlets	Citation
Unigene wa1	454	3,804,253	Newbler	37,455	Not Stated	*Pisum sativum unigene wa1*
Unigene v1	ESTs	17,717	Cap3 (-p 90)	3,095	6,950	*Pisum sativum unigene v1*
Unigene v2	(combined wa1 and v1)	47,500	Cap3 (-p 90)	8,817	26,029	*Pisum sativum unigene v2*
Franssen	454	2,209,755	MIRA (first assembly) TGICL pipeline (Second assembly)	42,000	26,622	Franssen et al. (2011)
Kaur	454	720,324	NextGene	13,583	57,099	Kaur et al. (2012)
Duarte	454	3,826,797	MIRA(-est)	68,850	Not Stated	Duarte et al. (2014)
Sindhu	454	4,008,648	Ngen(DNAstar)	29,725	Not Stated	Sindhu et al. (2014)
CSFL RefTrans v1	454 and Illumina publicly downloaded	18,576 ESTs 2.9bill RNAseq (2.6B PE; 300m SE)	Trinity v2.0.6 (RNA) Cap3 (ESTs)	45,727	Not Stated	*P. sativum CSFL RefTrans V1*
CSFL RefTrans v2	454 and Illumina and ESTs publicly downloaded	2.4 billion RNA-Seq reads and 18,576 ESTs	Mainlab RefTrans pipeline (unpublished)	63,990	Not Stated	*P. sativum CSFL RefTrans V2*

contigs in a transcriptomic assembly can have potential to mistakenly identify redundant contigs from the same gene as differentially expressed genes in comparative RNA-seq studies (Ono et al., 2015). Nevertheless, this novel contribution of the transcriptomic dataset is useful to the pea community and could be valuable to this project at a later date. Franssen et al. (2011) have identified that upon the release of the pea reference genome; the combination of a reference genome and transcriptome dataset will be a powerful combination for the pea community.

Kaur et al. (2012) also used Roche 454 reads to produce a transcriptome with almost as many unigenes and annotated unigenes in comparison to *Medicago truncatula* as Franssen et al. (2011). This has 45,161 overlapping hits with Franssen et al. (2011), showing how these studies are complementary. Furthermore, they also discovered some SSR markers in the pea. This work has shown how transcriptomic datasets have potential to help with marker assisted selection.

More recently, another study used Roche 454 to identify SNPs (Duarte et al., 2014). It has produced a relevant and up-to-date genetic map, identified 35,000 SNPs and made new cDNA contigs based on 92 genotypes. Again, this too, has potential to be useful for Marker Assisted Selection.

The most current and up-to-date pea transcriptome, the Cool Season Food Legumes pea reference transcriptome, publicly released pea RNA-seq read data from 454 and Illumina sequencing technologies as well as ESTs and combining them to form the most updated transcriptome so far, to our knowledge (*P. sativum CSFL RefTrans V2*).

These key works (Duarte et al., 2014; Franssen et al., 2011; Kaur et al., 2012; *P. sativum CSFL RefTrans V2*) have made a valuable contribution to the pea community and the resulting datasets may assist the Pea Genome Sequencing Project.

Legume Phylogenetics and Synteny

Zhu et al. (2005) have illustrated how pea fits in within the phylogenetic tree of legumes (Figure 1.8 obtained from Zhu et al. (2005)).

[Redacted]

FIGURE 1.8: **Legume phylogenetics of cool and tropical season legumes.** Figure redacted.

Figure 1.8 demonstrates organisms in the Galegoid clade are cool season legumes and organisms in the Phaseoloid clade are tropical season legumes. It is based on previous work on maximum parsimony phylogenetic analysis of the *matK* gene found in plastids (Steele and Wojciechowski, 2003).

This indicates *Medicago truncatula* and *Lotus japonicus* may serve as a better comparison with pea, whereas other legumes such as *Glycine max* may not serve so well. Such information can help to order and orientate genetic markers in the absence of a pea genome reference through **synteny** with a closely related legume. Additionally, it is commonly known that pea has conserved gene order with *Medicago truncatula* and *Lotus japonicas* (Simon and MuehIbauer, 1997): two types of cool season legumes which also have reference sequences.

Pea has 7 chromosomes organised into linkage groups and is known to share synteny and high levels of **collinearity** with *Medicago truncatula*, the closest model organism to pea (Smýkal et al., 2012). *Medicago truncatula* has 8 chromosomes and a genome size of approximately 500Mb (Smýkal et al., 2012). Given the synteny between pea and *M.truncatula* it may be possible to position pea genes with respect to Medicago (Aubert et al., 2006).

Several studies have investigated synteny of pea and *Medicago truncatula*, one of which has resulted in consensus maps (Figure 1.9 obtained from Zhu et al. (2005)).

[Redacted]

FIGURE 1.9: **Consensus map of several legumes**. Figure redacted.

Kalo et al. (2004) investigated the synteny between pea and *Medicago sativum* (a relative of *M.truncatula*). This study suggests that due to the high levels of synteny it can be inferred that pea did not come from a smaller genome and pea's large genome size was not caused by genome duplication or amplification due to few chromosomal rearrangements. This study also found high levels of collinearity between the two organisms.

Bordat et al. (2011) researched the translational genomics of pea with *M.truncatula*. They used SNPs to draw a pea functional consensus map, looking for macrosynteny in pea and *M.truncatula* (as well as other legumes). Interestingly, their dot-plot of pea and *M.truncatula* shows more syntenic conservation than their dot-plots with other legumes. This demonstrates that *Medicago truncatula* is a good choice of relative legume to use in syntenic studies with pea.

Background to genetical analysis of traits in pea

Height

Why is height important to pea?

Mendel observed that peas differed in stem length, they were either tall or short in height and that taller plants were the more dominant trait (Mendel, 1996). Plant height can help improve the plant's access to light, thus making it competitive and more likely to photosynthesise, however, plant height is biologically "expensive" as maintenance and upkeep of the stem can be difficult (Falster and Westoby, 2003).

Lodging is a common problem where plants fall over and this reduces the harvest and increases cost (Tar'an et al., 2003). One pea variety that resists lodging, *afila* semi-leafless plants, are available. These mutants have reduced leaves and an excess of tendrils that intertwine to help hold up the plant (Tar'an et al., 2003). Furthermore, dwarf (*le*) mutants are also available. These dwarf mutants have short internode length so are less likely to succumb to lodging (Tar'an et al., 2003). Height is a domesticated trait and tall plant height is commonly found in wild material and dwarf height in domesticated material (Weeden, 2007).

Genetic analysis of height in pea

The *Le* gene is also known as the Mendel's stem length gene and controls the length of the internodes along the stem (Lester et al., 1997). The *Le* gene encodes for Gibberellin 3B-hydroxylase (GA2bOH) which converts gibberellin (GA20) to an active form (GA1) (Lester et al., 1997). The *le* mutants have impaired gibberellin metabolism resulting in shorter internode length (Sherriff et al., 1994). Tar'an et al. (2003) have found that the *Le* gene is responsible for height and its role in lodging resistance and the authors have also found several other markers (cttg7, caag4 and cagg5) for the plant height trait.

Another QTL found on linkage group IV, is positioned near the gene for gibberellin 2B-hydroxylase (Weeden, 2007). Presence of this QTL causes shortness of height, even if the *Le* allele for increased height is present (Weeden, 2007).

Three QTLs named *ht1*, *2* and *3* are found on linkage groups II, III, and IV respectively. Of these, the largest contributor to height was *ht2* (Prioul et al., 2004).

Seed traits

Why are seed traits important to pea?

Seed weight is a heritable trait in pea (Singh, Singh, and Babu, 2011) and is linked to vigour, in other words, how likely a seed is to grow into a plant and develop. Peksen et al. (2004) showed that low seed weight leads to high germination in pea. Understanding which varieties have better seed weight may assist towards better understanding and usage in the context of food security. Wild seeds are known to be smaller and domesticated seeds are larger in size (Weeden, 2007).

Mendel's work observed that seed roundness or wrinkledness is heritable and that round seeds are the dominant trait. There are commercial uses for petite pois which are smaller and sweeter peas, and for marrowfat peas which are larger and less tasty like those commonly found in mushy peas. Commercially, there may be more interest in rounder peas than less aesthetically pleasing shapes and understanding the genetic causes for this can better inform breeding choices.

The wrinkling is known to be caused by changes in the branching of starch molecules, giving these a perceived sweeter taste (Rayner et al., 2017).

Genetic analysis of seed traits in pea

The *rugosus* locus has been identified as causative gene for the wrinkled seed phenotype observed by Mendel (Bhattacharyya et al., 1990). The allele *r* is a mutation in the enzyme, starch branching enzyme I, which affects amylopectin synthesis (Rayner et al., 2017), a polysaccharide that in combination with amylose creates starch (Casey et al., 1998). This increases the sucrose concentration resulting in a sweeter tasting pea (Rayner et al., 2017). The high sugar concentration creates high osmotic pressure and initially there is high water uptake, but water loss will occur leaving a wrinkled phenotype (Wang and Hedley, 1991).

The *rb* allele affects the structure and activity of an enzyme called ADP-Glucose Pyrophosphorylase and this changes pea seed shape by increasing the sucrose concentration from 5% to 9% (Hylton and Smith, 1992).

The *rug3* mutations affects the phosphoglucomutase (PGM) which catalyses the reaction of Glucose-6-P into Glucose-1-P and mutants are near-starchless resulting in an extremely wrinkled phenotype (Harrison et al., 2000).

The *rug4* mutations affect the enzyme sucrose synthase which converts sucrose into UDP-glucose and fructose. This is the only gene in the pea starch metabolism pathway which occurs in the cytosol and not the plastid (Casey et al., 1998). Therefore, since not all of the downstream substrates will be used to create starch, the efficacy of this gene on starch metabolism is low resulting in the *rug4* mutants holding 20% less starch (Casey et al., 1998).

The *rug5* mutants affect the enzyme, starch synthase II. This mutation affects starch content by 30-40% with large effects on the structure of amylopectin but does not affect other enzymes in this pathway (Craig et al., 1998).

Another gene that affects the starch synthesis in pea is *lam* which is a low amylose mutant, however, it has similar levels of starch as wild type peas (Bogracheva et al., 1999). This affects the granule bound starch synthase I enzyme that converts ADP-glucose into amylose (Casey et al., 1998).

These 5 genes are best known for their effect on seed wrinkledness, but this does not take into consideration other genes that have pleiotropic effects on seed roundness. For example, Mendel's flower colour gene, A, is thought to have an effect on seed dormancy and quality (Weeden, 2007).

For seed size, there are many QTLs found. Pleiotropic genes such as *Np* which corresponds to presence of pod neoplasms but loss of this gene increases seed size (Weeden, 2007). The *r* gene improves sweetness but also reduces seed size (Weeden, 2007). There are 3 QTLs for seed size (QTL I, IV, VII) (Weeden, 2007). In linkage group V, a QTL corresponding to seed weight was found (Krajewski et al., 2012).

Leaflet

Why are leaflet traits important to pea?

Observations based on leaflet phenotypes have often been used to classify plants. Previous studies such as those by Xu et al. (2009) and Royer et al. (2005) found an increased presence of leaflet teeth correlated with plants grown in cold temperatures. Peppe et al. (2011) identified a low presence of teeth in wet climates.

Some peas have round (also known as "entire") margins whilst others have serrated margins. It is known that wild pea varieties are more serrated than cultivated material (Holdsworth et al., 2017). Leaves with teeth tend to hold greater photosynthetic capacity and are more active in terms of transpiration, additionally, plants grown in shade tend to have more teeth - this makes round margins more beneficial in a drought environment (Xu et al., 2009).

Leaf shape is important and in addition to teeth has also been used to help classify plants. The primary function of the leaf is to photosynthesise and changes in leaf shape can affect the photosynthetic capacity (Kidner and Umbreen, 2010). The way in which the leaf lamina grows affects it shape leading to impacts different function of the leaf such as photosynthesis, transpiration and even temperature regulation (Kidner and Umbreen, 2010).

Genetic analysis of leaflet traits in pea

The afila (*af*) mutant contribute to a semi-leafless phenotype (*afaf*++) and in combination with the stipule reduced (*st*) mutant contribute to a leafless phenotype (*afafstst*) (Lafond, Evans, and Ali-Khan, 1981). Scientists at The John Innes Centre created these mutants in the 1970s (Snoad, 1974). These mutants have fewer leaves and instead have more tendrils which aid support by holding the plant upright (Heath and Hebblethwaite, 1985). These increased tendrils increase the likelihood of knitting together to other peas and to knit to any supportive netting, as a result, these mutants have reduced lodging (a process where the plant falls over) and better standing ability.

Knotted1 (*Kn1*) is a gene which can lead to lobes (Hofer and Ellis, 1998) whereas the Serrated (*Ser*) gene can lead to leaflet teeth (Weeden and Ambrose, 2004). The presence of teeth or lobes can affect the area of the leaflet.

The gene unifolia (*uni*) changes the number of leaflets to one leaflet only (Hofer and Ellis, 1998). The *uni* gene is down-regulated by the hormone auxin (Bai and DeMason, 2006)

Pod

Why are pod traits important to pea?

Pods are the fruit of *Pisum sativum* that hold the pea seed. The shape of the pod is known to affect yield of the plant. Pods with a pointed apex leads to increased seed abortion (Lee, 1988). Wild pods are known to be shorter and thinner than domesticated varieties.

In some wild pods, another purpose of the pod is seed dehiscence (also known as “seed shatter”) - a mechanism for seed dispersal, whereas domesticated peas have been bred for seed retention (Weeden, 2007). Pods can also help protect the seed from pests and infection, for example, in some conditions, peas may develop growths on the pod, called neoplasms which enable the plant to resist pea bruchid infection (Doss et al., 2000).

Genetic analysis of pod traits in pea

This thesis intends to explore pod size and shape. Pods with the *n* gene have thick walls but they also appear to have greater curvature to the pod (Wehner and Gritton, 1981). Pods with the *nn* genotype are also smaller in length and width as well as a smaller area (Wehner and Gritton, 1981). This gene is one of interest to this thesis with regards to pod traits.

Genetic analysis into pod dehiscence have identified the genes *Dpo1* and *2* as responsible for pods undergoing seed shatter in wild peas (Weeden, 2007).

The *Np* gene is responsible for the growth of pod neoplasms that protect the plant from insect attack (Doss et al., 2000), but interestingly loss of the *Np* gene has an effect on increasing seed size (Weeden, 2007). Other genes involved in pod texture is the example of the *sin-2* gene and this creates a stringless pod which is a commercially desirable trait in edible pods (McGee and Baggett, 1992).

1.4 Analysis of Germplasm Collections

Chapter 2 outlines how a core collection of peas from the John Innes Centre *Pisum* germplasm was obtained. When considering the accessions to be included in a core collection, the population structure is a consideration. There is a trade-off to be made, either to include cultivars with desired traits which may reduce the richness in genetic diversity, or to include wild relatives that are more diverse but are harder to grow and to phenotype.

Chapter 2 presents a representative core collection and chooses from all groups and subgroups of the population that have been outlined in Jing et al. (2010). Random sampling of all subgroups would be beneficial to remove biases, paying particular attention to groups 1 and 2 which are the landrace and cultivars, respectively.

There are many ways to evaluate a collection: Summary statistics, Principle Components Analysis, Shannon Diversity Index, Class coverage and Chi-square (Odong et al., 2013).

Chapter 2 opted for one and two-sample Kolmogorov-Smirnov tests (K-S tests) to compare the JIC and total selection against a normal distribution and against each other.

K-S tests detect if the samples being tested are significantly different from a normal distribution (Ennos, 2007). However, with small samples sizes, the K-S test may produce a type II error. This means it will fail to find a significant difference when it is actually present (Ennos, 2007).

Phenotyping

Phenotyping can be performed manually or in an automated manner. Currently, manual phenotyping is laborious but requires little input. Automated phenotyping has made massive advances and is hugely beneficial for Genome-wide Association Studies (GWAS) as shown by the recent work in rice (Yang et al., 2014). Automated phenotyping requires more resources and is more expensive. In Chapter 3, phenotyping both in the field and in the greenhouse is performed to account for the "Genotyping by Environment" (Pieruschka and Poorter, 2012).

However, the difficulties do not end with plant phenotyping, but also lie further downstream in data management (Billiau et al., 2012) and data standardisation. The **transPLANT** consortium is a European consortium with the purpose of improving software infrastructure, data management and data standardisation, specifically for plants. They have created a standardised method for meta-data and data exchange (Ćwiek-Kupczyńska et al., 2016).

The transPLANT project has produced a more standardised way of recording phenotypic metadata, called **Minimum Information about Plant Phenotyping Experiment (MIAPPE)**. MIAPPE is a list of minimum information that is needed to do a phenotyping experiment (Ćwiek-Kupczyńska et al., 2016). One alternative to MIAPPE is **ISA-Tab** - a data format for phenotypic metadata. It is a way to organise experiments into Investigation, Study and Assay levels, to create a standardised way of documenting and sharing phenotypic data

(Rocca-Serra et al., 2010). Both MIAPPE and ISA-Tab are ways to share phenotypic data in a consistent and standardised manner.

The John Innes Centre holds some historic phenotyping for pea, which is not always complete or consistent. This is a cause for concern when comparing phenotypes, so whilst still a useful addition, this cannot be solely relied on. Chapter 3 of this project begins to explore Simple Shape and Morphological Descriptors (SMSDs) in pea, namely, phenotypes that relate to plant height, seed weight as well as leaflet, pod and seed morphology.

1.5 Background to Methods

1.5.1 Image Analysis

Chapter 3 outlines a new image analysis tool called "MktStall". Image analysis tools come in two main forms: general and specific tools. General tools require computational expertise and rely on the user's technical ability. Specific tools do not require expertise to run.

General Tools

A good example of a general tool is ImageJ (Schneider, Rasband, and Eliceiri, 2012). Historically, it was created for microscopy images however over recent years it has been used in a multitude of different applications including organ images. The different distributions of ImageJ include: FIJI (Fiji is Just ImageJ) (Schindelin et al., 2012), ImageJ1, ImageJ2, ImageJA as well as obsolete versions, such as ImageJX. ImageJ provides distributions for Windows, Mac and Linux making it a cross-platform tool. It is supported by a large community base that often creating their own plugins that can be used by others. Some of these plugins have been published (Polder, Blokker, and Heijden, 2012), but others have not, raising questions over whether the source code is provable since the results have not been peer reviewed. Since ImageJ relies heavily on the user customising their own scripts, this will depend on their level of computational expertise and does not always mean that it provides reproducibility of results. Therefore, several scientists can measure the same image, but given different techniques, it has the potential to arrive at different answers. Also, there is a growing trend amongst scientists to simply state that ImageJ was used in their analysis within the methods section without explanation of how this was conducted, further contributing to a lack of reproducibility. Merely stating the use of ImageJ in the method is not a reproducible statement. Instead, it should state how ImageJ was used by the user, unless a citable plugin was used. Although ImageJ can be automated and batch processed, some key plugins remain tightly bound to the Graphical User Interface (GUI) (Newton and Deugo, 2007). This means there would be issues batch processing the images which would require user intervention and prevent full automation. Such issues have arisen from ImageJ's community base because these users are not necessarily computational experts and inevitably, there are some shortcomings with their plugins.

Chapter 3 aims to produce images that are human readable. It is far more convenient for users to read a printed label than it is to read a QR code, particularly when working in field environments where QR readers would be another piece of equipment to carry and would require additional considerations, such as battery life when considering usage of smartphones. Also, scale is an important issue. Whilst many different objects of known size could be used as a scale, these are not necessarily easily recognisable, hence, a ruler of known SI measurement is far more appropriate. Furthermore, Section 1.5.2 establishes the importance of using ontology definitions to provide accurate measurements. ImageJ can be inaccurate with the optical character recognition (OCR) to read labels, it cannot measure according to an ontology and does not provide an automated scale reader that is human readable.

Specific Tools

Specific tools are designed to address a particular biological question, for example, can insect damage be quantified in leaves (Machado et al., 2016; Bakr, 2005) or can grain colour be quantified to provide information of grain quality (Whan et al., 2014), to name but a few examples of specific tools answering such questions. In specific tools, the code solution to answer this particular question has been encapsulated into the software. Therefore, often, specific tools do not require computational expertise from the user. The "Plant Image Analysis" database (Lobet, 2017; Lobet, Draye, and Périlleux, 2013) is a database of many different categories of both general and specific image analysis tools. Categories for these specific tools include canopy, cell, flower, fruit, hypocotyl, leaf, plant, root system, rosette, seed, shoot and single root. Of these, only fruit, leaf and seed are of interest to the work outlined in this thesis.

A review of all relevant categories (fruit, leaf and seed) from the "Plant Image Analysis" database was conducted. Figure 1.10 shows how these do not apply to our context. For example, there is currently no specific fruit software tool for pea. Those which are specific to leaf and seed are either commercialised, require specialist equipment, apply to other plants and other phenotypes or are not cross platform. Furthermore, only a few tools are considered useable, but are not truly usable to this specific context because not all phenotypes that need to be quantified can be measured using these tools. Here, we summarise the review performed from the "Plant Image Analysis" database (Lobet, 2017).

Commercial Image Analysis Tools

Commercial Licensing requires users to buy a licence to use the tool. Commercial software can be also be sold in various grades and so higher-grade software have higher financial costs associated with them but provide the user with more features. Whilst commercial tools can be useful, it must be noted that these tools are less accessible. A list of such software tools for leaf and seed and provided in Table A.1.

Specialised Image Analysis Tools

Some image analysis tools are designed to work with specialist equipment. Specialised tools require such equipment because they can be incompatible with other pieces of equipment or reach sub-optimal performance for accurate measurements. Examples of such tools for leaf and seed are provided in Table A.1.

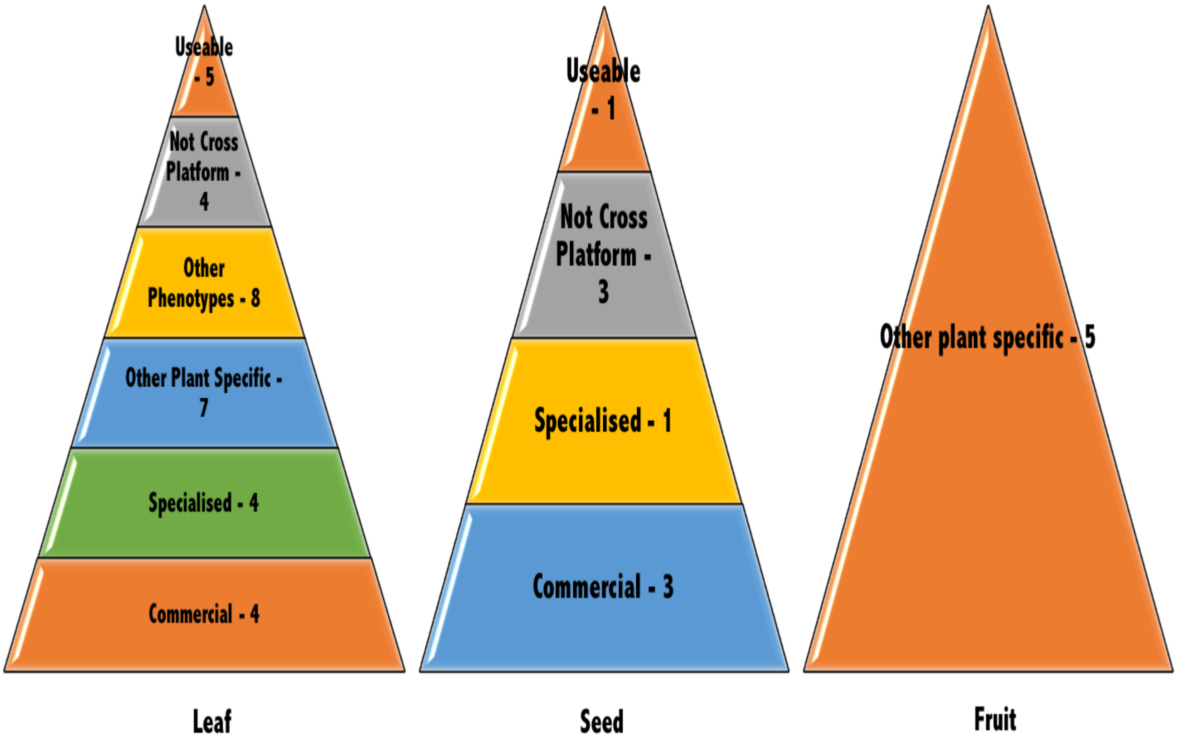


FIGURE 1.10: **Plant image analysis tools for leaflet, seed and fruit.** The distribution for each organ is categorised into Commercial tools, Specialised tools, Other Plant Specific, Other phenotypes, Not Cross-Platform and Usable tools. The tools for each organ was obtained from the Plant Image Analysis Database (Lobet, 2017).

Plant specific Image Analysis Tools

Some image analysis tools are developed to analyse a particular plant and therefore it will not necessarily perform well in the organism of choice, pea. There are image analysis tools for Arabidopsis (Remmler and Rolland-Lagan, 2012), for monocots (Dornbusch and Andrieu, 2010), for cash crops like maize (Miller et al., 2017), rice (Faroq et al., 2013), potato (*Potatosize*), tomato (Brewer et al., 2006), grape (Hill et al., 2011), as well as woody plants (Novotnỳ and Suk, 2013) and trees (Kumar et al., 2012). A list of tools for leaf and fruit are available in Table A.1.

Phenotype specific Image Analysis Tools

Other tools are bespoke to a particular phenotype, which would render it unusable if the it does not match our phenotypes of interest. Examples of phenotypes which the tools in Table A.1 address which do not match the interests of this thesis include leaf venation (Bühler et al., 2015; Max Planck Institute, 2018; Price et al., 2010), leaf disease (Pethybridge and Nelson, 2015), leaf damage by insects (Machado et al., 2016; Bakr, 2005) and recognition of plants through leaves (Joly et al., 2014). All tools from this literature review are present in Table A.1.

Cross Platform Image Analysis Tools

The user's **operating system (OS)** should be of some consideration. Key operating systems include Windows, Mac and Linux operating systems. Tools that are not cross-platform will only work on the native OS and cannot be used on other operating systems. Some tools are written for a particular OS but can be used on other platforms using an emulator (Varma and Osuri, 2013) but others will only work on particular platforms (Easlon and Bloom, 2014; Joosen et al., 2010; Tanabata et al., 2012; Whan et al., 2014). A list of all tools from this literature review are presented in Table A.1.

Usable Image Analysis Tools

There is no single specific tool for all of the phenotypes or for all of the organs that this thesis aims to observe. To observe multiple phenotypes would require passing images through many pieces of software and this would take up significant runtime. For example, leaflet dimensions could be measured with certain tools (Bylesjö et al., 2008; Green et al., 2012) but teeth would have to be measured with another tool (Biot et al., 2016; Backhaus et al., 2010). However, since there are no specific tools for fruit which are usable for pea, this means pod phenotypes cannot be measured with a specific tool. For seed, a usable tool is available but still requires coding proficiency from the user to develop a pipeline (Gehan et al., 2017).

Such information shows us that there is a need for specific tools capable of analysing multiple phenotypes of multiple organs that is "usable". In Chapter 3, a cross-platform tool with no commercial licence, that is designed to work without additional specialised equipment that will work on pea and the phenotypes of interest in this thesis is described.

1.5.2 Ontologies

To phenotype accurately MktStall uses the Plant Trait Ontology (Arnaud et al., 2012). Phenotype Ontologies are sets of words and nomenclature that describe a phenotype. They help describe different types of elements and the relationships and interactions between them (Mabee et al., 2007). Phenotype Ontologies allow the description and characterisation of phenotypes in a standardized fashion (Mungall et al., 2010).

The OBO Foundry [Open and (Biological and) Biomedical Foundry] contains more than 60 ontologies (Smith et al., 2007). Each ontology comprises of the same core values; the ontology must be open, formal and well-documented. All ambiguity must be removed through use of unique identifier codes, clear content and textual definitions.

The EQ syntax describes the entity-quality value, providing a biological feature with a value, to describe the phenotype. EQ values allow the integration of phenotypic data and enables comparison. The cROP initiative is the common reference ontologies for plants (Consortium, 2015). Within this, is the Plant Trait Ontology and this uses the EQ syntax.

Genome-wide Association Studies produces genotype-to-phenotype data (Thorisson, Muilu, and Brookes, 2009) through identifying statistically associated associations. It would be wise for MktStall to include ontologies such as the Plant Trait Ontology in order to make the best use of this genotype-to-phenotype data produced from a GWAS. Phenotype ontologies assist in the annotation and description of genotype-to-phenotype associations. Despite their uses, it can be difficult to integrate phenotype ontologies across multiple species (Mungall et al., 2010).

1.5.3 Sequencing Technology

Sequencing has been around for over forty years and has undergone many advances in this time (Shendure et al., 2017). Sequencing elucidates the order of the nucleotides found in

DNA. It is this that gives scientists an understanding of genes that may encode for particular phenotypes and which maybe evolutionarily conserved with other organisms through sequence homology, synteny and collinearity.

Second Generation Sequencing

The Human Genome Project paved the way for massively parallelized sequencing - **Next Generation Sequencing (NGS)**. Though there are many platforms available, this thesis focuses on the Illumina sequencing platforms which are well-known for a high-throughput sequencing output and capacity for massively parallelized sequencing (Shendure et al., 2017).

Figure 1.11, reproduced from Shendure et al. (2017) illustrates how this works. Adapters are ligated onto the fragmented DNA (shown in red and blue). On the flowcell, there is a lawn of oligos, short sequences that are complementary to the adapter sequence. A DNA polymerase enzyme creates the complementary sequence of the fragmented DNA. Once the template is washed away, **bridge amplification** can occur, where the strand folds and binds to the other oligo. The DNA polymerase can create the complementary sequence forming a double stranded bridge, which upon denaturing, creates two forms of the template. During sequencing, nucleotides are added and imaged at the same time, a process called **sequencing-by-synthesis (SBS)** (Shendure et al., 2017). During each imaging cycle, a light source excites the fluorescently tagged nucleotide, and the emission is detected.

The benefit of Illumina HiSeq platforms is the ability to sequence two flowcells simultaneously resulting in a large sequencing output. Another benefit of the Illumina sequencers is the high accuracy rate (98%), and the ability to multiplex (Dijk et al., 2014). However, they produce short reads which are difficult to assemble, and can produce reads with GC and Polymerase Chain Reaction (PCR) biases. Additionally, Illumina reads have a tendency for low accuracy towards the 3′ end of the read (Liu et al., 2012).

Library Construction is a difficult, labour-intensive but necessary step prior to sequencing. It requires careful consideration for multiplexing and adapters (Liu et al., 2012), but it is this

[Redacted]

FIGURE 1.11: **Schematic of Next Generation Sequencing.** Figure redacted

stage that has made Next Generation Sequencing so versatile for sequencing many samples in a high-throughput manner.

1.5.4 Common characteristics of Reduced Representation Sequencing

There are many different types of Reduced Representation Sequencing which are compared in Section 1.5.5. The following sections explore the common aspects of reduced representation techniques and how they can be optimised so researchers can obtain the most out of their datasets. Here, the common characteristics of this type of sequencing are explained.

DNA digestion

Reduced representation sequencing methods make use of the properties of restriction enzymes and exploit them to their advantage. Two ways in which restriction enzymes can affect number of cut sites is the GC bias and the cutting frequency. Restriction enzymes can be chosen accordingly if the GC content of the genome is known or if the enzyme is known to be a common or rare cutter to provide a desired number of cut sites (Davey et al., 2011; Andrews et al., 2016).

The type of information obtained from the marker can be selected for by the use of restriction enzymes that are methylation sensitive (Elshire et al., 2011). This means they do not cut methylated cytosines. This can be exploited with a highly repetitive genome such as the pea because repetitive regions are silenced through methylation. By selecting a methylation-sensitive enzyme it will not cut repetitive regions and instead cut gene-rich regions, resulting in a marker panel which may possess more useful information.

The density of the marker panel can be selected for, as some protocols use two different restriction enzymes (Peterson et al., 2012) to increase the marker panel. Additionally, fewer cuts would increase marker density which would improve coverage and allow for greater confidence in SNP calling.

Adaptor ligation and Barcoding

All reduced representation sequencing methods ligate adaptors to the sticky end of the cut site in the construction of libraries (Andrews et al., 2016; Davey et al., 2011). Some methods such as restriction site-associated DNA sequencing (RAD-seq), have an additional library construction stage which adds Y-shaped adaptors to ensure the exclusive PCR amplification of DNA fragments which contain the adaptor (Andrews et al., 2016).

Adaptors can contain unique barcode sequences to help distinguish each sample (Andrews et al., 2016). This allows for multiplexing, meaning that many samples can be sequenced on a lane. Once sequenced, it is possible to demultiplex each sample from the lane due to its unique barcode (Andrews et al., 2016).

Phasing of the sequence is important to accurate base calling - the imaging cycle cannot ascertain the nucleotide position because it is being read out of phase - as shown in Figure 1.12, reproduced from Shendure et al. (2017). Therefore, to reduce the likelihood of phasing, variable length adaptors can be used to ensures even nucleotide distribution during imaging of sequencing-by-synthesis (Davey et al., 2011).

[Redacted]

FIGURE 1.12: **Phasing in Next Generation Sequencing.** Figure redacted.

Adaptors should have barcodes which have at least 2-3bp difference from the others so that sequencing error does not cause confusion when demultiplexing (Davey et al., 2011).

Size selection

In some methods, such as RAD-seq, the digestion of DNA with a restriction enzyme can cause the resulting fragments to be of various lengths and some DNA fragments can be shorter or longer than others. Size selection takes only a portion of these fragments whose size is best suited for sequencing but some forms of reduced representation sequencing, such as, genotyping by sequencing (GBS) do not utilise size selection steps (Andrews et al., 2016).

Sequence data

Depending on the type of sequencing platform used, the resulting read length will differ. The type of sequencing data such as single-end or paired-end sequencing will result in different data output. Paired-end data provides more confidence in genotype calling and greater coverage if the ends overlap, or it can provide greater contig length if they are contiguous (Andrews et al., 2016).

Problems with Reduced Representation Sequencing

The main issues with reduced representation sequencing is the need for high molecular weight and some protocols require greater input DNA to reduce PCR cycles and therefore reduce amplification bias (Andrews et al., 2016).

Allele dropouts can be caused by a mutation in the recognition site meaning that the enzyme does not cut and will not be sequenced. These are null alleles which are alleles that are present but are not identified. This means that for SNPs in null alleles, individuals heterozygous at this locus may appear homozygous, affecting other downstream analyses such as the **fixation index (F_{ST})** by increasing false positives and false negatives (Andrews et al., 2016).

This can be alleviated through the removal of null alleles and this can be identified by abnormally high variance in coverage across samples (Andrews et al., 2016). Filtering loci genotyped across a percentage of samples may also help reduce null alleles (Andrews et al., 2016) and this is also known as missing data. There is a tendency for allele dropout with the increased of length of restriction enzyme recognition site, so choice of restriction enzyme can be important here (Andrews et al., 2016).

PCR duplicates are caused by PCR preferentially amplifying smaller fragments. This means some alleles are more amplified than others. Additionally, with each PCR cycle the error rate can increase, this is because PCR errors may occur and then be further amplified. PCR may favour GC rich content causing PCR bias. This can seriously affect downstream analysis, since heterozygous loci may appear homozygous and PCR errors might be considered a true allele (Andrews et al., 2016).

1.5.5 Reduced Representation Sequencing Overview

There are three main classes of reduced representation sequencing which share similar techniques by using a restriction enzyme to create a series of loci to be sequenced (Davey et al., 2011). Historically, the terms RAD-seq and Genotyping-By-Sequencing (GBS) refer to a specific method or protocol (Andrews et al., 2016). These days, RAD-seq or GBS is used as an all encompassing term to describe all of the these different types of sequencing. In this thesis, for clarity and continuity, reduced representation sequencing is referred to as the all encompassing terminology for these different types of sequencing, and use the terms, RAD-seq and GBS in its historical definition to describe a particular protocol.

Class I: Reduced Representation Libraries (RRLs) or Complexity Reduction of Polymorphic Sequences (CRoPS)

The protocol is quite simple, the genomic DNA is taken and cut with a restriction enzyme, before pooling, size selecting and sequencing (Table 1.3). Complexity Reduction of Polymorphic Sequences (CRoPS), however, is an adaptation of a marker technology called Amplified Fragment Length Polymorphisms (AFLP) which was used before NGS technology was developed (Davey et al., 2011). The small amount of input DNA is helpful but the use of PCR to amplify introduces bias (Van Orsouw et al., 2007).

TABLE 1.3: **The library preparation stages of all three classes of Reduced Representation Sequencing.** Table drawn from data found in a figure in (Davey et al., 2011).

Library Construction Stage	CRoPS	RAD-seq	GBS
Digestion with Restriction Enzyme	✓	✓	✓
Ligate Adaptors	✗	✓	✓
Pooling	✓	✓	✓
Random Shearing	✗	✓	✗
Size Selection	✓	✓	✗
Ligate Adaptors	✓	✓	✗
PCR amplification	✗	✓	✓

Class II: Restriction-site associated DNA sequencing (RAD-seq)

RAD-seq is an attractive method to *de novo* sequencing projects since it is able to produce single-end or paired-end reads to genotype at a reduced cost (Davey et al., 2011). When using paired-end reads, read one can help to identify SNPs but read two can be assembled into 300-600bp RAD contigs (Davey et al., 2011) which can be used to help remove PCR duplication (Kagale et al., 2016) and the GC bias caused by PCR. Table 1.3 shows RAD-seq is quite a complex protocol. It sequences the flanking sequences around the cut site which is useful in building RAD contigs, but not all of the sequence may be informative to variant calling (Davey et al., 2011).

There are also different types of RAD-sequencing. Double digest RAD (ddRAD) uses 2 enzymes to give greater control over reduction and eliminates a random shearing step. This provides more data but this risks greater allele drop out (Peterson et al., 2012). Another type of RAD-sequencing, 2bRAD is where the genomic DNA is cut with a Type 2b restriction enzyme. This too, is a cheap and quick RAD-seq protocol with high allele dropout (Wang et al., 2012).

Class III: Low coverage Genotyping

This class includes sequencing such as Genotyping by Sequencing (GBS). GBS has the added benefit of having a library preparation stage which is easier than RAD-seq (Elshire et al., 2011). However, control over complexity reduction is compromised and subsequently results in more missing data. This missing data can be imputed but this would require the availability of a reference (Davey et al., 2011). Furthermore, this method requires less input DNA than traditional RAD-seq methods (Davey et al., 2011). There is also a 2 enzyme GBS approach (Poland et al., 2012) which provides greater control of the reduction of sequencing.

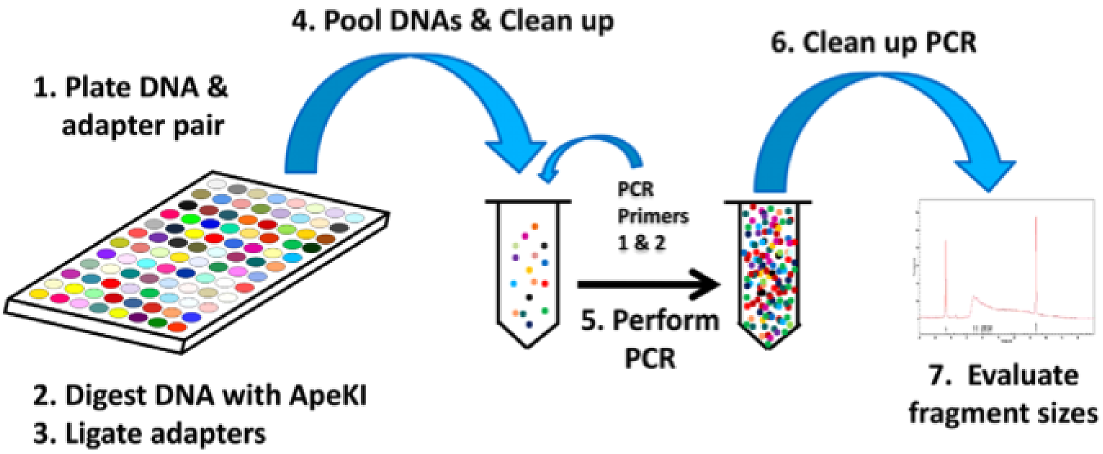


FIGURE 1.13: **Protocol for Genotyping by Sequencing.** Figure taken from Elshire et al. (2011). This is an open access article under the Creative Commons CC0 public domain dedication. Original in colour.

For this project, GBS is a suitable option (Figure 1.13), as an approach that can be used *de novo* or with a reference. When considering important factors such as cost, sample numbers, number of SNPs desired, ease of library preparation and comparing with existing pea GBS data cut with ApeKI (Holdsworth et al., 2017), GBS is an appropriate method to be used in this project.

In summary, reduced representation methods are cheap, efficient and produce a high density of markers which make it beneficial for population genotyping (Davey et al., 2011). Additionally, the use of a reference genome allows imputation of any missing data.

1.5.6 *De novo* Assembly for reference

Assembly involves taking short reads from the sequencing instrument and pieces them together (Pop, 2009). The sequencing read can be assembled into a contig, which in turn can be assembled into a scaffold using scaffolding tools such as SSAKE (Warren et al., 2006), SSPACE (Boetzer et al., 2010) or OPERA (Gao, Nagarajan, and Sung, 2011) and gap closing using tools such as GAPFILLER (Nadalin, Vezzi, and Policriti, 2012). The assembly of sequencing reads into contiguous sequences are usually performed by assembly algorithms, of which, two main types exist, **de Bruijn graphs** and **Overlap-Layout-Consensus (OLC)**.

De Bruijn graphs

To better understand De Bruijn Graphs, a real-life problem provides an excellent illustration into how this algorithm works. In a city called Königsberg, there were 7 bridges, and the people of Königsberg wanted to cross each bridge once and still return to their starting point (Compeau, Pevzner, and Tesler, 2011). In 1735, Leonard Euler came up with a mathematical solution to this problem by making each part of the city a "node" and each bridge an "edge" to form a "graph". In 1946, Nicholas de Bruijn took the Eulerian principal and applied it to superstrings (Compeau, Pevzner, and Tesler, 2011). To resolve the complex problem of genome assembly, the sequencing reads are broken down into **k-mers** (shorter sequences of length *k*) and the Bruijn graph principles are applied to solve this problem.

There are 3 assumptions made when using de Bruijn graphs (Compeau, Pevzner, and Tesler, 2011). Firstly, that each k-mer appears a maximum of once because otherwise it can lead to inaccurate assembly. Secondly, it is assumed that each k-mer is present in the genome because this could introduce gaps into the assembly and thirdly, that there are no errors in the k-mer because otherwise this could cause a bulge in the de Bruijn graph, which can be a common occurrence as a result of sequencing errors (Compeau, Pevzner, and Tesler, 2011). In genome assembly, sometimes these assumptions are not met and an explanation of the consequences of not meeting these assumptions is outlined next. K-mers can occur more than once, particularly in a repetitive genome such as pea, this is called **k-mer multiplicity**, which can sometimes cause repeat collapse within the final assembly (Compeau, Pevzner, and Tesler, 2011). Gaps in the sequence will mean not all k-mers are present in the genome, this will result in a less contiguous and more fragmented genome - this issue can be overcome by either more sequencing or by scaffolding the contigs (Compeau, Pevzner, and Tesler, 2011).

Examples of de Bruijn graph assemblers are: ALLPATHS (Butler et al., 2008), SOAPdenovo2 (Luo et al., 2012), ABySS (Simpson et al., 2009), Velvet (Zerbino and Birney, 2008).

Overlap-Layout-Consensus (OLC)

The alternative algorithm to the de Bruijn graph is the overlap-layout-consensus algorithm. The overlap-layout-consensus algorithm is a three step process: first the algorithm overlaps the reads, this helps to layout the reads onto a graph from which the consensus sequence can be obtained (Li et al., 2012). This algorithm is less susceptible to sequencing errors and other common issues such as heterozygosity and coverage, however, this approach can be considered to be computationally complex (Wang et al., 2018).

Examples of OLC assemblers are: Celera Assembler (Denisov et al., 2008), CAP3 (Huang and Madan, 1999) and PCAP (Huang et al., 2003).

1.5.7 GBS Assembly

Section 1.5.5 described the GBS library protocol as well as the advantages and disadvantages of using this sequencing method. The assembly of GBS reads using bespoke GBS assembly tools is outlined below. There are a variety of tools available to assemble GBS reads such as Stacks (Catchen et al., 2011), Rainbow (Chong, Ruan, and Wu, 2012), Universal Network-Enabled Analysis Kit (UNEAK) (Lu et al., 2013), Tassel (Bradbury et al., 2007) and PyRad

(Eaton, 2014) to name but a few. The choice of tool should be selected based on some important considerations which are outlined below (Andrews et al., 2016; Davey et al., 2011; Mastretta-Yanes et al., 2015).

The genome is a key consideration because some tools fail to detect paralogous sequences that are induced by a history of duplication in the genome's history and other tools do not perform well with highly heterozygous or repetitive genomes (Mastretta-Yanes et al., 2015). In addition to this, the availability of a reference sequence should be considered (Davey et al., 2011) because some tools only work *de novo* or with a reference sequence. The sequencing data itself is another consideration (Davey et al., 2011) because some tools only work on single-end or paired-end data and the performance of some tools is affected if there is low coverage (Andrews et al., 2016; Mastretta-Yanes et al., 2015). Furthermore, the number of samples is also key consideration (Andrews et al., 2016) because it can cause performance issues if the tool is not fast or memory efficient (*Stacks denovo map*).

Here in Table 1.4 is a summary of a few of the most widely available tools and each tool is evaluated for their strengths and weaknesses.

TABLE 1.4: **A comparison of GBS assembly tools.**

Software	*De novo*	Reference	SE reads	PE reads	Pros	Cons	Citation
Stacks	✓	✓	✓	✓	• Calls genotypes with maximum likelihood model • Can assist with phylogeography	• Struggles with PCR errors and over-merges paralogs • Assumes organism is a diploid.	(Catchen et al., 2011)
UNEAK	✓	✗	✓	✓	• Works well on high and low coverage	• Designed for *de novo*	(Lu et al., 2013)
PyRAD	✓	✗	✓	✓	• Allows for indel variation and incomplete overlap • Parallelized and Fast • Good for large datasets • Low memory usage	• It may exclude loci with low coverage	(Eaton, 2014)
Rainbow	✓	✗	✗	✓	• Fast and memory efficient • Works well with highly heterozygous genomes	• Does not work with SE reads or reference genome.	(Chong, Ruan, and Wu, 2012)

A recent publication (Torkamaneh, Laroche, and Belzile, 2016) has trialled several pipelines on GBS data cut with ApeKI using the Illumina platform for soybean. Torkamaneh, Laroche, and Belzile (2016) trialled both *de novo* and reference methods and found that reference sequence-based methods produced more SNPs. The authors found that Tassel v1 produced the least accurate SNP calls (due to **paralogs** - genes related from duplication within the genome's history) and Fast-GBS produced the most SNPs and was the most accurate. Stacks produced the least number of SNPs for *de novo* and reference based methods due to inefficient demultiplexing and had the most missing data for reference-based methods. From this paper, it is possible to draw the conclusion that for *de novo* methods, UNEAK should be used because it calls more SNPs and is marginally more accurate (Torkamaneh, Laroche, and Belzile, 2016) and unlike other tools; there is no need for input from a transcriptome or other forms of sequencing data (Lu et al., 2013). However, it trims the reads to 64 bp before calling SNPs which means a third of the read is discarded.

The *de novo* assembly tool, Stacks (Catchen et al., 2011), is capable of taking any type of reduced representation reads, and assembles them into "stacks" before identifying loci and calling the genotype. Stacks operates differently to UNEAK by keeping the reads flush (all the same length) and does not trim the read. Instead, it discards entire reads that are considered of low quality, resulting in fewer reads of greater quality. Paris, Stevens, and Catchen (2017) have shown that Stacks may be best used *de novo* and then used to integrate the reference afterwards. One of the issues with Stacks is the over-merging of paralogs (Mastretta-Yanes et al., 2015) and whilst studies suggest that there has not been a recent duplication event in pea due to synteny and collinearity studies with **Medicago truncatula** (Section 1.3), Stacks is limited by its ability to handle a high number of samples in the catalog (*Stacks denovo map*). Alternatively, pyRAD (Eaton, 2014) is another *de novo* assembly tool which works for GBS and for reference-based methods, Fast-GBS would be the best (Torkamaneh, Laroche, and Belzile, 2016).

1.5.8 Genome-wide Association Study (GWAS)

Chapters 3 and 4 will produce sequencing and phenotypic data for a variety of pea accessions, enabling a Genome-wide Association Study (GWAS) to be performed.

GWAS statistically associate traits to SNPs identified in the sequencing data. The SNPs will act as genetic markers to help identify candidate regions that maybe associated with a trait of interest (Ayling, 2012). These **Genotype-to-phenotype (G2P)** associations may also enable the prediction of this phenotype being found in other closely related accessions. Furthermore, when the genotype-to-phenotype associations are combined with a reference genome, it might be possible identify the causative gene of the trait through a direct association which is rare or to find a locus indirectly associated with the gene which could be some distance away from the causative gene itself (Bush and Moore, 2012).

Previously, **Quantitative Trait Loci (QTL) mapping** has been used to identify genetic loci involved in a trait of interest using linkage analysis (Zargar et al., 2015). In other words, it identifies regions of co-segregation with a trait (Korte and Farlow, 2013) using a marker panel and a mapping population. It is limited to looking at allelic diversity in the parents and recombination limits its mapping resolution (Korte and Farlow, 2013).

GWAS overcomes the issues with QTL mapping but also has its own difficulties. Since GWAS statistically associates markers with a phenotype from many individuals it can be used to provide insights into the genetic architecture of trait. Simple genetic architecture is ideal for GWAS, whereby a small number of loci have high penetrance (effect of the gene). Whereas, complex genetic architecture, where many loci have a small penetrance or many rare variants have high penetrance is not ideal for GWAS (Korte and Farlow, 2013). GWAS can help to identify candidate parents for QTL analysis, identify candidates for mutagenesis or can complement QTL mapping (Korte and Farlow, 2013).

Considerations of GWAS

When performing a GWAS, there are many considerations that must be made (Korte and Farlow, 2013; Bush and Moore, 2012; Ayling, 2012). Firstly, the sample size of the GWAS is an important factor because greater sample sizes can increase the statistical power of associations being made (Korte and Farlow, 2013; Bush and Moore, 2012; Ayling, 2012). Incomplete genotyping is an important consideration because missing data is a common problem, however, this can be addressed by use of imputation to help infer SNPs that are missing and can tackle low coverage issues in sequencing (Korte and Farlow, 2013; Bush and Moore, 2012; Ayling, 2012). Genetic heterogeneity can be an important issue (Bush and Moore, 2012) as it can cause non-causative markers to be associated with the phenotype (Korte and Farlow, 2013). Additionally, the genetic background of the organism is a key consideration and can be accounted for by the use of Mixed Models which can take into account population structure and by reducing false positives using False Discovery Rate corrections (Korte and Farlow, 2013; Bush and Moore, 2012).

In order to perform a GWAS, requirements include a high-density marker panel and a large enough sample size to obtain statistically significant associations (Smýkal et al., 2012). GWAS is a method best suited to the investigation of common variance with low penetrance (Bush and Moore, 2012). Rare traits are hard to identify because sufficient sample size is needed to find enough samples that carry this trait to obtain sufficient allele frequency (Bush and Moore, 2012; Visscher et al., 2012).

Linkage Disequilibrium (LD)

With **Genome Wide Association Studies (GWAS)**, it is important to consider the **linkage disequilibrium (LD)** which is defined as the non-random association of loci (Slatkin, 2008). Linkage disequilibrium is an important consideration with GWAS because genes positioned close together are more likely to be inherited together as a result of high LD, whereas, recombination causes markers to end up in linkage equilibrium (Bush and Moore, 2012). Linkage disequilibrium can indicate the marker density required for GWAS because with low LD decay rates fewer markers need to be obtained but at higher decay rate, a greater density of SNPs is necessary (Ayling, 2012). LD decay is affected by the size of the population, the number of chromosomes that the SNPs occur on and the number of generations (Bush and Moore, 2012).

Linkage Disequilibrium is measured by two metrics, D′ or r^2. D′ is a measurement of recombination between markers, where 0 denotes complete linkage equilibrium (high recombination) and 1 signifies complete linkage disequilibrium (no recombination) (Bush and Moore, 2012). Another measure is correlation, r^2, whereby high r^2 indicates that two 2 SNPs are observed together (Bush and Moore, 2012).

GWAS Association Tests

GWAS can be analysed in different ways: Single Locus Analysis, Population Stratification adjustment, Multiple Testing correction and Multi-Locus Analysis (Bush and Moore, 2012). These association tests are described below.

Analysis 1: Single Locus Analysis

Single locus analysis considers each SNP independently of all other SNPs and tests for statistical association, such analysis can be performed through a **Generalised Linear Model (GLM)**, one example of this GLM is **Analysis of Variance (ANOVA)** (Bush and Moore, 2012). Another example of a GLM tool is GenABEL (Aulchenko et al., 2007). This is the most basic of tests available and does not consider *a priori* information such as the structure of the population being investigated or the spuriousness of statistics, unless performed by the user.

Analysis 2: Adjusting for population stratification.

Population stratification is the differences in allele frequencies that are not caused by association with the trait of interest but by ancestry (Freedman et al., 2004). Ancestry can be measured using tools such as STRUCTURE (Pritchard, Stephens, and Donnelly, 2000). The results can be used to exclude samples or to adjust for a covariate (in this case ancestry) (Bush and Moore, 2012) before applying GLM methods. Alternatively, **Mixed Linear Models (MLM)** are methods that take into account population structure (sometimes referred to as "Q") and genetic differences in the relatedness of individuals (sometimes referred to as kinship, "K") (Zhang et al., 2010). An alternative is the genomic inflation factor, that corrects for the inflation of test statistics caused by population structure (Zeng et al., 2015; Yang et al., 2011). It adjusts the test statistic by dividing each Chi-Squared statistic by the genomic inflation factor. A genomic inflation factor of more than 1 suggests population structure or genotyping error (Zeng et al., 2015).

Analysis 3: Corrections for multiple testing

Correcting for multiple testing is important since there are many statistical associations being made, any errors made can compound, therefore, GWAS is likely to suffer from over-inflation of false positives (Lipka et al., 2012). One way to correct for multiple testing is to use the Bonferroni correction, which takes a p-value of 0.05 and divides it by the number of statistical tests (in this case the number of SNPs) (Bush and Moore, 2012). It assumes that all statistical tests are independent which is not necessarily the case when considering LD. A different correction method is called the **False Discovery Rate (FDR)** which can be used to measure how many of the results are false positives. The Benjamini and Hochberg FDR method (Benjamini and Hochberg, 1995) determines both an estimated false discoveries rate and provides an estimated number of true significant associations. An alternative correction method is permutation testing which is a computationally expensive method of finding the distribution of statistical tests when the null hypothesis is true (Bush and Moore, 2012). The reason for this computational cost is because it does so by recharacterizing the phenotype of one individual as belonging to another multiple times, in order to remove the genotype-to-phenotype relationships (Bush and Moore, 2012). Another correction method called genome-wide significance is a threshold decided by observing the changes in linkage disequilbrium across the genome to detect independent regions and it is this threshold that is used to correct for statistical tests (Bush and Moore, 2012).

Analysis 4: Multi-locus Analysis

Multi-locus Analysis is a computationally and statistically complex problem. It involves reducing the marker panel to remove redundancy. There are two ways to do this: restrict SNP combinations to those with biological pathways (which requires prior knowledge) or to use single SNP analysis and a threshold to obtain the reduced subset (Bush and Moore, 2012).

Notable Works

There have been several key works contributing to advancement of GWAS in a number of organisms. In rice (*Oryza sativa*), GWAS has been used to explore 34 agronomically important traits relating to flowering, morphology, yield, stress and quality (Zhao et al., 2011). It used 44,100 SNPs and 413 accessions from 82 countries and selected across 4 subgroups of *O.sativa*. This was a landmark project, because not only did it seek to understand more about the genetic architecture in rice it also released the seeds as a public resource, meaning that others do not need to spend money on sequencing.

In elite lines of rice, GWAS in combination with genomic selection was used for the first time in rice to find new QTLs for plant height and flowering time (Spindel et al., 2015). Another GWAS on elite lines, used over 71,000 GBS markers and found 52 QTLs for a variety of agronomic traits to improve breeding in rice (Begum et al., 2015). Huang et al. (2010) performed a GWAS on 517 landraces of rice *Oryza sativa indica*, using approximately 3.6 million SNPs across 14 agronomic traits. Together, the work done in rice has been key to improving breeding in rice and to better understand the allelic richness in the organism itself.

Maize (*Zea mays*) has also had many successful genome-wide association studies. One GWAS study (Suwarno et al., 2015) has identified candidate genes for vitamin A in maize including well known alleles such as lycopene epsilon cyclase (*LcyE*) (Harjes et al., 2008) and beta-carotene hydroxylase 1 (*crtRB1*) (Yan et al., 2010) and have been bred into elite lines of maize resulting in increased vitamin A content through biofortification within the human diet (Harjes et al., 2008; Yan et al., 2010). This shows how GWAS have potential to inform breeding practices.

Much work has been done on the plant architecture of maize, one GWAS study found *liguless 1* and *2* (*lg 1 and 2*) have been found to correspond with leaf angle, a trait that can help capture more sunlight and with potential to increase photosynthesis and therefore has potential yield gains (Tian et al., 2011). GWAS has been performed on leaf architecture in maize (Pan et al., 2017) and have identified QTLs for ear architecture and grain size (Xiao et al., 2016; Dell'Acqua et al., 2015). This information in maize shows how genome-wide association studies can be performed on plant architecture and can help to provide information that could be used to improve food security.

It is evident how each key project builds on the foundation of another, providing vital genomic resources which have been used to hone GWAS methods and inform the scientific community of the advantages and limitations of GWAS in practice. Section 4.1.2 explores in great detail key genome-wide association studies in pea by Annicchiarico et al. (2017) and Holdsworth et al. (2017) and compares and contrasts their methodology.

Markers, traits and loci in GWAS

Improvements in sequencing outputs and cost have resulted in the use of genomics to identify markers to assist breeding (Varshney, Graner, and Sorrells, 2005). GWAS produces Genotype-to-Phenotype associations which may assist in crop improvement because they have potential to allow breeders to select for desirable phenotypes (Varshney, Graner, and Sorrells, 2005).

The output of GWAS may also have potential much further downstream to assist in translational genomics in agriculture (TGA). TGA is the use of genomics to produce better products and whilst much work has occurred in this field with rice and maize, there has also been some initial efforts have occurred in pea (Varshney et al., 2015).

The output of GWAS also has potential to assist with **Marker Assisted Selection (MAS)** and Genomic Selection (GS). MAS makes use of genetic markers and the phenotypes associated with these markers to enable better informed breeding of plants with desirable traits (Ayling, 2012). Genomic selection utilises markers across the genome to select for desired genotypes according to the Genomic Estimated Breeding Value (GEBV) which is a prediction of statistical likelihood of an allele being present at any given loci (Ayling, 2012) and which may be calculated first on a training population before extending this to the breeding population (Varshney et al., 2015).

1.6 The novel contribution of this research project

This project aims to develop a core collection of the John Innes *Pisum* Collection in Chapter 2 with a view to exploring domestication. This project also aims to phenotype morphological traits in Chapter 3, to create a publicly accessible pea genome reference and sequence each accession in the core collection with GBS and use this to create a Genome-wide Association Study in Chapter 4.

This project is particularly challenging since the genome of the pea is large and repetitive. There is currently no reference genome released yet. Assembling a new reference and publicly releasing it, would be of great use to the pea community.

Genome Wide Association Studies have their limitations and rely on statistical power, so a large sample size is needed. This thesis addresses this by obtaining a core collection of 350 individuals in Chapter 2. The data generated from this will be novel, and will help identify markers associated with certain traits.

Developing a new tool in image analysis in Chapter 3 will be of great benefit to the community as a fully automated multi-organ image analysis tool that uses the Plant Trait Ontology. The results of this will be a quantification of phenotypes for the pea. This novel contribution may help crop improvement.

Chapter 2

Developing a core collection

2.1 Introduction

Core collections are proposed by researchers to be a set of accessions formed from a gene bank. Researchers design core collections to be an appropriate sample size (less than 10% of a gene bank and less than 2000 accessions), with minimal repetitiveness and maximum genetic diversity (Frankel, 1984). Studies have found that core collections can capture approximately as much as 90% of genetic diversity of the original gene bank (Charmet and Balfourier, 1995). Since there can be many germplasm accessions within a gene bank, a core collection represents a more manageable sample size to analyse (Odong et al., 2013).

Core collections are created by first obtaining a **domain**, in other words, a set of germplasm accessions. Then the size of the core collection is decided, in other words, what percentage of the germplasm accessions (domain) will make up the core collection. This core collection in then divided into a number of groups. The number of entries to be found in each group is decided, and then finally, the entries are selected to enter the core collection(Hintum et al., 2000).

When creating a core collection, you must consider the size of the germplasm collection under investigation and have prior understanding of the state of current knowledge regarding the organism's genetic diversity or the accessions of key priority (Hintum et al., 2000).

Addressing the concerns listed above can allow for the creation of a potentially useful core collection which retains as much genetic diversity as possible.

2.1.1 Core collection in pea

Core collections have been made in many crop plants such as barley (Muñoz-Amatriaín et al., 2014), wheat (Balfourier et al., 2007), rice (Yan et al., 2007) and soybean (Oliveira et al., 2010) to name but a few. Core collections in pea (Smýkal et al., 2012) have been developed in 5 of the 16 major germplasm collections: United States Department of Agriculture (USDA) in USA, Australian Temperate Field Crop (ATFC) in Australia, John Innes Centre (JIC) in UK, Institut National de la Recherche Agronomique (INRA) in France, AGRITECH (CZE) in the Czech Republic. Many studies have been performed on these core collections investigating a number of phenotypes. For example, in the USDA collection, seed nutrients (Kwon et al., 2012) and seedling root architecture (McPhee, 2005) have been investigated in this core collection and one study explored 25 traits such as disease resistance traits, seed type and flower colour (Cheng et al., 2014).

In pea, there is an attempt to collate a world core collection using molecular markers and storing phenotypic information and other accession data in a database for world-wide use (Smỳkal et al., 2008). If achieved, this will be an important resource for users as it will allow greater scope for diversity because accessions found in different germplasms will be available for all to use.

2.1.2 Core collection methods

There are several methods designed to develop core collections. These include random or random stratified strategies as well as methods based on location or genetic markers to explore allelic representativeness or allelic diversity respectively (Thachuk et al., 2009). Each method holds advantages and disadvantages to developing a core collection that holds the most genetic diversity with the least amount of repetitiveness.

Random Strategy

Choosing accessions at random through a random number generator without replacement is one method which is truly random. Other methods are more systematic such as obtaining every nth accession (Zeuli and Qualset, 1993). Whilst the easiest and quickest option, the resulting core collection is not based on any *a priori* information that will make it more relevant to a study.

Random Stratified Strategies

Random Stratified Strategies first group accessions by origin, passport, phenotype or genotype data. These strategies then allocate the numbers of accessions in the groups according to the proportion (P-strategy) or the logarithmic proportion (L-strategy) required (De Beukelaer et al., 2012). The L-strategy has been used in a forming a USDA barley core collection (Muñoz-Amatriaín et al., 2014). These methods ensure the sample size is proportionate to the weighting of groups in a domain (Zeuli and Qualset, 1993) and are particularly useful in collections that may contain high genetic redundancy (Groth and Roelfs, 1987).

Another stratified method, the maximisation strategy (M-strategy) uses markers to increase allelic richness aiming to capture the most number of markers at each loci (Thachuk et al., 2009) and progressively iterates to ensure allelic richness and maximise diversity. It does make the assumption that maximising marker diversity will correlate with maximum allelic diversity at the loci of interest (Groth and Roelfs, 1987). These are methods of the most interest to geneticists due to the resultant core collection being rich allelic diversity (Balfourier et al., 2007).

Additionally, the D-method determines how many accessions are in each cluster and this is designed to increase diversity (Thachuk et al., 2009) through genetic distance. The most diverse clusters contribute more accessions (De Beukelaer et al., 2012) to the core collection, resulting in a diverse collection using genetic distance. This is of most interest to breeders as the resultant core collections are more representative (Balfourier et al., 2007).

Tools for developing a Core Collection

Focused Identification of Germplasm Strategy (FIGS) uses information such as geographic location and environmental profiles in order to predict the likelihood of other accessions in the gene bank having a similar target trait (Bari et al., 2012). This is based on the assumption that similar environments are probably affected by similar selection pressures and proposes that the resultant accession will have a similar phenotype or genotype (Bari et al., 2012). In simpler terms, FIGS mines the gene bank to find accessions with similar traits. This makes it easier to select accessions for a core collection based on location. This has been a beneficial contribution to plant science, since large germplasm collections are difficult to access and screen and this makes the identification of agronomically and agriculturally useful traits difficult. The FIGS tool does have its own limitations because it needs accurate geo-referencing for its accessions (Endresen et al., 2012). It can be argued that FIGS is not capable of finding more genetically diverse accessions and instead only finds similar varieties, due to its underlying assumption that desirable alleles are more likely to be found in similar environments.

There are other tools that do not consider location but instead use genetic markers. Core Hunter is a tool that uses to genetic markers to collate a core collection and comes in three versions, Core Hunter (I, II, III) (Thachuk et al., 2009; De Beukelaer et al., 2012; De Beukelaer, Davenport, and Fack, 2018). Core Hunter I can obtain a core collection based on allelic richness, allelic representativeness or a combination of both (Thachuk et al., 2009). When tested against other tools, Core Hunter provides a core collection just as good or better for a number of genetic measures (Thachuk et al., 2009). Core Hunter III is an improvement by being quicker (De Beukelaer, Davenport, and Fack, 2018). PowerCore, another tool that uses genetic markers, uses the M-strategy to provide diversity in as few accessions as possible (Kim et al., 2007). However, this tool is not cross-platform and is outperformed by Core Hunter (Thachuk et al., 2009). Genocore claims to be a memory-efficient way to obtain the smallest possible and most representative core collection with the greatest coverage (Jeong et al., 2017) using genetic markers and shows comparable results to other tools in terms of genetic measures.

Here, in Chapter 2, a random strategy aiming to be as statistically representative of the domain as possible based on seed weight is outlined.

2.2 Materials and Methods

2.2.1 Creating the domain from the germplasm

The John Innes Centre *Pisum* contains 2776 accessions. These were filtered to remove group 3 (the unknown group) and to keep the accessions that were seed harvested - seed was harvested from the individual plant that was DNA typed and this filtering resulted in 2443 accessions. These accessions formed the domain.

2.2.2 Creating groups from the domain

Previous work (Jing et al., 2010) has used RBIP markers to characterise the JIC germplasm into 3 main groups and 16 sub-groups based on seed weight and the characterisation of this core collection is based on this work. A sample of 350 accessions were taken (approximately 14% of the collection - allowing room for sample drop out). These accessions were grouped in accordance to Jing et al. (2010). Figure 2.1 shows the main groups formed by this work. It is important to note, the differences in nomenclature. Work in this paper characterises the germplasm into 3 groups, with an additional unknown group (group U). The John Innes Centre database records peas into 4 groups, with group 3 being classed as the unknown group. For clarity, from herein, this thesis recognises the 4-group nomenclature as recorded by the John Innes Centre database.

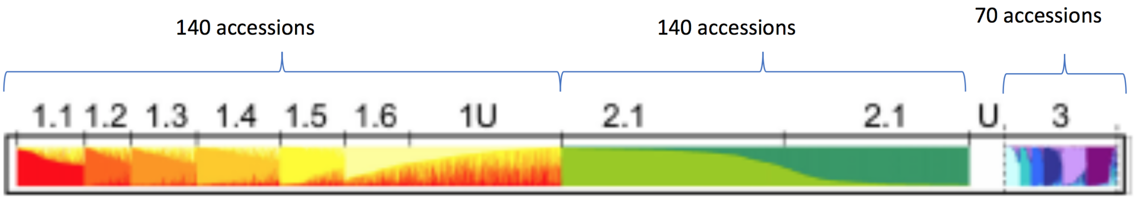


FIGURE 2.1: **Accessions to be added to each group to form a core collection.** Adapted figure from (Jing et al., 2010) and cropped. This shows the Structure analysis of the JIC collection using RBIP markers based on seed weight. The figures above show how many accessions from each group are in this collection. This is an Open Access article under Creative Commons Attribution License (`http://creativecommons.org/licenses/by/2.0`). Original in colour.

2.2.3 Randomly selecting accessions to be added to the group

A Random Accession generator was developed in Java to randomly select accessions from the domain, ensuring 20 accessions were selected from each subgroup of Group 1, 70 accessions were selected from each subgroup of Group 2 and 10 accessions selected from each subgroup of Group 3. For each group, accessions were chosen evenly across all sub-groups, in order to capture the genetic diversity of this gene bank. This brings the total number of peas in the randomly chosen selection to 350. Figure 2.1 shows the distribution of accessions across the groups.

2.2.4 Statistical analysis: ensuring selection is statistically representative of the domain

One sample Kolmogorov-Smirnov test: random selection

For the randomly chosen selection, a one sample Kolmogorov-Smirnov test based on seed weight on the randomly chosen selection was performed to see if it is normally distributed, using SPSS version 22 (IBM Corp, 2013). The null hypothesis was the selection is normally distributed.

One sample Kolmogorov-Smirnov test: domain

This was also then performed on the JIC domain, also using SPSS version 22 (IBM Corp, 2013). The null hypothesis was that the JIC domain is normally distributed.

Two sample Kolmogorov-Smirnov test: domain vs random selection

For both the JIC domain and the selection, a two-sample Kolmogorov-Smirnov test based on seed weight was performed to see if the selection was representative of the JIC domain. For selection subgroups which are not representative of the JIC domain, (i.e. reject null hypothesis p <0.05), the random Java generator was re-used and the statistics tests re-performed until all tests retained the null hypothesis. This formed the core collection.

2.2.5 Stock availability

In some cases, the accessions chosen were not available so they were substituted with another accession based on nearest seed weight.

2.2.6 Sowing (and chipping)

Some wild material required chipping (drilling a hole opposite the embryonic axis) to ensure water penetration and germination in the wild material as shown in Figure 2.2.

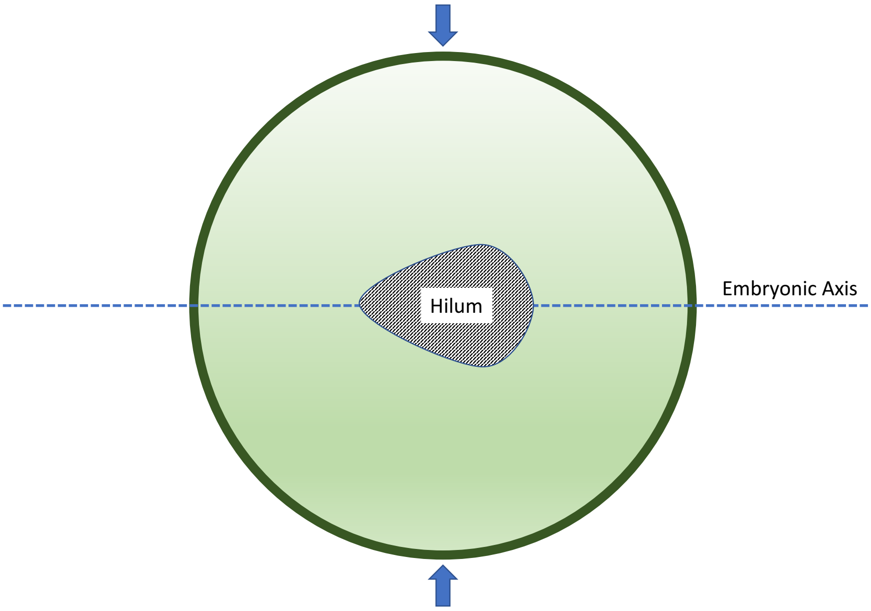


FIGURE 2.2: **Drilling (chipping) a pea seed**. The arrows denoted in blue are where the chip can be placed. The hilum denoted in shaded black (the abscission scar from the pea being separated from the funicle of the pod) points towards the embryonic axis denoted as a dashed line.

A drill with a moving clamp was used to chip the peas with the aid of a metal guide if needed as shown in Figure 2.3. All peas were sown in triplicate in glasshouse and field environments in 2015.

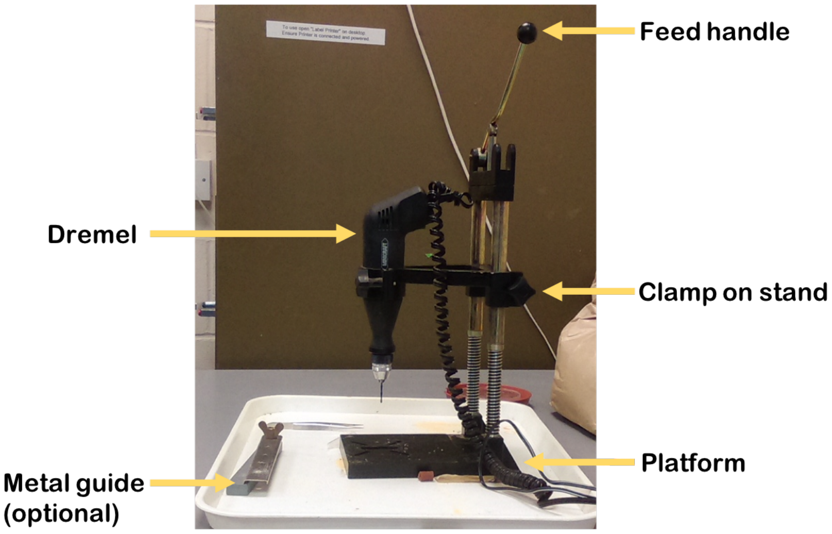


FIGURE 2.3: **Equipment setup for drilling peas**. A dremel placed on a moving clamp was used to chip wild peas. Metal guide was used to hold the pea in place during drilling.

2.3 Results

2.3.1 Investigating the distribution of the JIC collection

A one sample Kolmogorov-Smirnov test on the JIC domain (for accessions that held a seed weight) was used to see if it is normally distributed, where the null hypothesis was: the JIC domain is normally distributed.

The results show (Table 2.1) show that Groups 1 and 2 are similar in seed weight. However, Group 4 is almost half the weight in comparison. Jing et al. (2010) indicates that Group 1 are mainly landrace material, Group 2 are cultivars and Group 4 as wild. The observation found in Table 2.1 is linked to the literature which suggests that wild material has a lower seed weight than cultivated varieties (Weeden, 2007). Whilst, Groups 1 and 2 are similar in seed weight, Group 2 (cultivars) is larger than Group 1 (landraces).

Table 2.2 shows that the JIC domain is not normally distributed since p <0.05 the null hypothesis must be rejected. This one-sample Kolmogorov-Smirnov test compares against a reference distribution and this shows the domain does not follow this pattern.

TABLE 2.1: **Background distribution of the JIC domain.** Where n denotes number of samples that hold a seed weight. Out of 2444 accessions in the domain, 2312 had a seed weight recorded.

Group	Mean ± SD	n	Subgroup	Mean ± SD	n
1	214.6 ± 70.346	1267	1.1	218.92 ± 99.757	157
			1.2	223.58 ± 61.897	108
			1.3	210.87 ± 65.850	152
			1.4	209.04 ± 59.498	194
			1.5	215.41 ± 58.220	148
			1.6	221.83 ± 72.546	120
			1.7	212.04 ± 68.412	388
2	240.84 ± 65.427	791	2.1	248.53 ± 65.75	424
			2.2	232.92 ± 63.647	380
4	120.22 ± 56.885	254	3.1	153.81 ± 43.830	36
			3.2	98.21 ± 27.263	19
			3.3	102.88 ± 33.711	34
			3.4	107.14 ± 53.964	43
			3.5	156.33 ± 72.928	61
			3.6	77.85 ± 11.816	47
			3.7	131.0 ± 39.019	14

TABLE 2.2: **One-sample Kolmogorov-Smirnov test of the JIC domain.** The John Innes domain and the test statistic (Kolmogorov-Smirnov test), p-value, and Absolute Extreme Difference (AED).

Collection	Test statistic	p-value	AED	Decision
JIC collection	0.046	0	0.046	Reject Null Hypothesis

2.3.2 Investigating the distribution of the randomly selected core collection

For the randomly chosen selection, a one sample Kolmogorov-Smirnov test on the randomly chosen selection was performed to see if it is normally distributed. The null hypothesis was the total selection is normally distributed.

The results show (Table 2.3) that the total selection is very similar to the mean seed weight seen in the JIC collection (Table 2.1).

TABLE 2.3: **Distribution of the Total Selection**. Where n denotes number of samples that hold a seed weight. Only 344 of 350 accessions held a seed weight.

Group	Mean ± SD	n	Subgroup	Mean ± SD	n
1	209.75 ± 58.248	138	1.1	204.65 ± 58.064	20
			1.2	228 ± 61.897	20
			1.3	195.11 ± 40.883	19
			1.4	192.65 ± 59.282	20
			1.5	207.65 ± 64.89	20
			1.6	220.2 ± 66.819	20
			1.7	219.63 ± 49.985	19
2	240.72 ± 71.575	138	2.1	238.79 ± 70.886	68
			2.2	242.59 ± 72.701	70
4	116.66 ± 51.476	68	4.1	150.800 ± 55.467	10
			4.2	92.3 ± 14.682	10
			4.3	95.1 ± 17.515	10
			4.4	101.44 ± 62.234	9
			4.5	164.7 ± 61.364	10
			4.6	80.111 ± 18.731	9
			4.7	127.0 ± 43.4255	10

The results also show that the null hypothesis can be can retained, meaning the total selection is normally distributed (Table 2.4).

TABLE 2.4: **One-sample Kolmogorov-Smirnov test of the total selection**. The total selection and the test statistic (Kolmogorov-Smirnov test), p-value, and Absolute Extreme Difference (AED).

Selection	Test statistic	p-value	AED	Decision
Total Selection	0.039	0.2	0.039	Retain H^0

2.3.3 Comparing the JIC collection with the randomly selected core collection

For both the JIC collection and the selection, a two-sample Kolmogorov-Smirnov test was performed to see if this selection was representative of the JIC domain. The null hypothesis was that the total selection was representative of the JIC collection. Table 2.5 shows that the null hypothesis can be retained and therefore this selection is representative of the JIC domain.

TABLE 2.5: **Two sample Kolmogorov-Smirnov Test of the JIC domain against the Total Selection.** The John Innes domain and the total selection with the test statistic (Kolmogorov-Smirnov test), z-value, p-value, and Absolute Extreme Difference (AED).

Collection	Selection	z-value	p-value	AED	Decision
JIC	Total selection	1.106	0.173	0.064	Retain Null hypothesis

Then for each subgroup of the randomly chosen selection, each subgroup was compared against its corresponding group of the JIC collection. The null hypothesis was that each subgroup of the total selection was representative of the corresponding group in the JIC collection. The results show that in each case, the total selection subgroup is representative of the JIC collection group (Table 2.6).

TABLE 2.6: **Two-sample Kolmogorov-Smirnov Test of JIC domain groups against the Total Selection subgroups.** The subgroups of the selection are statistically representative of the domain groups.

Group	Subgroup	Z-value	p-value	AED	Decision
1	1	0.638	0.81	0.144	Retain H^0
	2	0.886	0.413	0.2	Retain H^0
	3	0.802	0.541	0.185	Retain H^0
	4	0.726	0.668	0.164	Retain H^0
	5	0.915	0.373	0.206	Retain H^0
	6	0.717	0.682	0.126	Retain H^0
	7	0.788	0.564	0.182	Retain H^0
2	1	0.695	0.72	0.088	Retain H^0
	2	0.642	0.805	0.08	Retain H^0
4	1	1.316	0.062	0.424	Retain H^0
	2	1.209	0.108	0.39	Retain H^0
	3	1.033	0.236	0.333	Retain H^0
	4	1.211	0.106	0.411	Retain H^0
	5	1.336	0.056	0.431	Retain H^0
	6	1.321	0.061	0.448	Retain H^0
	7	0.955	0.321	0.308	Retain H^0

2.3.4 Ensuring the selection is representative

The above results are from the random number generator, which was re-performed until all selection subgroups were representative of the JIC collection and until all retained the null hypothesis. This formed the statistically representative core collection. After the core collection was found to be statistically representative of the JIC list, the accessions were obtained from the *Pisum* germplasm. However, three accessions were not in stock, so the closest available seed weight in the group was chosen to substitute these accessions. These were as follows:

JI 2292, Group 1.2, Seed weight:331mg $\rightarrow$ JI 2570, Group 1.2, Seed weight:331mg

JI 2539, Group 4.2, Seed weight:96mg → JI 262, Group 4.2, Seed weight:88mg

JI 2215, Group 4.7, Seed weight:200mg → JI 1870, Group 4.7, Seed weight:182mg

This forms the final list of accessions for the core collection.

2.4 Discussion

Here in Chapter 2, a core collection of pea based on seed weight was developed, since seed weight is one of this study's phenotypes of interest. Currently, most core collections represent other germplasms world wide in the USA, Australia, France and the Czech Republic (Smýkal et al., 2012). A core collection has been made in pea at the JIC, according to Smýkal et al. (2012) but this is difficult to find. This core collection was based on a phenotype of interest and therefore made for the additional studies to be carried out in this thesis.

The approach taken to develop this core collection is a random method with 350 accessions from the stratification obtained by RBIP markers from Jing et al. (2010) in a landrace, cultivar and wild split of 140:140:70. The JIC *Pisum* germplasm contained 2776 accessions, the domain contained 2444 accessions and whilst the core collection was aimed to be as representative as possible, but 6 accessions did not have a seed weight, meaning the remaining 344 accessions were representative and due to issues in the availability in germplasm stock, 3 were replaced with the nearest available seed weight. This core collection is largely representative of the entire germplasm collection and is a good sample size. A core collection should be 10% of a germplasm collection (Charmet and Balfourier, 1995), and this core collection contains 14% of the domain, to make allowances for sample drop out in further downstream analysis, for example, during phenotyping (because samples are unable to be phenotyped) or during genotyping (because samples are lost during DNA extraction). The use of population stratification from Jing et al. (2010) to ensure representation of each group matters because cultivars are of great interest to breeders, but also, by incorporating the wild varieties, this provides diversity due to allelic richness and provides insights into taxonomy (Thachuk et al., 2009). Table 2.1 shows that the JIC list has more landraces compared to all other groups. A landrace:cultivar:wild split of 140:140:70 was chosen to explore diversity but also to look at representativeness.

The findings of this chapter fit the literature indicating seed weight as a hallmark of domestication as shown in Table 2.1. In the domain, seed weight of landrace and cultivars is approximately double that found in the wild varieties. It is important to note, sample size is far smaller in the JIC domain of wild material than cultivated or landrace. This chapter found that the JIC domain is not normally distributed (Table 2.2). Since sample size of wild is smaller than domesticated material, this may also be a possibility for failing the test of normality. This statistically representative core collection consists of 344 statistically representative samples of the domain (and 6 additional samples). The seed weights of each group of the statistically representative core collection that had a seed weight (344 samples) are almost identical to those of the groups found in the domain (Tables 2.1 and 2.3).

This work differs from some others because it does not use *a priori* information such as genetic markers or agroclimatic datasets. FIGS is a possibility of developing a core collection based on trait and environmental parameters. However, the germplasm database, SeedStor (Horler et al., 2017), does not have agroclimatics data for many accessions, but it does have some historical phenotypic data. This inconsistency and lack of location data means that the *a priori* information that FIGS would need in order to develop a core collection is absent. If location data were available, FIGS would use this alongside trait data to find accessions that have a particular trait in a particular agroclimatic environment (Bari et al., 2012).

Alternative options such as Core Hunter or PowerCore require a marker panel. Jing's data has been made available (but is difficult to get hold of in the "Germinate" database). It would have been good to have used this data, but if genetic markers were to be used, it provides no guarantee that those markers are in genes that correspond with this study's phenotypes of interest. Other studies such as one generating SNPs (Burstin et al., 2015), would not have been of use since the panel of accessions should be the same.

Furthermore, this work differs from others because an existing core collection could not be used. Firstly, they may be different core collections from different germplasm collections, therefore, this study would not have had access to such stocks.

There are certainly limitations of this core collection in the context of a genome-wide association study. Primarily, since it was selected on seed weight, further GWAS analysis may not show association with other phenotypes, unless the genetic markers lie in linkage disequilibrium with those associated with seed weight. Secondly, this core collection is largely statistically representative rather than entirely representative.

This core collection can be considered to be a random sampling of the domain using stratification from the Jing et al. (2010) dataset. This is slightly larger than a typical core collection to allow for sample drop-out, so there may be some repetitiveness depending on how many samples drop out. However, it should give a good indication of diversity and representation of the domain.

In hindsight, it would have been prudent to exclude accessions that do not hold a seed weight in the domain so downstream analysis would be representative of the domain. Furthermore, RBIP markers and CoreHunter could have been used to develop the core collection - but the RBIPs were difficult to find. Furthermore, perhaps, if bearing the downstream GWAS in mind, it may have helped to have changed strategy from a representation of the domain to one which takes the top and bottom seed weight of each group or sub-group since that is a phenotype of interest. The method used did not take a systematic approach through chronology such as obtaining every nth accession (Zeuli and Qualset, 1993). This approach would give a systematic layout of the structure combine order if performed on this variable and perhaps would have been a lot simpler in method.

This chapter has obtained a random stratified core collection for a phenotype of interest, seed weight. The work done here can contribute to the field: people can use the 344 accessions that hold a seed weight by filtering the statistically representative samples by seed weight as a core collection for the John Innes Centre *Pisum* germplasm and this can also help towards building a world pea core collection if used in conjunction with data from other world germplasm collections.

All subsequent research performed in this thesis uses this core collection.

Chapter 3

Results I: Phenotyping

3.1 Introduction

3.1.1 Image Analysis

Image analysis (IA) is defined as acquiring data from an image. When working with biological material, IA is of benefit since it can be a fairly non-invasive method which does not harm the biological material being observed. Image analysis has the potential to be upscaled in a high throughput manner. It can be easy to replicate and can lead to reproducible results when using the same method on the same image. The images can also serve as historical documentation, in other words it is a snapshot in time of what the organ looked like and can help to form a horticultural library.

Previous research using image analysis in seeds

Image analysis (IA) has been used by Sandeep, Kanaka, and Keshavulu (2013) in seed certification to detect the variety and purity of the seed using **Distinctness Uniformity and Stability (DUS)** characters. Seed analysis can identify new varieties using DUS assessment to determine if the seeds are actually different to existing varieties. Image analysis is a non-invasive method and uses metrics such as size, colour, shape and texture to help grade and sort seed.

Iva et al. (2013) measured flax seeds using IA in a highly reproducible manner. They did this by arranging the seeds in such a fashion as to not touch one another, in accordance to Bacchetta et al. (2008). This arrangement was used in order to prevent errors. The authors calibrated the scanner according to the Venora, Grillo, and Saccone (2009) method and this attention to detail of its methodology makes it highly reproducible. It makes note of the scanner type, resolution (measured in **dots per inch (dpi)**), colour depth (measured in **bit**), scanning area (measured in number of **pixels** x number of pixels) and ensured the scanner was calibrated, before analysis was performed using a macro called, "flaxseed.mcr" for the software package KC-400 v.3.0 (Carl Zeiss Vision, Oberkochen, Germany). Although, they used specialist software designed for a dedicated machine and therefore, can only be considered highly reproducible if the same equipment is present in another laboratory, this method ensured that all images were acquired in exactly the same way.

Prasad, Mukherjee, and Gangopadhyay (2014) looked at a germplasm of sesame seeds using a digital camera and Scanning Electron Microscope before analysing the images with Image Pro. They created several "character states", which are categories of qualitative and quantitative measurements and the authors then used DARwin to construct a phenogram with pictures of each sesame group placed next to it. This work demonstrates how the shape of seed changes between groups and placed the measurements obtained from the image analysis into evolutionary context.

Smykalova et al. (2011) identified stability, differences and effect of experimental location using 33 colorimetric and morphological traits in pea by using a training set to teach a statistical classifier how to classify seeds and test sets of unknown groups to validate the accuracy of the classifier. They used the linear discriminate analysis (LDA) algorithm to classify groups by qualitative and quantitative variables in order to find differences caused by changes in stability and effect of location on the peas.

Seeds are often analysed because they are a good way to identify type, using basic measurements based on colour, shape and spots. This can be done manually but this is time consuming and challenging because it requires specialised technicians to carry out such work.

Hence, image analysis can expedite this process.

Previous research using image analysis in leaf

Historically, leaves have always been an important way to help classify plants taxonomically. Leaves have also been analysed for classification purposes using region-based features (such as aspect ratio) rather than contour-based features (such as leaf margin) to classify leaves as it has been stated that region-based features are easier to find (Lee and Chen, 2006). Image analysis has also been used to classify leaves using the Linear Discriminant Classifier, making particular use of contour features such as the leaf margin (Kalyoncu and Toygar, 2015). Corney et al. (2012) developed a new algorithm for detecting leaf teeth using Linear Discriminant Analysis to classify the genus *Tilia* into different species.

Plant recognition is an example of an advancement in leaf image analysis. Plant recognition is the ability to recognise plants (Caglayan, Guclu, and Can, 2013) through the use of machine learning classification algorithms such as k-Nearest Neighbour, Support Vector Machines as well as Naive Bayes and Random Forest methods in supervised forms using leaf images.

Royer et al. (2005) used leaf image analysis in fossil leaves to explore palaeoecology - the ecology of the past. The image analysis performed here is more difficult because on fossilised material, the leaf must be manipulated first to recover the original leaf outline. Leaf teeth were manually counted because at the time of this paper, it was considered that leaf teeth could not be measured through image analysis. The authors of this paper state that at the time of publishing their findings, no computer algorithm could count leaf teeth accurately given the resolution used. Royer et al. (2005) also calculated features such as lengths, widths and areas through image analysis. Royer et al. (2005) found that leaves with fewer teeth tend to grow in warmer climates and larger tooth area correlates with low leaf mass per area and larger nitrogen content. Peppe et al. (2011) explored the palaeoclimate using leaf size and shape but first manipulated the fossil leaf to obtain the margin before using ImageJ to analyse the images. They found that teeth are larger in wet climates and this corresponds with Royer et al. (2005) by also concluding that plants in cold climates have a tendency for more leaf teeth.

Migicovsky et al. (2018) explored leaf shape in apple using Elliptical Fourier Descriptors (EFD) on leaf images with the software package ImageJ. The authors found leaf shape can be an indicator of flowering time and fruit quality in apple before linking these traits to SNPs obtained from sequencing data.

Here, we can see how image analysis on leaves can be used to classify and recognise plants, how it can be used to correlate phenotypes with climates and ecology of the past but is still relevant today for exploring links between important agricultural traits and genetic datasets.

Previous research using image analysis in pods

Vooren and Van Der Heijden (1993) measured french bean pods using image analysis and compared it to manual measurements for length, width and apex of the bean pod. This particular analysis took a video recording of these pods, before breaking them down into single shot frames for the image analysis process. They used the software packages TCL-Image and Acuity. The authors of this paper noted that the time to complete their analysis took 2 days, whereas, manual measurements were subjective and took several weeks. Furthermore, the authors observed close coefficients of variation between manual and image analysis methods.

A study focussing on peanut used a US quarter as a scale of known width (Wu et al., 2015). The authors used a camera with the pod on a black textile background, taking great care to remove background light or shadows by using bulbs inside a flow hood. The images were analysed using ImageJ for pod area and the authors used the images to build a predictive model for measuring volume. This great attention to detail, ensures that image analysis process is easier, however, they did limit their study since the use of a coin as a scale is not conventional or recognised by the International System of Units (SI) nor is easy to measure by eye.

Polder, Blokker, and Heijden (2012) developed an ImageJ plugin for a variety of crops such as flax seed, pea pod and carrot cotyledons. Whilst this tool tries to measure multiple organs, they are not of the same plant. This tool requires QR codes which are not human readable and are inconvenient in the field. The tool also requires calibration discs to be made which are not human readable and to use this tool the user must install other ImageJ plugins.

MktStall - a fully automated multi organ image analysis tool

In Section 1.5.1, the limitations of image analysis tools were discussed. Here, we present a specific tool which is a novel multi-organ image analysis tool. It is user friendly because it requires no computational expertise to run. MktStall quantifies phenotypes in accordance to the plant trait ontology, in order to remove ambiguity when referring to phenotypes. This tool ensures high reproducibility since the method is coded in the tool and therefore highly automated. MktStall is developed to analyse images that are human readable in terms of labelling and relatable in terms of analysing rulers as scales.

MktStall has been developed to incorporate multi-threading functionality, this means parallel processing can occur so it can process multiple images at a time, significantly reducing runtime. Additionally, it provides both an interactive output as well as an output in flat file format making it easy for the user to compare and contrast between samples. This tool also annotates the images provided using the definitions of the controlled vocabulary found in the Plant Trait Ontology to allow for consistency when comparing multiple images.

This tool is named, **MktStall**, pronounced "market stall". The name is derived from the concept of attending a market where you can pick up the produce and assess their phenotypes without being hindered by cumbersome packaging. MktStall was designed to remove the “packaging” which is the computational expertise and provides a new and interactive way to engage with the phenotypic data. There can be many stalls on a market, each selling different produce. In our analogy, there is a market stall for leaf, another market stall for pods, and another market stall for seeds - but also room for expansion, for additional market stalls to be added for other organs in future work. MktStall is a tool designed and developed by

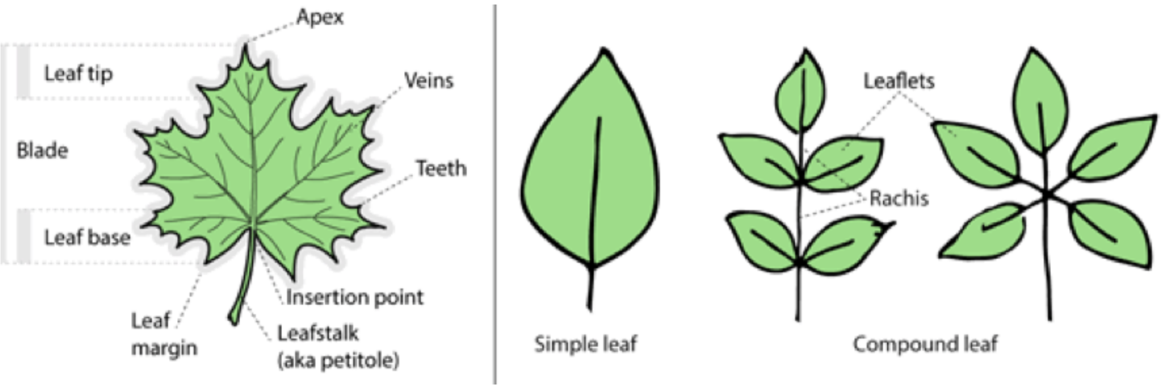


FIGURE 3.1: **Illustration of Leaf Morphology.** Figure obtained from (Wäldchen and Mäder, 2017), under the Creative Commons Attribution 4.0 International License (`http://creativecommons.org/licenses/by/4.0/`)

the author of this thesis with contributions to the code base from Dr Leighton Folkes (LF), Dr Ji Zhou (JZ) and Mr Luis Yanes (LY).

3.1.2 Introduction to common image processing strategies

There are several stages to image processing: image acquisition, image pre-processing, image segmentation, feature extraction and classification (Wäldchen and Mäder, 2017; Chitradevi and Srimathi, 2014). Here, we introduce the stages of image processing in the context of MktStall and the work outlined in this chapter.

Image Acquisition

Image acquisition is the first step of image analysis. Great consideration must be made at this stage, as any issues with this step will make the later stages of image processing much harder, if not, render the image unusable. The first consideration is the part of the organ that will be imaged. For example, in these samples, we chose not to keep the petiole of single leaflets (Figure 3.1), but to keep the calyx of the pod and to remove the seeds from the pod for imaging.

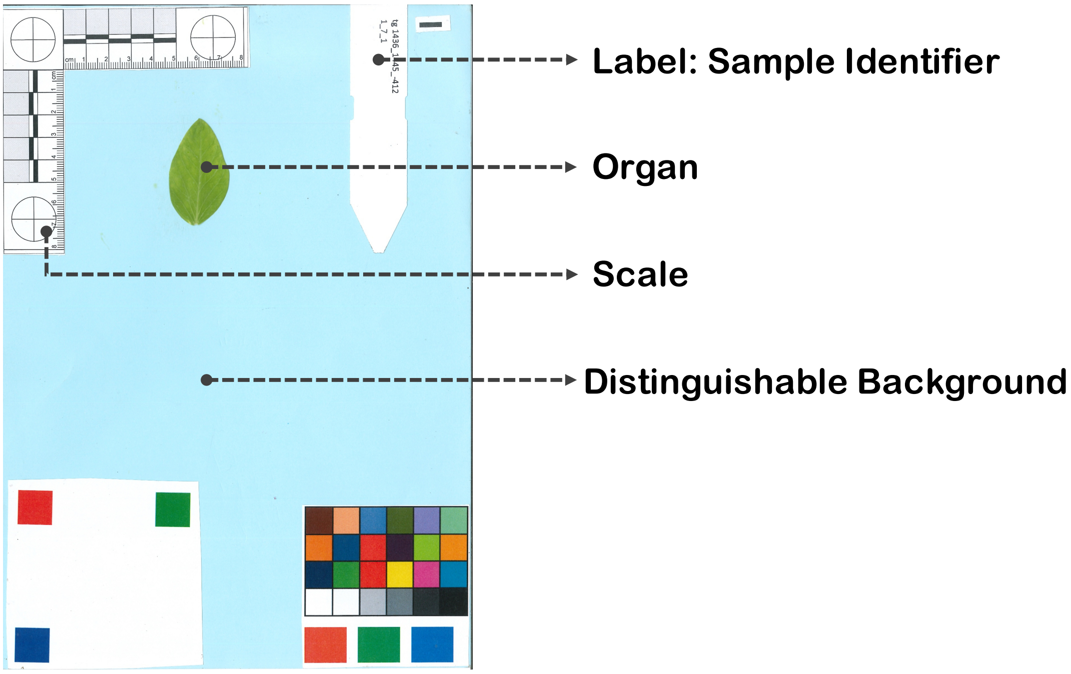


FIGURE 3.2: **Image Acquisition Design.** The design of the image acquisition includes a human readable label for sample identification, the organ to be analysed, a human readable scale and a distringuishable background to extract the organ feature. Original in colour.

Furthermore, designing how the image is to be acquired is important. There must be a scale of known size present in the image as well as clear labelling to identify the sample. Figure 3.2 shows how the design of image acquisition in this thesis is set up.

Image Pre-processing

The pre-processing stage is important, it includes **image enhancement** to make the image easier to work with and **image denoising** to filter out noise that can cause issues further downstream. The input is the image obtained from image acquisition and the output is a modified image of the input.

For image enhancement to be possible, **colourspace** must be considered. The Red Green Blue (RGB) colour model is a colour value that quantifies the amount of red, green or blue in a particular colour. Hue, Saturation and Value (HSV) is a colour transformation of the RGB colour space, where Hue is based on RGB colour value, saturation is the colourfulness of the colour, and Value describes the brightness of the colour (Figure 3.3). Under the colour model RGB, the colours are tightly coupled whereas the elements of HSV can be individually changed, for example, Hue can be kept but Saturation and Value changed for image enhancement (Hanmandlu, Jha, and Sharma, 2003). This enables downstream processing without computational complexity.

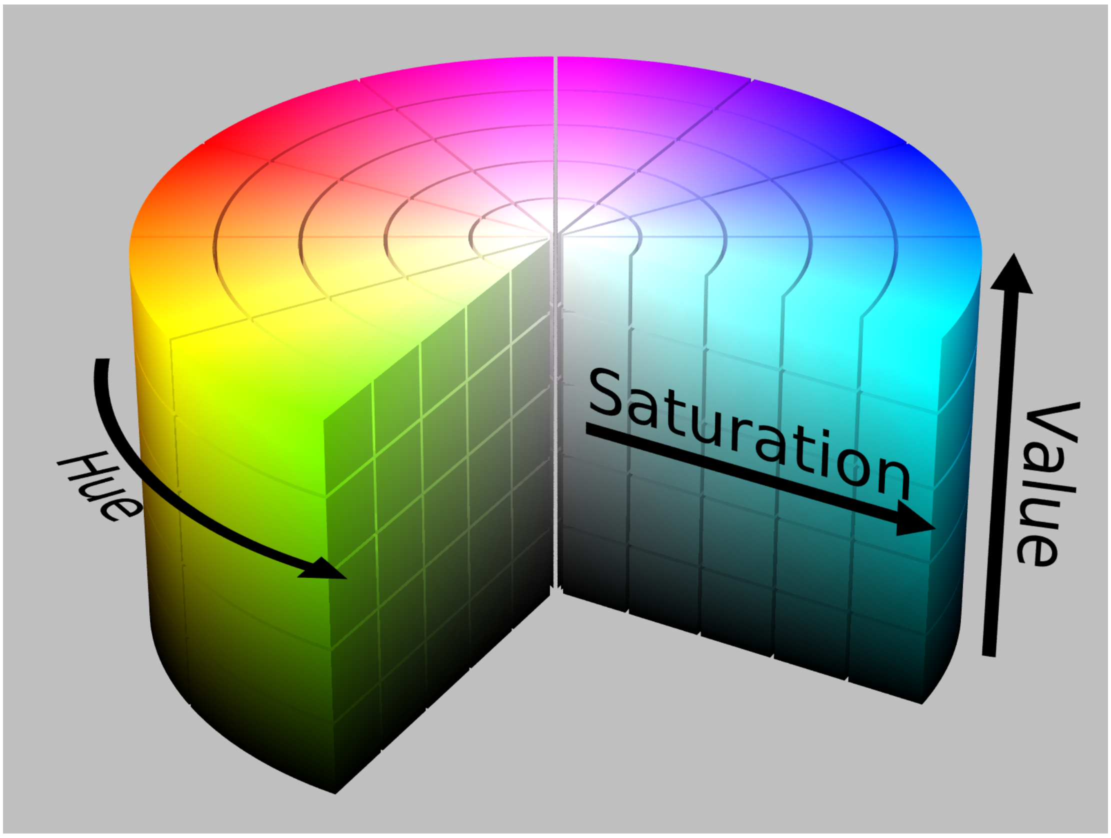


FIGURE 3.3: **Hue Saturation Value Colourspace.** Images obtained from Wikimedia Commons (2010) under the Creative Commons Attribution ShareAlike 3.0 Unported Licence with `https://creativecommons.org/licenses/by-sa/3.0/` and therefore appears with permissions. Original in colour.

Noise is unwanted variation in an image. Denoising algorithms remove this either through local-based spatial means, such as the Gaussian filter, where for each pixel, noise reduction methods are implemented based on neighbouring pixels to provide an estimate of the value for the denoised pixel or through non-local means such as fast non-local means denoising algorithm (Liu et al., 2008).

Image Segmentation

Image segmentation helps to find and divide the image into the different parts (Chitradevi and Srimathi, 2014).

Image thresholding are techniques used to segment the image. Thresholding techniques are able to distinguish between the foreground and background to create a binary image of the foreground (Al-Amri and Kalyankar, 2010). The binary image simplifies the computation complexity of the image and makes further downstream analysis easier (Al-Amri and Kalyankar, 2010).

Common thresholding techniques include **adaptive thresholding** which considers illumination of an image (Bradley and Roth, 2007), by setting the threshold to the mean of neighbouring pixels or a **gaussian** version of this where the weighted sum of neighbouring pixels is used as the threshold. **Otsu's binarisation** is designed to separate out foreground and background pixels and reduces variance within the foreground pixels and within the background pixels (Otsu, 1979).

Image segmentation can also be performed on a region level, one such example is the **watershed algorithm**. The origin of this name comes from its original application of simulating flooding (Vincent and Soille, 1991). In this instance the image has high intensity peaks and low intensity valleys, and using this analogy of flooding, as you fill the valley with different coloured water eventually the colours will merge. To prevent merging of water, artificial "dams" must be built, in other words, lines which form the segmentation (Sun, Yang, and Ren, 2005).

Feature extraction

Once segmented, features (in this case, the organ) can be measured such as length, width or area. Examples of feature detection include detecting contours using Canny edge detection (Canny, 1986). Once the edges of the organ are identified, we can describe the features using Simple and Morphological Shape Descriptors (SMSDs) (Wäldchen and Mäder, 2017).

Classification and Recognition

In classification, all extracted features form a feature vector to be classified (Wäldchen and Mäder, 2017). These can be used to classify the feature, and furthermore for recognition using supervised or unsupervised machine learning methods.

Optical Character Recognition (OCR) is a type of recognition that provides a digital output of printed characters such as those printed on a label. Providing a quick, consistent way of reading characters and giving a searchable output. It is commonly used in Automatic Number Plate Recognition and other uses include those on smartphones, smartglasses, data entry of printed records and applications for the visually impaired (Kaur and Banga, 2013; Burie et al., 2015; Depari et al., 2015; Dumitras et al., 2006).

Pytesseract is a Python wrapper for Tesseract (Google's OCR engine) (Smith, 2007). It uses a two-step classification process, first creating a list of characters that might match and then from this list those sharing similarity are found. This was trained using 60,160 training samples, making it a deep learning type of machine learning tool.

3.1.3 Simple Morphometric and Shape Descriptors (SMSDs)

Length, Width and Aspect Ratio

Major Axis Length (**L**) is defined as the vertical length of the leaf, seed or pod. Minor Axis Length (**W**) is the largest horizontal width of the leaf, seed or pod which runs parallel to the Major Axis Length (Wäldchen and Mäder, 2017). These are shown in Figures 3.4a and 3.4b. Major and Minor Axis Lengths are commonly known and length and width, respectively and it is imperative to get these correct as they form many other SMSDs.

Aspect Ratio (**AR**) is a ratio of these two lengths (Figure 3.4c). According to Wäldchen and Mäder (2017) it is alternatively referred to as slimness and is the Major to Minor Axis Length ratio given by the formula:

$$\text{Aspect Ratio} = \frac{\text{Major Axis Length}}{\text{Minor Axis Length}} \tag{3.1}$$

Area, Perimeter and Perimeter Ratios

Area (**A**) is described as the number of pixels found in the region (Wäldchen and Mäder, 2017). In flat leaf images, it is described as the surface area of a 2-Dimensional image (Figure 3.4d) and seed or pod it is the area found within the perimeter. Perimeter (**P**) is described as the sum of the distances between adjoining pixels around the leaf margin (Wäldchen and Mäder, 2017) or pod or seed margin. This feature is described as the length of the boundary of the leaf, pod, or seed margin (Figure 3.4e). There are also different ratios of perimeter, Perimeter Ratio of Major Axis Length ($\mathbf{P_L}$) and Perimeter Ratio of Major Axis Length and Minor Axis Length ($\mathbf{P_{LW}}$) (Figures 3.4f and 3.4g). These perimeter ratios have different definitions (Wäldchen and Mäder, 2017):

$$\text{Perimeter Ratio of Major Axis Length} = \frac{\text{Perimeter}}{\text{Length}} \tag{3.2}$$

$$\text{Perimeter Ratio of Major Axis Length and Minor Axis Length} = \frac{\text{Perimeter}}{(\text{Length} + \text{Width})} \tag{3.3}$$

Convexity

Convex Hull (**CH**) is also referred to Convex Area and according to Wäldchen and Mäder (2017) it is considered to be the smallest convex region containing the leaf or in other words the tightest boundary containing all of the extremities of the organ (Figure 3.4h). Area convexity ($\mathbf{A_{C1}}$) is also known as Entirety (Figure 3.4i) and defined as the normalised difference of area between the convex hull and the actual organ area (Wäldchen and Mäder, 2017).

$$\text{Area Convexity} = \frac{(\text{Convex Hull Area} - \text{Area})}{\text{Area}} \tag{3.4}$$

Area ratio of convexity (**A_{C2}**) is also defined as Solidity (Figure 3.4j), according to Wäldchen and Mäder (2017) is considered the ratio of the area and convex hull area, given by the formula:

$$\text{Area ratio of convexity} = \frac{\text{Area}}{\text{Convex Hull Area}} \tag{3.5}$$

Perimeter convexity (**P_C**) is the ratio of the perimeter of the convex hull and the organ perimeter (Figure 3.4k) (Wäldchen and Mäder, 2017)

$$\text{Perimeter convexity} = \frac{\text{Perimeter of Convex Hull}}{\text{Perimeter}} \tag{3.6}$$

3.1.4 Shape Descriptors

Equivalent circular diameter is defined by Wäldchen and Mäder (2017) as the diameter of a circle that has the same calculated area as that of the organ (Figure 3.4l).

$$\text{Equivalent circular diameter} = \sqrt{\frac{4 \cdot \text{Area}}{\pi}} \tag{3.7}$$

Roundness, explains how similar the organ is to a circle (Wäldchen and Mäder, 2017) and is known by alternative names, form factor, circularity or isoperimetric factor (Figure 3.4m).

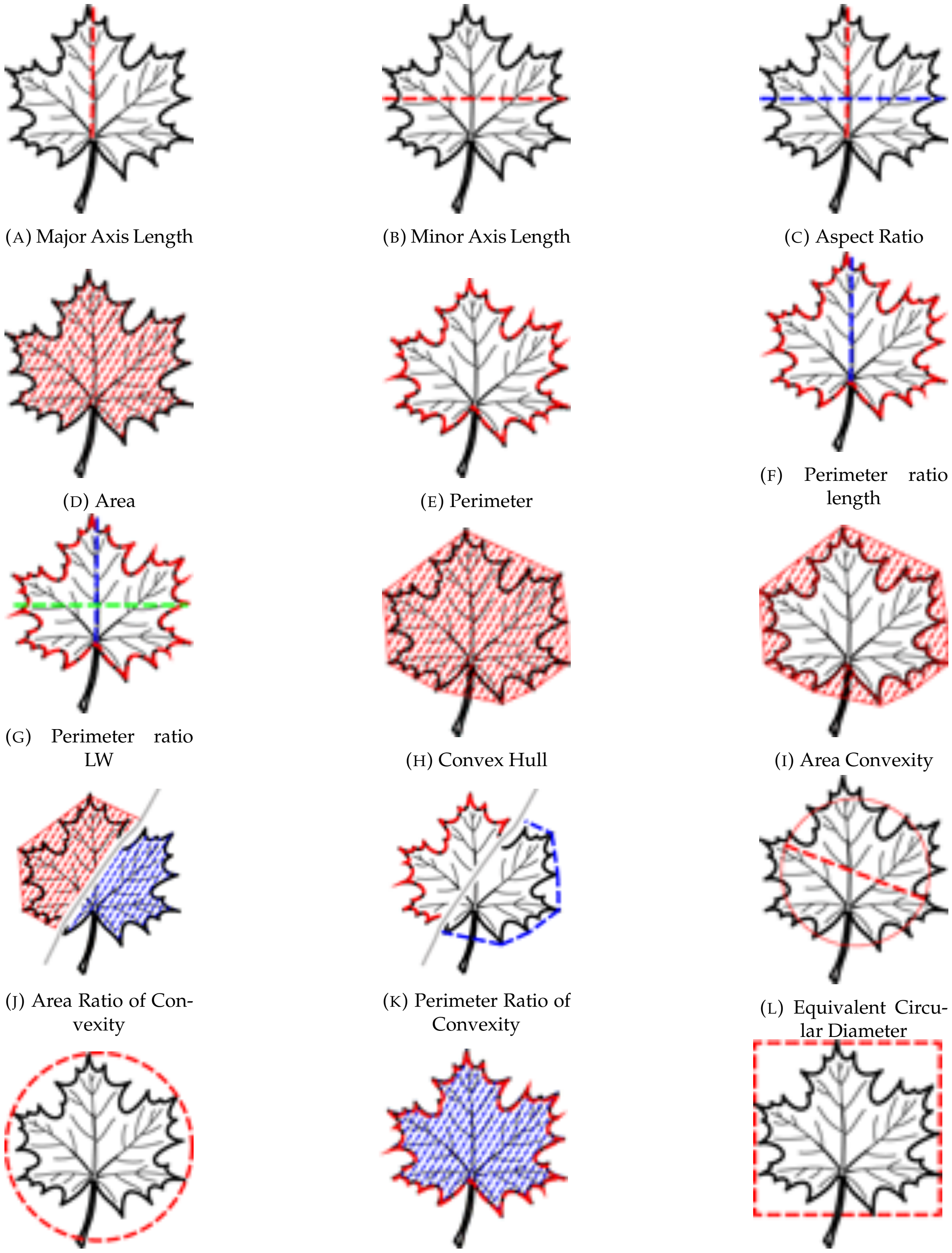
$$\text{Roundness} = \frac{4 \cdot \text{Area} \cdot \pi}{\text{Perimeter}^2} \tag{3.8}$$

Compactness is defined by Wäldchen and Mäder (2017) as the ratio of organ perimeter and organ area (Figure 3.4n).

$$\text{Compactness} = \frac{\text{Perimeter}^2}{\text{Area}} \tag{3.9}$$

Rectangularity is also called extent (Wäldchen and Mäder, 2017) and provides information on how rectangular the organ is (Figure 3.4o).

$$\text{Rectangularity} = \frac{\text{Area}}{\text{Length} \cdot \text{Width}} \tag{3.10}$$



(A) Major Axis Length

(B) Minor Axis Length

(C) Aspect Ratio

(D) Area

(E) Perimeter

(F) Perimeter ratio length

(G) Perimeter ratio LW

(H) Convex Hull

(I) Area Convexity

(J) Area Ratio of Convexity

(K) Perimeter Ratio of Convexity

(L) Equivalent Circular Diameter

(M) Roundness

(N) Compactness

(O) Rectangularity

FIGURE 3.4: **Simple Morphological and Shape Descriptors (SMSDs) using the leaf as an example organ.** Figures have been obtained from Wäldchen and Mäder (2017) under the Creative Commons Attribution 4.0 International License (`http://creativecommons.org/licenses/by/4.0/`). Assignment of images to different names have been changed for this context. Original in colour.

Proposed application of Simple Morphological and Shape Descriptors (SMSDs) in pea

The intention of obtaining these Simple Morphological and Shape Descriptors (SMSDs) in pea is to obtain not just phenotypic measurements in leaflet, pod and seed but to help better understand variation and domestication across all samples.

For example, this thesis proposes that understanding SMSDs in seed can help quantify how round or wrinkled these seeds are. Seed wrinkledness is important as wrinkled peas are known to be sweeter (Rayner et al., 2017) due to changes in starch structure. Since domesticated varieties of pea are known to have larger seeds (Weeden, 2007) this thesis aims to use SMSDs to measure seed size.

Leaves have often been used to classify plants taxonomically, but this thesis aims to quantify the phenotypes to provide a better understanding of domestication in pea. Xu et al. (2009) and Royer et al. (2005) found an increased presence of leaflet teeth correlated with growth in cold temperatures and Peppe et al. (2011) identified a low presence of teeth in wet climates. Therefore, in this application, SMSDs may be able to provide insights into adaptations of leaflet size and shape of certain pea varieties.

Pod shape is an important trait because there appears to be a correlation between blunt pods and higher rates of seed abortion in the pod (Lee, 1988). SMSDs in this application may help in identifying varieties more prone to seed abortion and therefore lower yields.

3.1.5 Objectives of the work

The objective of this chapter is to (i) develop a novel multi-organ image analysis tool for measuring Simple Morphometric and Shape Descriptors in leaflet, seed and pod (ii) obtain quantitative and morphometric measurements for: seed weight, plant height, seed size, leaflet margin and pod shape (iii) illustrate the distribution of phenotypes across the core collection, between groups and where applicable, between this dataset and a historical JIC dataset as well as between environments.

The first objective is the development of a novel tool called "MktStall". The image acquisition serves as a snapshot of the observed organ in time - a horticultural library. This objective is important because this will be the novel multi-organ image analysis tool to measure phenotypes in a high throughput and consistent manner, that could be replicable by other scientists and without prior computational expertise, serving as a useful tool for the community to use.

The second objective of obtaining phenotypic measurements includes the manual measurements of seed weight and plant height as well as the output of MktStall for leaflet margin, seed size and pod shape. This objective is important to the study of evolution and domestication across the core collection to explore observable differences. By using SMSDs, it provides phenotypes that are easily measurable and identifiable to see how these change through domestication.

The third objective to explore changes between groups and environments as well as between this dataset and the historical JIC dataset is an important one. Through the exploration of the pea's characteristics or traits across the population stratification of this species, variations for each phenotype can be observed. This assumes that under the process of natural selection, different phenotypes will help the organism to adapt to its environment and therefore survive. It assumes that with humans creating artificial selection pressures, different phenotypes may occur leading to domestication. These plants can be grouped together based on phenotype. The end goal is to perform a GWAS study which will help associate the phenotype with its genotype for all accessions. The aim is to identify SNPs which statistically associate with these phenotypes and may be in linkage with a causal variant. Finding SNPs associated with the phenotype has potential to create new tools for breeding.

In short, it is hypothesised that wild plants found in group 4 will have smaller seed weight, shorter plant height, smaller seeds, more serrated leaflet margins, smaller and thinner pods with blunt ends in comparison to the landrace and cultivar material in groups 1 and 2. It is also hypothesised that field varieties have lower measurements compared to glasshouse across all yield related traits observed.

3.2 Materials and Methods

3.2.1 Experimental Design and Sampling Strategy

The peas were grown in glasshouse (GH) and field environments during Spring 2015. Peas in the glasshouse were grown in JIC structure order in replicates of three as shown in Figure 3.5.

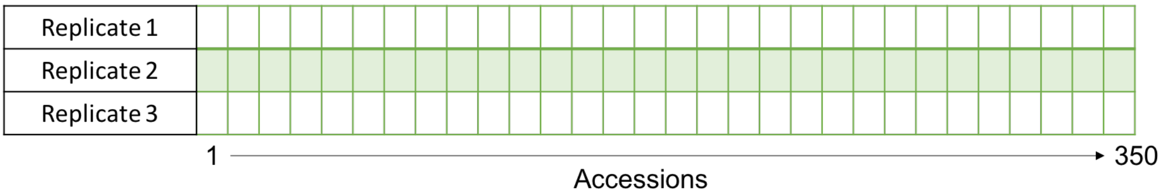


FIGURE 3.5: **Experimental Design of the Glasshouse**. Seeds were sown in replicates of three for 350 accessions

Peas grown in the field were also grown in JIC structure order in replicates of three as shown in Figure 3.6. All plants were grown to full maturity prior to organ harvesting. Seed weight was the only trait to be phenotyped in both glasshouse and field environments but all phenotypes were observed in the glasshouse environments.

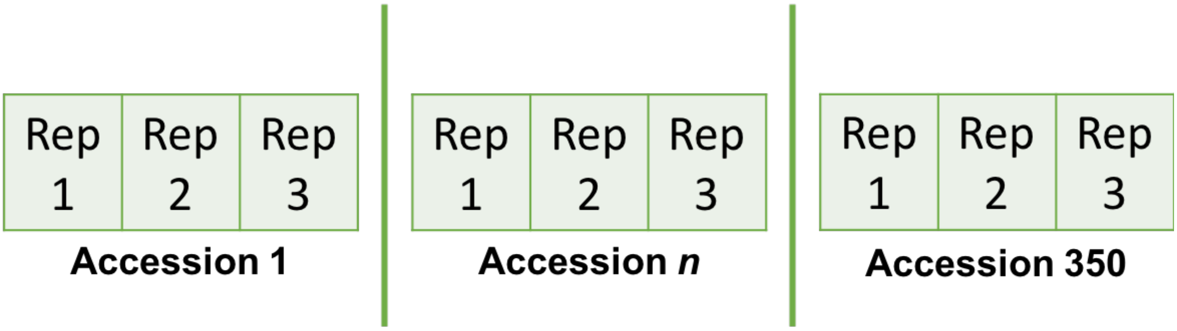


FIGURE 3.6: **Experimental Design of the Field**. Seeds were sown in replicates of three for 350 accessions. Seed weight was the only phenotype observed in the field for this study.

3.2.2 Material Collection and Image Acquisition

For each environment of GH and Field, different phenotypes were measured. For glasshouse peas, the seed weight and plant height were measured manually and images taken for seed, leaflet and pod. For field peas, only seed weight was measured and this was performed manually.

The appropriate organs were first harvested from the glasshouse peas prior to image acquisition. For leaf, the most viable leaf was harvested from each plant and the petiole was removed prior to imaging on the MktStall background using a scanner. For pod, a random pod was harvested with the calyx from each plant intact and placed onto the MktStall background before imaging with a scanner. All seeds were harvested from the total number of pods and any aborted seeds or those thought to be infected by pea moth were removed prior to imaging on the MktStall background using a camera.

For the image acquisition, a camera tripod was set up for imaging seeds and an office photocopier was used to image the leaves and pods. All images taken against a blue background with ruler measurement.

Then the remaining manual phenotypes were measured. The plant height of the glasshouse peas were measured manually with a tape measure. The plants were measured plants from the top of the soil to the last growing point after the removal of cane sticks and other supports. Seed weight of both Glasshouse and Field were weighed using a weighing scale and counted with a tablet counter.

3.2.3 Phenotyping - Image Analysis

Seed

All seed images were analysed using the MktStall "Seed" module. For the image preprocessing, MktStall ensures that the image is converted to JPEG format. From the image, it calculated the RGB values of each pixel and converted it to the HSV colour model. Using the HSV image, MktStall masked and segmented the image to identify the ruler, organ and label sections.

To ensure all measurements could be measured in centimetres, MktStall measured the number of pixels in each centimetre of the ruler section.

To measure the organs, a green mask was applied to the HSV images and the images were denoised using a bilateral filter using the parameters supplied in Appendix B. Each seed was segmented out with the watershed algorithm and was measured in accordance to the Wäldchen and Mäder (2017) paper and with respect to the Plant Trait Ontology (Arnaud et al., 2012).

The label regions were identified by rotating the image and then transforming the image, first to grayscale, then using adaptive thresholding and finally cropped to size. The cropped label image is progressively denoised, first by applying a Fast Non-Local Means denoising algorithm. The OCR-engine (Smith, 2007) attempts to read the label on the denoised image and if it fails to do so, the label image is then further denoised and checked after the use of each algorithm to see if it can be read. The progressive denoising algorithms used were the Global Thresholding algorithm, Otsu's thresholding algorithm and a combination of Gaussian and Otsu's thresholding using a 5x5 Gaussian kernel. The parameters for the above can be found in Appendix B.

For each accession, a Comma-Seperated Values (CSV) file is created with the results for each seed. The average seed phenotypes are calculated and outputted into a Comma-Seperated Values (CSV) file for all accessions.

Leaflet

All leaflet images were analysed using the MktStall "Leaflet" module. The leaflet module uses the same method as the seed module for the image pre-processing and analysis of the ruler and label regions but a different method for the feature extraction of the organ.

The organ feature was detected using Canny Edge detection and then orientated the correct way up. Region properties and SMSDs were calculated in accordance to the Wäldchen and Mäder (2017) paper and with respect to the Plant Trait Ontology (Arnaud et al., 2012). For leaflet teeth, the margin was plotted on a graph by traversing in an anti-clockwise direction from the apex and plotting margin co-ordinates. Leaflet teeth are detected in the local minima and maxima those that lie 2 standard deviations above or below the mean are removed, in accordance with the algorithm outlined in the Corney et al. (2012) paper with the exception of the area value of 15 being substituted to 2 standard deviations. The results are outputted into a Comma-Seperated Values (CSV) file.

Pod

All pod images were analysed using the MktStall "Pod" modules. The pod module uses the same method as the seed module for the image pre-processing and analysis of both the ruler and label regions but a different method for the feature extraction of the organ.

The organ feature was detected using Canny Edge detection and then orientated the correct way up. Region properties and SMSDs were calculated in accordance to the Wäldchen and Mäder (2017) paper and with respect to the Plant Trait Ontology (Arnaud et al., 2012). The results are outputted into a Comma-Seperated Values (CSV) file for each phenotype.

MktStall

The above forms the back-end of MktStall written in Python 2.7 requiring the following dependancies: pytesseract to read the label (Smith, 2007), scikit-image for feature extraction and measurements (Walt et al., 2014), OpenCV for image processing (Culjak et al., 2012), Python Imaging Library for opening and saving images, NumPy for storing large arrays (Walt, Colbert, and Varoquaux, 2011), scipy (*SciPy*), matplotlib.pyplot to draw plots (*Matplotlib - pyplot*) and shapely to draw the shapes for visualisation (*Shapely - Manipulation and analysis of geometric objects*).

The front-end Graphical User Interface was written in Java to provide an interactive way of visualising the dataset provided by the back-end of MktStall via the CSV files.

The MktStall source code has been made publicly available at following address: `https://github.com/kheth/MktStall`.

3.2.4 Statistical Analysis

The data generated by MktStall as well as manual measurements for plant height and seed weight were statistically analysed using a decision tree outlined in Figure 3.7. The statistical analysis was performed in R version 3.4.3 (R Core Team, 2013) using the following libraries: dplyr, gplots, PMCMR and ggplot2.

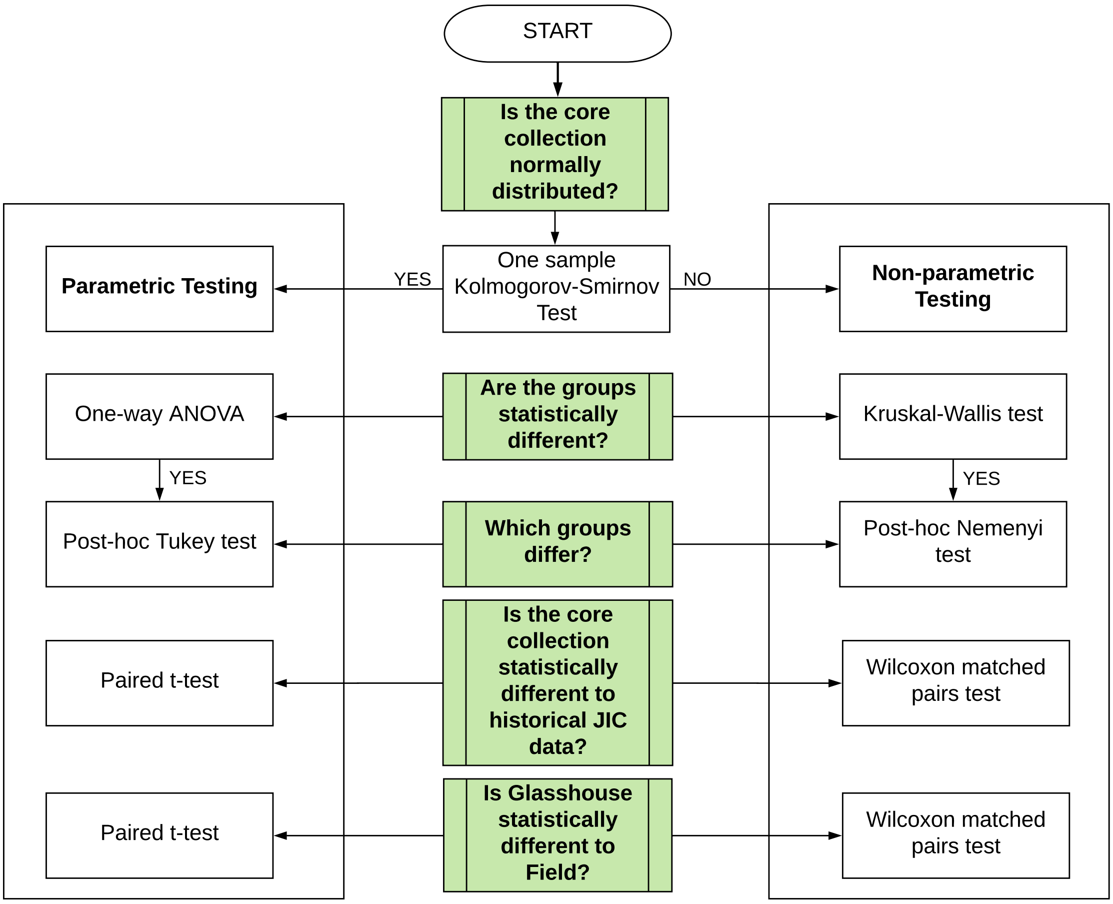

START
Is the core collection normally distributed?
One sample Kolmogorov-Smirnov Test
YES
NO
Parametric Testing
Non-parametric Testing
Are the groups statistically different?
One-way ANOVA
Kruskal-Wallis test
YES
YES
Which groups differ?
Post-hoc Tukey test
Post-hoc Nemenyi test
Is the core collection statistically different to historical JIC data?
Paired t-test
Wilcoxon matched pairs test
Is Glasshouse statistically different to Field?
Paired t-test
Wilcoxon matched pairs test

FIGURE 3.7: **Statistical decision tree for the analysis of each phenotype in our core collection**. Green boxes denote biological questions being answered. The left hand box denotes parametric tests to be used if the core collection is normally distributed. The right hand box denotes non-parametric tests to be used if the core core collection is not normally distributed.

Each phenotype was tested to see if the core collection was normally distributed. To determine the normality of each phenotype a one-sample Kolmogorov-Smirnov (K-S) statistical test was performed and supplemented by plotting histograms (for plant height only), kernel density plots and Quartile-Quartile (Q-Q) plots. The results of this will subsequently inform whether downstream analysis requires parametric or non-parametric testing. For phenotypes which do not possess a normal distribution, a non-parametric Kruskal-Wallis statistical test was used to see if the phenotype differed between groups. This statistical test provides a dichotomous result, either statistically or not statistically significant, therefore a post hoc Nemenyi test was performed to determine which groups differed significantly.

The variability of each phenotype was also tested using ANOVA to analyse between replicate and within replicate variability.

For seed weight of the core collection, the GH and field were compared against seed weights held in a historical dataset using Wilcoxon matched test. Field and GH were then compared to see how they differed statistically also using the Wilcoxon matched test. Correlation analysis between the glasshouse and field weights were performed using Kendall's Tau-b Rank Correlation.

Furthermore, a correlation matrix of the phenotypes was produced in R version 3.4.3 using the following libraries: corrplot, corrgram and RColorBrewer. A Biplot of the trait space was also created, in R version 3.4.3 using the following libraries: FactoMineR, factoextra and corrplot.

3.3 Results

3.3.1 Plant Height

Plant height from the glasshouse was tested to see if it followed a normal distribution, with the null hypothesis being that there is no difference from the normal distribution.

The Kolmogorov-Smirov test indicate that the null hypothesis should be rejected and suggest plant height of peas from glasshouse do not follow a normal distribution (Appendix C). The Quartile-Quartile (Q-Q) plots for this dataset (Figure D.1) displays distension away from the normal reference line (drawn in red) supporting the indication that this is not normally distributed. The extreme tails observed in the histogram indicate bimodality in the data (Figure 3.8). The data suggests the core collection is not normally distributed for plant height.

The Q-Q plots of Groups 1 and 2 (Figure D.1) suggest these groups are not normally distributed. The Kernel Density Plots indicate Group 1 (landrace) is skewed towards taller plant height and Group 2 (cultivars) skewed towards shorter plant height in Figure 3.9.

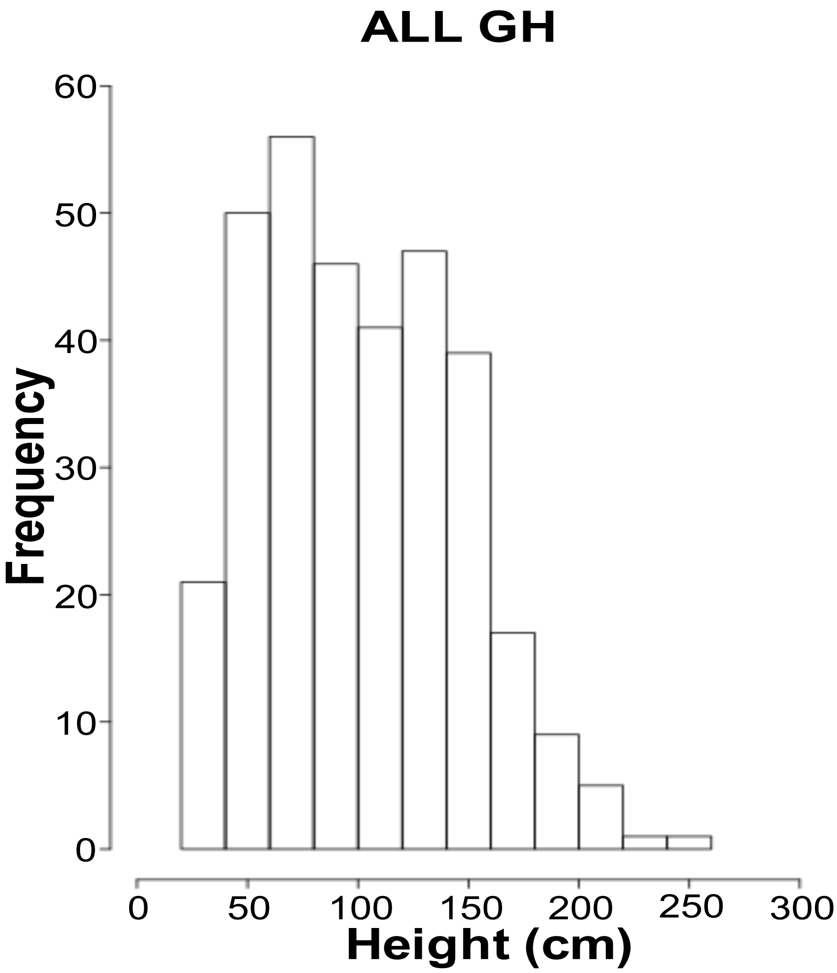


FIGURE 3.8: **Histogram for average plant height of each pea variety for peas grown in a glasshouse environment.**

Groups were tested to see if they were different with the null hypothesis being there is no difference between groups. The results indicate the null hypothesis should be rejected and groups are statistically different (Appendix E). A post hoc Nemenyi test (Appendix F) was used to test which of these groups differed and the results show there is no statistical difference between group 1 (landrace) and group 4 (wild) (p-value = 0.138) but there were differences between groups 1 and 2 (landrace and cultivar respectively) (p-value = 2.1e-10) and between groups 2 and 4 (cultivar and wild respectively) (p-value = 0.003). The results suggest cultivated varieties differ statistically significantly in plant height but the wild and landrace material do not differ statistically significantly in plant height. This may

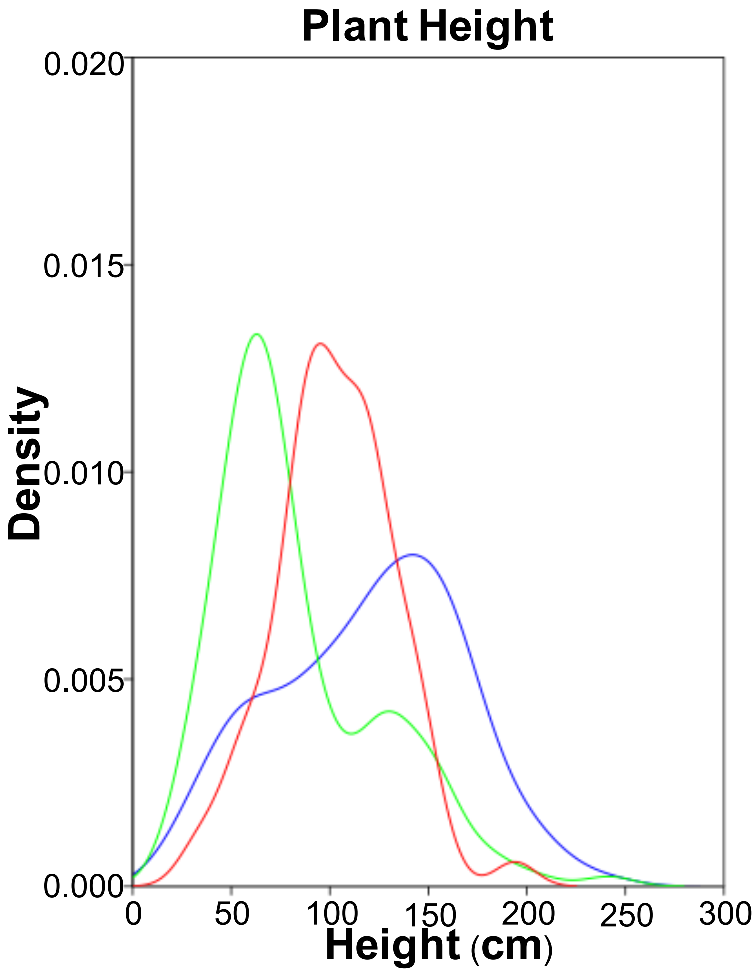


FIGURE 3.9: **Kernel Density Plot for average plant height of each pea variety for peas grown in a glasshouse environment.** The core collection was separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). Original in colour.

give insights into the understanding of plant height in pea domestication. Table G shows that Group 2 has the lowest median (median = 70.4cm (3 s.f.)) and the lowest mean (mean ± sd = 83.5 ± 41.6 cm (3 s.f.)) plant height out of all the groups suggesting tendencies towards shorter height in this cultivated group.

The variability of the plant height was tested using ANOVA with the null hypothesis being that there is no difference between replicates. The ANOVA results for plant height (p-value = 0.0673) suggest there are differences between the replicates (Table 3.1) and this suggests there maybe genetic variability for this trait.

TABLE 3.1: **ANOVA for plant height.** This denotes degrees of freedom (Df), sum of squares (SS), mean square (MS), F value test statistic and the corresponding p-value.

	Df	SS	MS	F value	p-value
Replicates	2	15425	7712	2.706	0.0673
Residuals	1044	2975937	2851		

3.3.2 Seed Weight

Seed weight for glasshouse and field environments were tested for normal distribution, with the null hypothesis being there is no difference from the normal distribution. The Kolmogorov-Smirnov test was used with results for glasshouse and field (Appendix C) indicate that the null hypothesis should be rejected, that both datasets are not normally distributed and therefore for this phenotype all subsequent statistical analysis should be non-parametric. Kernel density plots and Q-Q plots were drawn to visualise how normal the distribution of data was for GH and Field (Figures 3.10, 3.11, D.2, and D.3). All glasshouse and field datasets appear to follow the normal line, but distend away from this line at the end, indicating presence of some extreme tails. For the groups, there is greater distension away from the line, particularly at the high seed weight end, indicating this is also an extreme tail. Group 4 for field and glasshouse shows an upward curve, indicating skewness, and the kernel density plot shows it is skewed to the right. Out of the groups, group 2 (cultivar), appear to have the most bell-shaped kernel density plot, and closest fit to the normal reference line on the Q-Q plot. The next closest fit to the normal line is group 1 (landrace) whereas group 4 (wild) hold the most skewed histograms and are the furthest away from the normal line.

The data for the seed weight of the core collection grown in glasshouse and for field (Appendix G) show that the mean seed weight is lower when grown in the field, but both hold a similar amount of variability, so the data overlaps between both conditions.

Based on the Kolmogorov-Smirnov test and data shown in (Figures 3.10 and 3.11), a Kruskal-Wallis test was performed to see if the groups differed statistically. The null hypothesis was that there was no difference between the groups. The groups were statistically different for glasshouse and for field (Appendix E). A post hoc Nemenyi test was used to identify which groups are different and the results indicate that all groups were statistically different to one another in glasshouse and field (Appendix F).

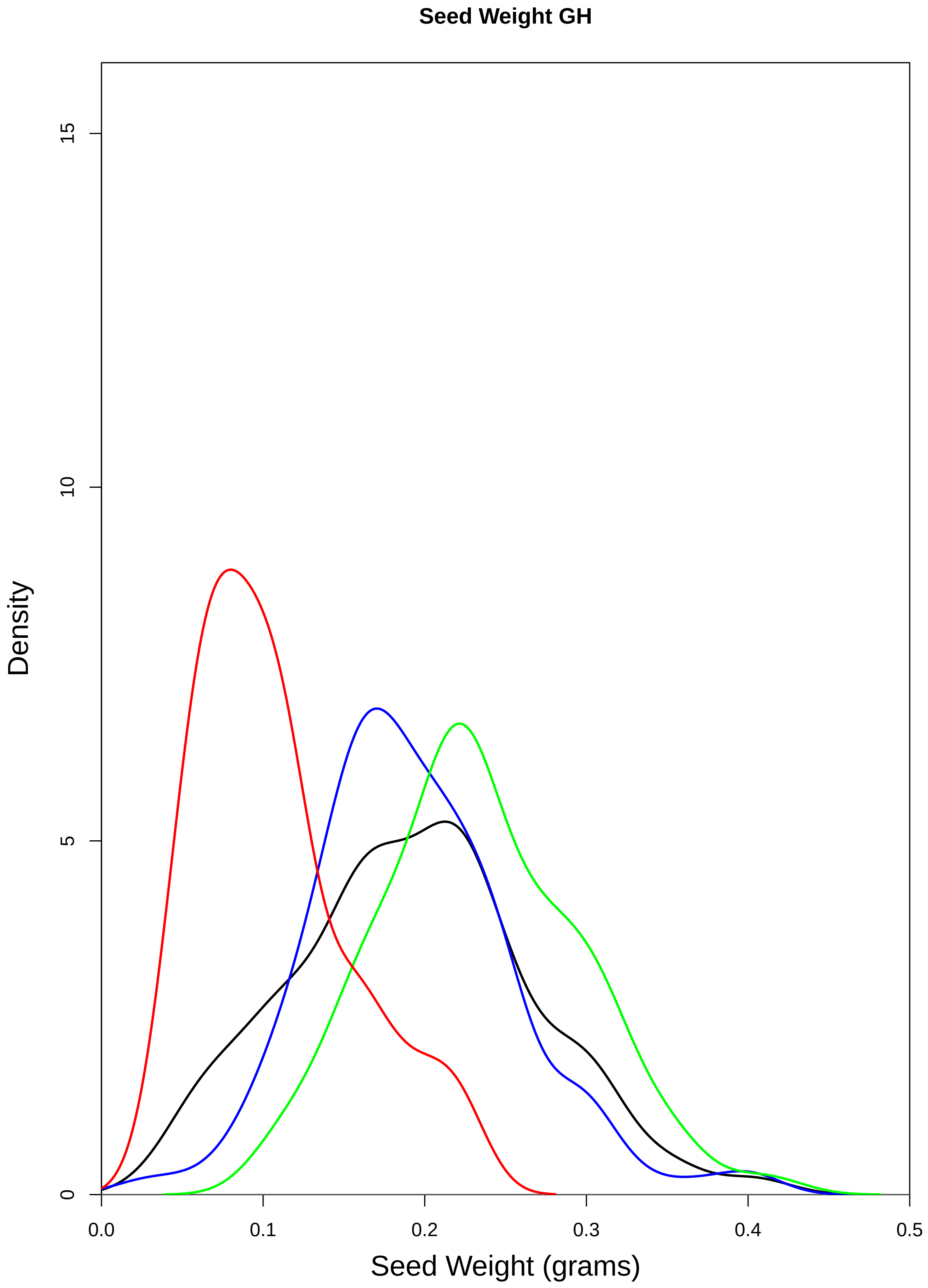


FIGURE 3.10: **Kernel Density Plot for seed weight of peas grown in a glasshouse environment.** The core collection (black) was also separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). Original in colour.

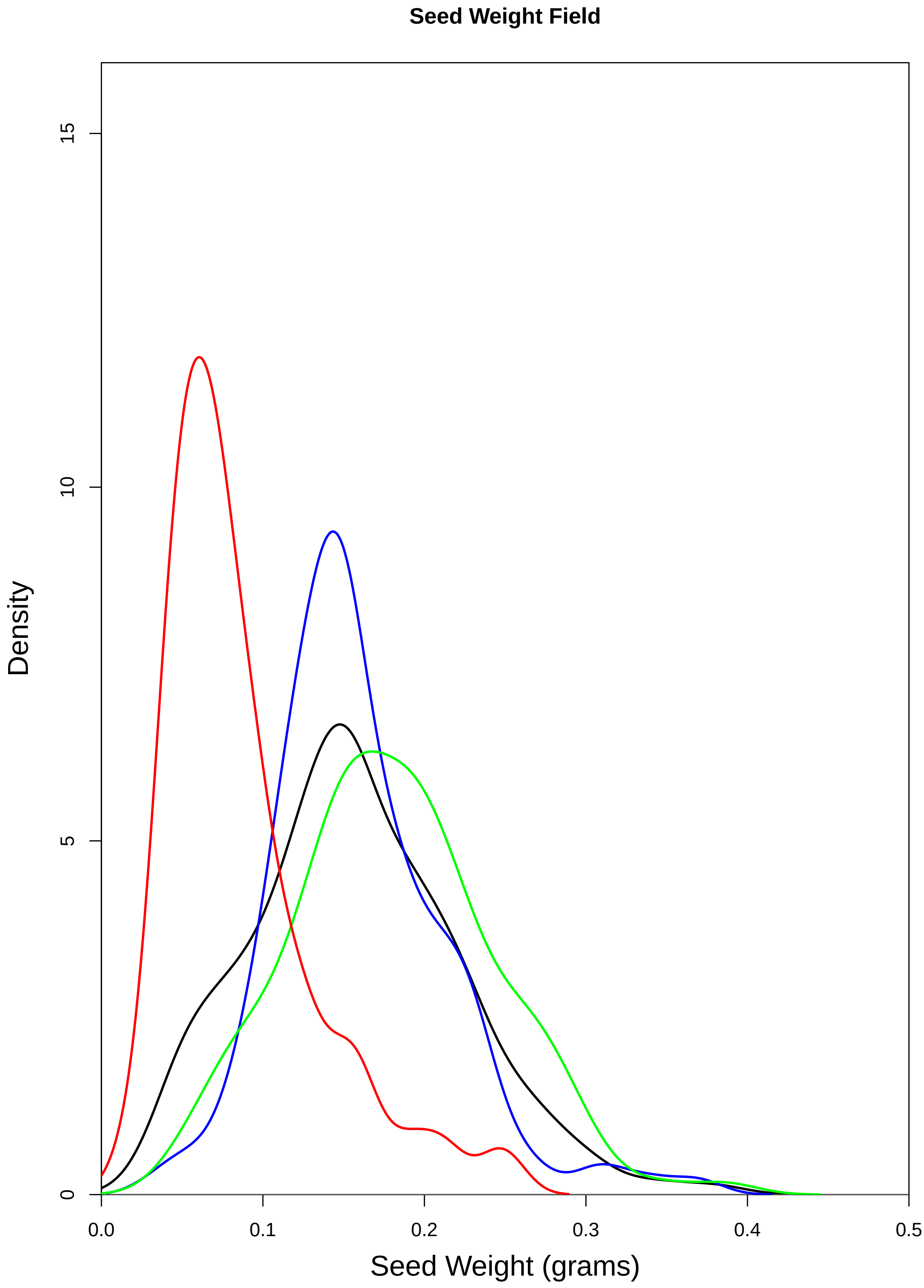


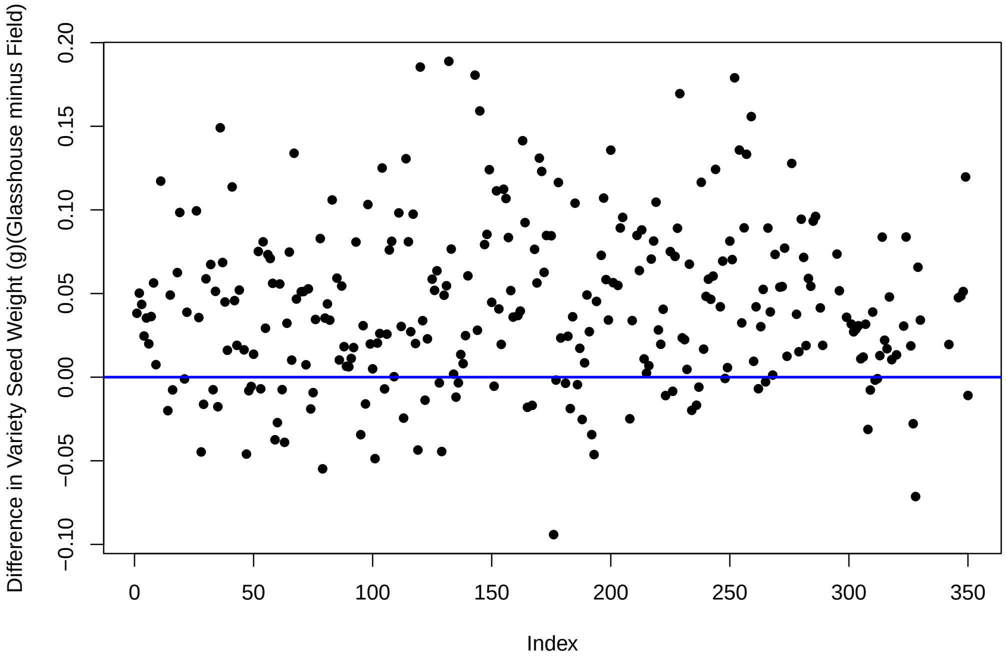
FIGURE 3.11: **Kernel Density Plot for seed weight of peas grown in a field environment.** The core collection (black) was also separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). Original in colour.

Appendix G alongside Figures 3.10 and 3.11 also show the distribution of data between groups. Under both conditions, group 2 (cultivar) has the highest seed weight, followed by group 1 (landrace) and group 4 (wild). Field grown peas have more extreme points in comparison to glasshouse grown peas. In glasshouse grown peas, the symmetry lies slightly to the lower seed weight end for groups 1 and 2. Whereas, in field grown peas, group 4 lies slightly more to the lower seed weight end.

Comparing the glasshouse dataset to the historical dataset using Wilcoxon matched pairs test suggests there is a difference between them (V = 18464, p-value = 9.318e-06, Wilcoxon Matched Pairs test). The field dataset also differed to the historical dataset (V = 2876, p-value < 2.2e-16, Wilcoxon Matched Pairs test). This may be due to changes in environment conditions in field but does not explain the changes in glasshouse environment.

Furthermore, comparing glasshouse and field seed weights also suggested a difference between them (V = 36805, p-value < 2.2e-16, Wilcoxon Matched Pairs test), suggesting seed weights differ between environments. Figure 3.12 shows that glasshouse has higher seed weight compared to field.

Since both the glasshouse and field seed weight were not normally distributed, this suggests a non-parametric correlation test should be used. The Spearman Rank Correlation or Kendall's Tau-b Rank Correlation are better suited to non-parametric tests since it uses the rank of the observations. Spearman Rank Correlation does not perform well under tied ranks, so Kendall's Tau-b Rank Correlation is used here. The null hypothesis is that there is no association between the glasshouse and field seed weights. The Kendall correlation coefficient is equal to 0.5437563 and the p-value is < 2.2e-16 so we must reject the null hypothesis, there is significant positive correlation between glasshouse and field seed weight.



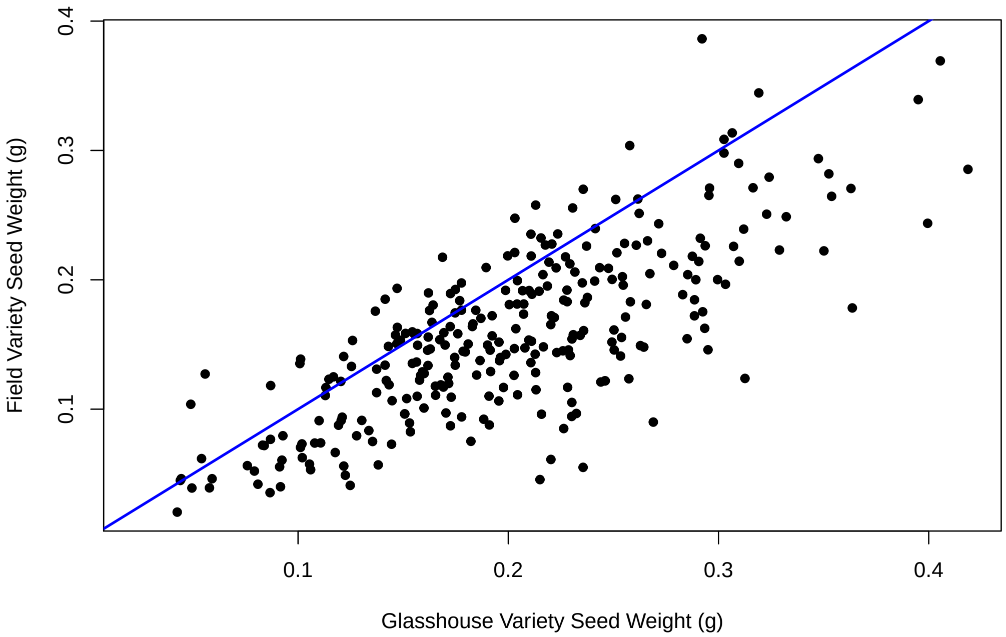


FIGURE 3.12: **Comparison of glasshouse and field seed weights.** The top graph shows each accession (Index) and the difference for that accessions seed weight (Glasshouse - Field). The blue line marks no difference between glasshouse and field, points below this line show accessions where field seed weight was larger than the glasshouse seed weight. The bottom graph plots the actual seed weights for glasshouse against field. The blue line shows accessions that lie beneath it hold higher glasshouse seed weight in comparison to field seed weight. Kendall's rank correlation shows tau = 0.5437563 (p-value < 2.2e-16)

TABLE 3.2: **ANOVA for seed weight.** This denotes degrees of freedom (Df), sum of squares (SS), mean square (MS), F value test statistic and the corresponding p-value.

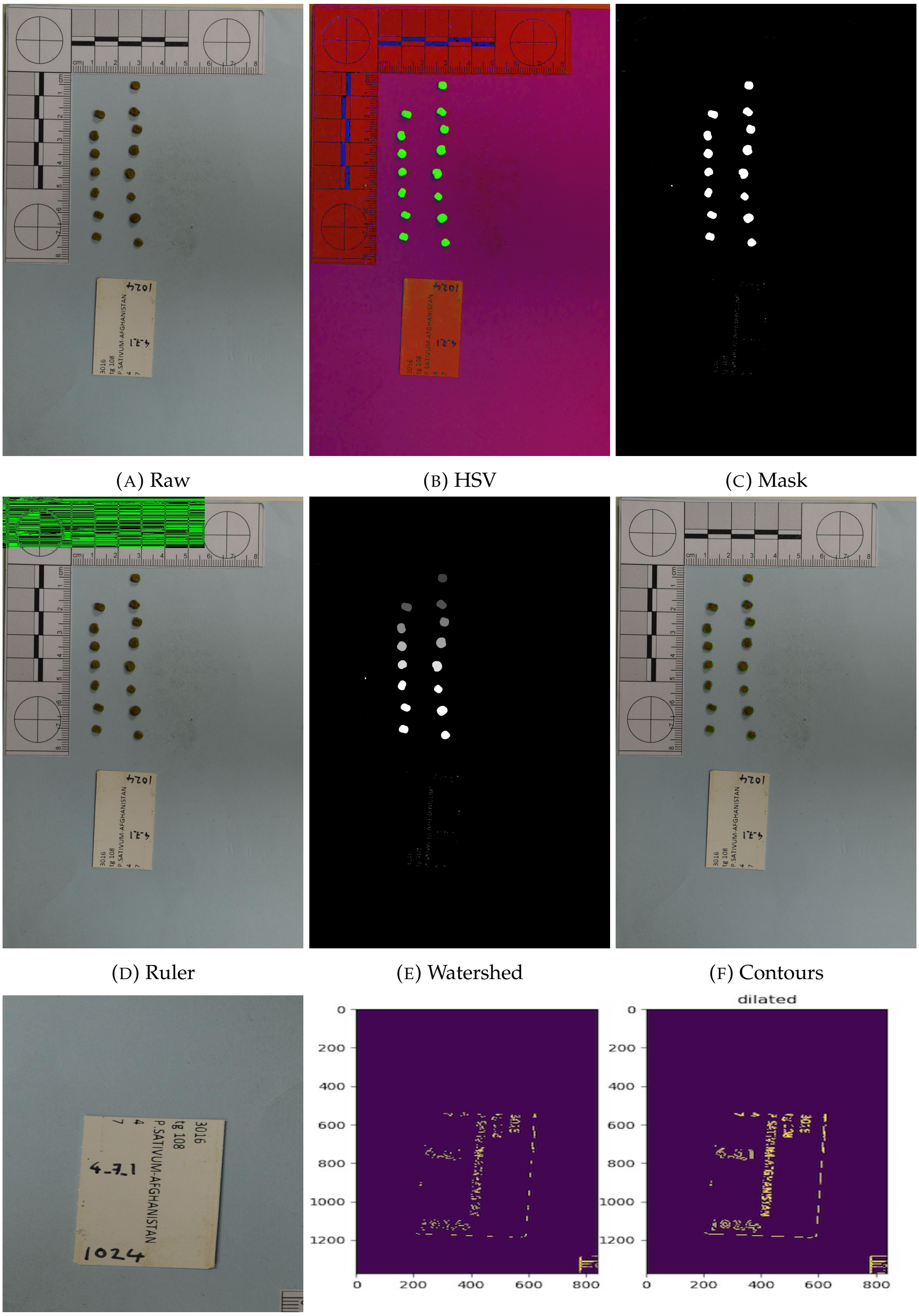
Phenotype		Df	SS	MS	F value	p-value
GH	Replicates	2	0.007	0.003661	0.387	0.679
	Residuals	1047	9.912	0.009467		
Field	Replicates	2	0.201	0.1007	11.98	7.14e-06
	Residuals	1047	8.799	0.0084		

Analysis of variance for seed weight

The variability of seed weight was analysed with the null hypothesis that there is no difference between replicates. The ANOVA results for suggest that for seed weight only those grown in the glasshouse displayed differences between the replicates (p-value = 0.679) whereas those grown in the field did not (Table 3.2).

3.3.3 MktStall - Seed

MktStall was used to detect seed area, perimeter, equivalent diameter and eccentricity. An example of the intermediary files is shown in Figure 3.13.



(A) Raw (B) HSV (C) Mask

(D) Ruler (E) Watershed (F) Contours

(G) Crop (H) Label

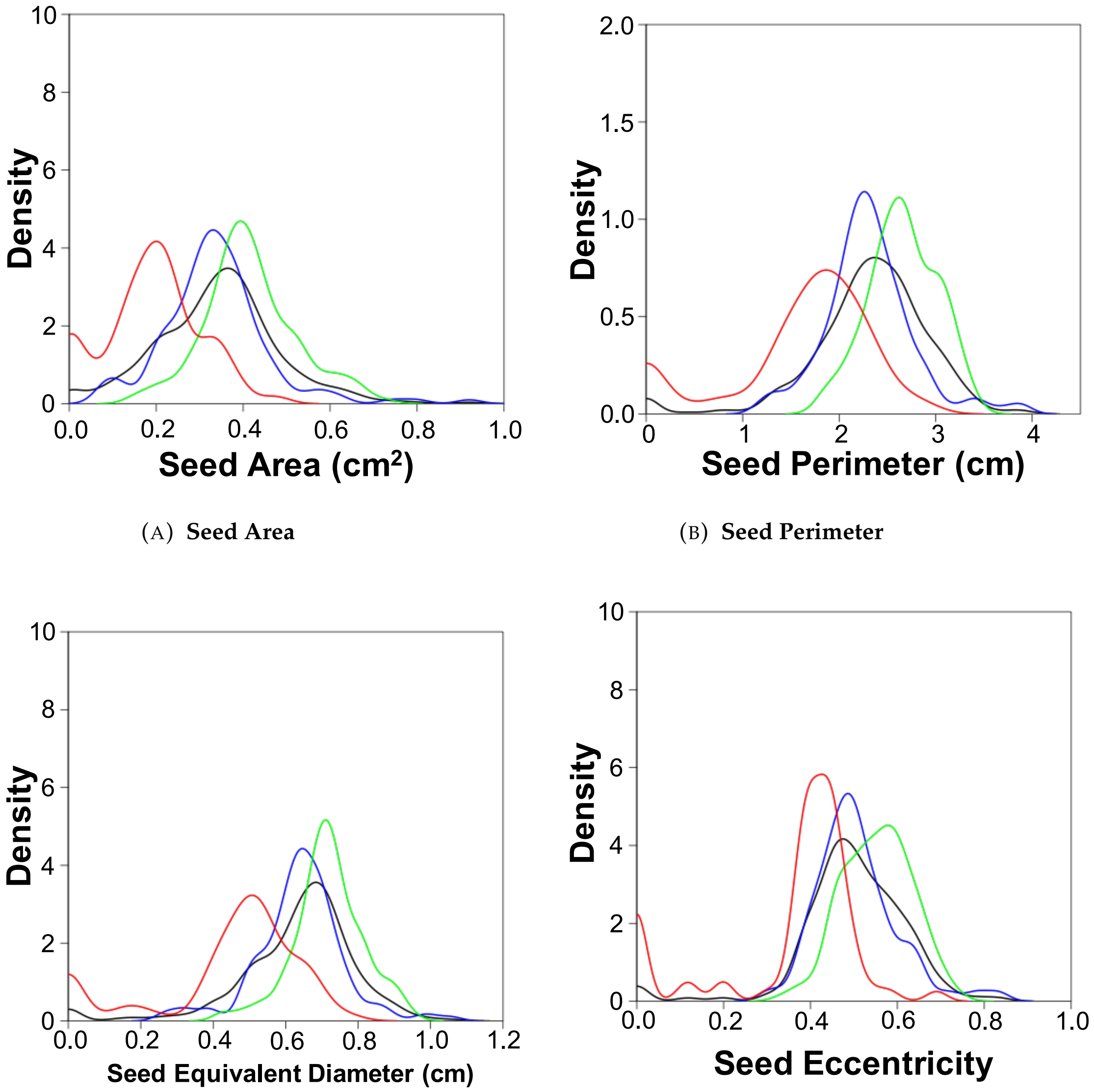
FIGURE 3.13: **MktStall Seed Module**. An example of the steps taken by MktStall Seed Module where each subfigure is an intermediary file produced. **(A)** Raw Input Image, **(B)** HSV conversion, **(C)** Green Mask, **(D)** Ruler Measurement, **(E)** Seed Segmentation by Watershed Algorithm, **(F)** Measuring contours around seed, **(G)** Cropping to Label, **(H)** Label Reading. Original in colour.

The following SMSDs have been measured by MktStall Seed module: area, perimeter, equivalent diameter and eccentricity.

Normality Testing of Seed

Seed area, perimeter, equivalent diameter and eccentricity for peas grown in the glasshouse were statistically tested for normality using the Kolmogorov-Smirnov test with the null hypothesis being there is no difference between the seed SMSDs of the core collection and a normal distribution. None of the seed SMSDs of this core collection were normally distributed (Appendix C).

To visually explore the distribution of each seed SMSD, a kernel density plot was drawn (Figure 3.14). It shows that in all phenotypes, the group to appear the most normal is the landraces. Cultivars are skewed to the right indicating the phenotypic value is larger than landraces and wilds are skewed to left indicating the phenotypic value is smaller than landraces. For all phenotypes, the core collection is not entirely normally distributed. Seed area does appear to be skewed to the left (Figure 3.14a). Seed perimeter does appear the most normal (Figure 3.14b). Seed equivalent diameter appears irregularly distributed (Figure 3.14c). Seed eccentricity appears to hold most of the data between 0.2 and 0.8 (Figure 3.14d). All phenotypes appear to have long tails for the core collection.



(A) **Seed Area**

(B) **Seed Perimeter**

(C) **Seed Equivalent Diameter**

(D) **Seed Eccentricity**

FIGURE 3.14: **Kernel Density Plot for seed phenotypes of peas grown in a glasshouse environment.** The core collection (black) was also separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). The phenotypes explored are (A) Seed Area, (B) Seed Perimeter, (C) Seed Equivalent Diameter and (D) Seed Eccentricity. Original in colour.

Furthermore, Quartile-Quartile (Q-Q) plots for each seed phenotype were drawn to further explore the normality of the seed SMSDs. The reference line shows the normal distribution.

Distension above the reference line on the right-hand side and below the line on the left-hand side indicates long tails in the data. Conversely, distension below the line on the right-hand side and above the line on the left-hand side indicate short tails in the data.

The Q-Q plots indicate long tails in the data for the core collection and each group in seed area (Appendix D.4). The Q-Q plots indicate that seed perimeter data holds long tails in the core collection and group 1 (landrace) whereas short tails are observed in cultivar and wild material (Appendix D.5). Seed Equivalent Diameter has long tails in landrace and cultivar material, but short tails in the core collection and wild material (Appendix D.6). Seed eccentricity has short tails in the cultivars and is close to the normal reference line but long tails are indicated in the core collection as well as landrace and wild material (Appendix D.7).

Comparison of Groups for Seed

For each SMSD, groups were statistically tested to see if groups differed using the non-parametric Kruskal-Wallis statistical test. This non-parametric test was used because the data was not normally distributed. All seed SMSDs (seed area, perimeter, equivalent diameter and eccentricity), were tested with the null hypothesis being that there was no difference between the groups. The Kruskal-Wallis test for all seed SMSDs did not support the null hypothesis (Appendix E), therefore all groups differed for all SMSDs tested.

To determine which groups differed, a post hoc Nemenyi test was used. For all seed SMSDs, all groups are statistically different to one another (Appendix F). The descriptive statistics provide more information on how they differ to one another. From Appendix G, for all SMSDs, Group 2 has the highest value, closely followed by Group 1 and the lowest value is observed in Group 4. This indicates that wild varieties have smaller seeds, in comparison to landrace and cultivars, with cultivars holding the largest seeds out of all of the groups.

TABLE 3.3: **ANOVA for seed.** This denotes degrees of freedom (Df), sum of squares (SS), mean square (MS), F value test statistic and the corresponding p-value.

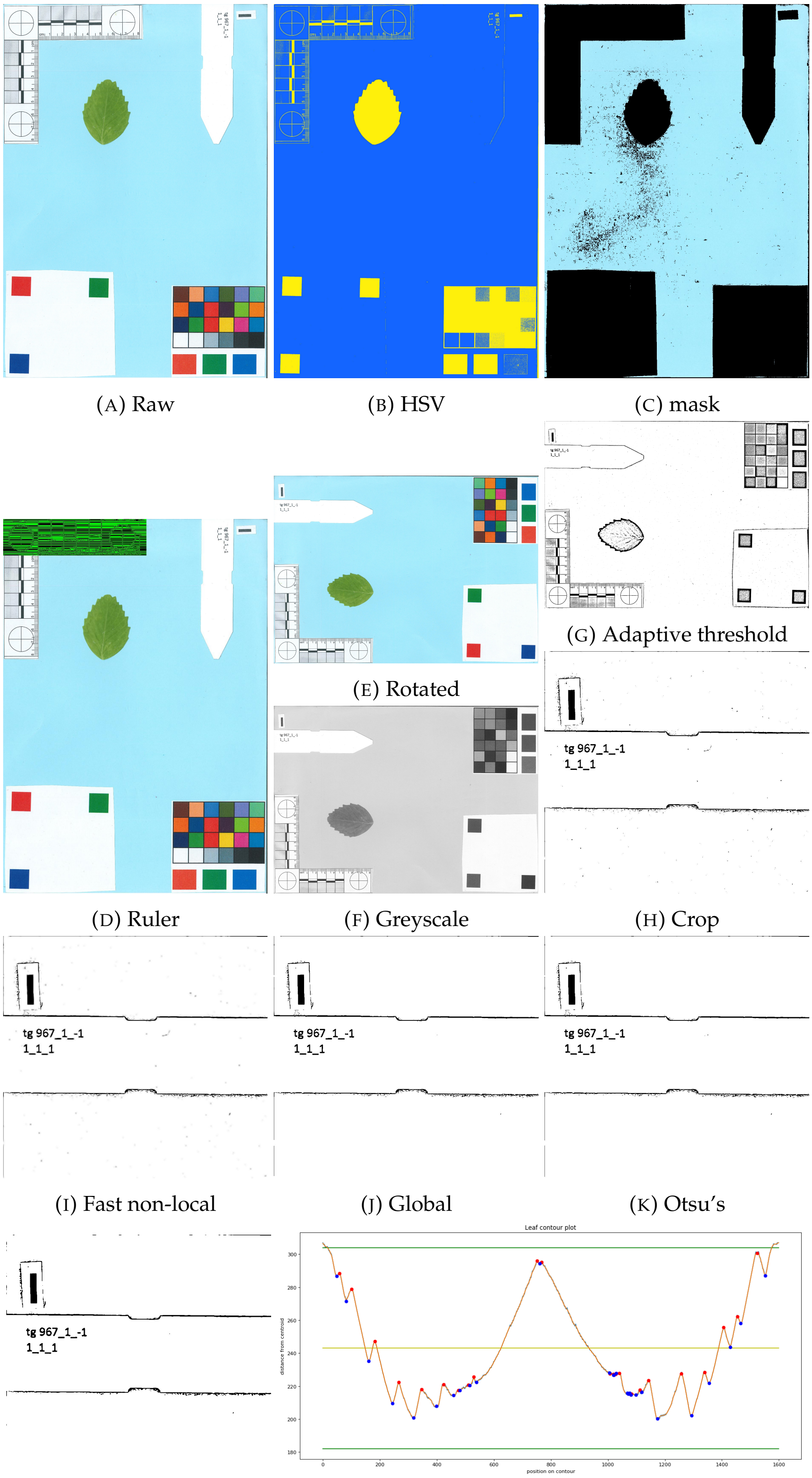
Phenotype		Df	SS	MS	F value	p-value
Area	Replicates	2	0.168	0.08395	2.733	0.0655
	Residuals	1017	31.241	0.03072		
Perimeter	Replicates	2	4.5	2.2337	2.548	0.0787
	Residuals	1017	891.4	0.8765		
Equivalent Diameter	Replicates	2	0.36	0.18098	2.842	0.0588
	Residuals	1017	64.76	0.06367		
Eccentricity	Replicates	2	0.14	0.06969	1.759	0.173
	Residuals	1017	40.28	0.03961		

Analysis of variance for seed

The variability of seed SMSDs was analysed with the null hypothesis being that there is no difference between replicates. The ANOVA results (Table 3.3) suggest that for all seed SMSDs there were differences between the replicates, area (p-value= 0.0655), eccentricity (p-value = 0.173), equivalent diameter (p-value = 0.0588) and perimeter (p-value = 0.0787). This suggests that there may be genetic variability for seed SMSDs.

3.3.4 MktStall - Leaflet

MktStall was used to detect leaf. An example of intermediary files is shown in Figure 3.15.



(A) Raw

(B) HSV

(C) mask

(D) Ruler

(E) Rotated

(F) Greyscale

(G) Adaptive threshold

(H) Crop

(I) Fast non-local

(J) Global

(K) Otsu's

(L) Gaussian and Otsu's

(M) Teeth

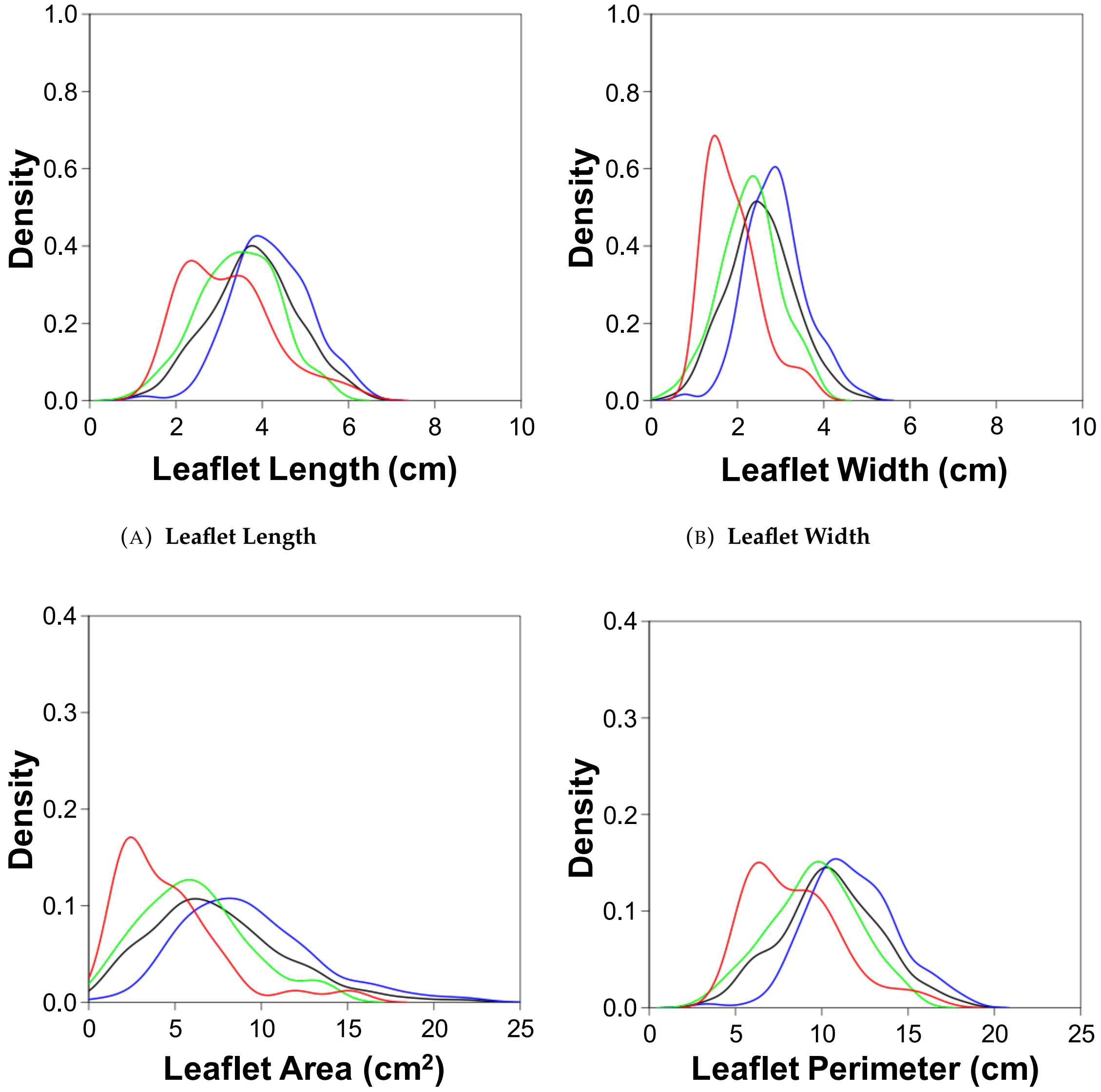
FIGURE 3.15: **MktStall Leaflet Module**. An example of the steps taken by MktStall Leaflet Module where each subfigure is an intermediary file produced. **(A)** Raw Input Image, **(B)** HSV conversion, **(C)** Mask Image, **(D)** Ruler Measurement, **(E)** Rotate, **(F)** Greyscale Conversion, **(G)** Adaptive Thresholding, **(H)** Crop the Label, **(I)** Fast non-local means denoising, **(J)** Global Thresholding, **(K)** Otsu's thresholding, **(L)** Gaussian and Otsu's Thresholding, **(M)** Leaflet Teeth Plot. Original in colour.

Normality testing of leaflet

Leaflet SMSDs for peas grown in the glasshouse were statistically tested for normality using the Kolmogorov-Smirnov test with the null hypothesis being there is no difference between the leaflet SMSDs of the core collection and a normal distribution. Appendix C shows that the statistical tests do not support the null hypothesis.

Kernel density plots were drawn to visualise the distribution of leaflet SMSDs (Figures 3.16, 3.17 and 3.18). Leaflet teeth and leaflet perimeter ratio of length and width (PRLW) display many irregularities. Whilst none of the SMSDs appeared normal - length, width, perimeter, area, roundness, rectangularity, perimeter ratio of length (PRL) and perimeter ratio of length and width (PRLW) show the wild varieties are distributed to the left indicating they are smaller in value, the landrace varieties in the middle and the cultivars are distributed to the right indicating they are larger. Whereas, aspect ratio and compactness display landrace to be distributed to the left indicating they are smaller, cultivars in the middle and wild distributed to the right suggesting they are larger in value.

Furthermore, Quartile-Quartile (Q-Q) plots for each seed phenotype were drawn to further explore the normality of the leaflet SMSDs. For all SMSDs, the wild varieties displayed a sparse distribution with leaflet aspect ratio and leaflet rectangularity closely following the reference line for Group 4 peas (Appendix D). For the core collection, both leaflet roundness and compactness display a curved Q-Q plot (Figures D.14 and D.15) suggesting that the data for these SMSDs are skewed, this is also corroborated in the kernel density plots (Figure 3.17). The Q-Q plots show that leaflet length width, perimeter, rectangularity, PRL and PRLW and equivalent diameter closely follow the reference line suggesting they are close to a normal distribution (Figures D.8, D.9, D.10, D.16, D.17, D.18, and D.19) but of these, width, PRL, PRLW have a long right-hand tail whereas rectangularity has a long left-hand tail. Leaflet area has long tails on either side of the distribution (Figure D.11) and aspect ratio (Figure D.13) only a long right-hand tail. Leaflet teeth has a plateau and a staircasing distribution, due to most leaves having very few teeth and the fact that teeth number is a discrete variable (Figure D.12).



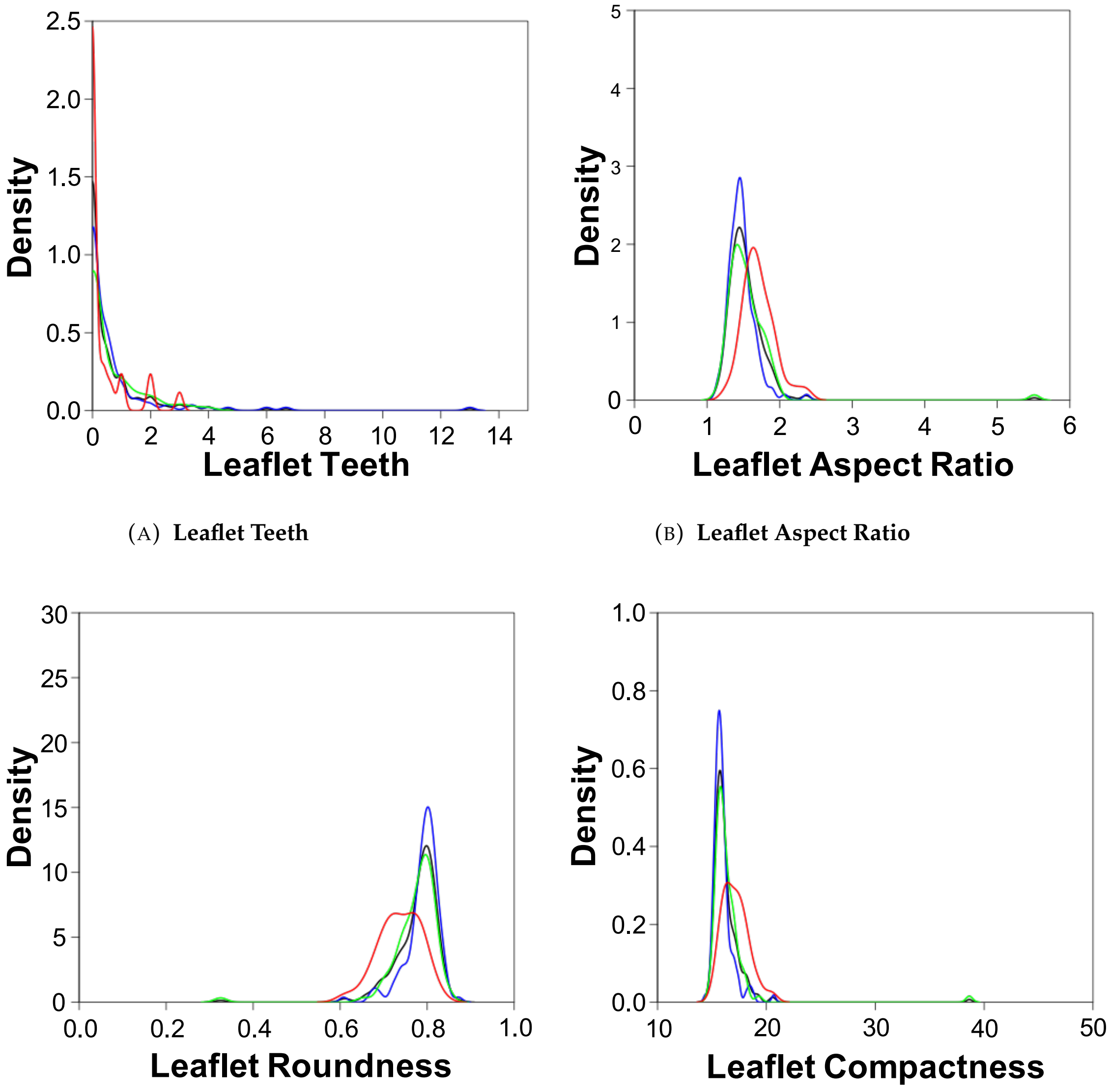
(A) **Leaflet Length**

(B) **Leaflet Width**

(C) **Leaflet Area**

(D) **Leaflet Perimeter**

FIGURE 3.16: **Kernel Density Plot for leaflet SMSDs of peas grown in a glasshouse environment.** The core collection (black) was also separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). The SMSDs shown are (A) Leaflet Length, (B) Leaflet Width, (C) Leaflet Area and (D) Leaflet Perimeter. Original in colour.



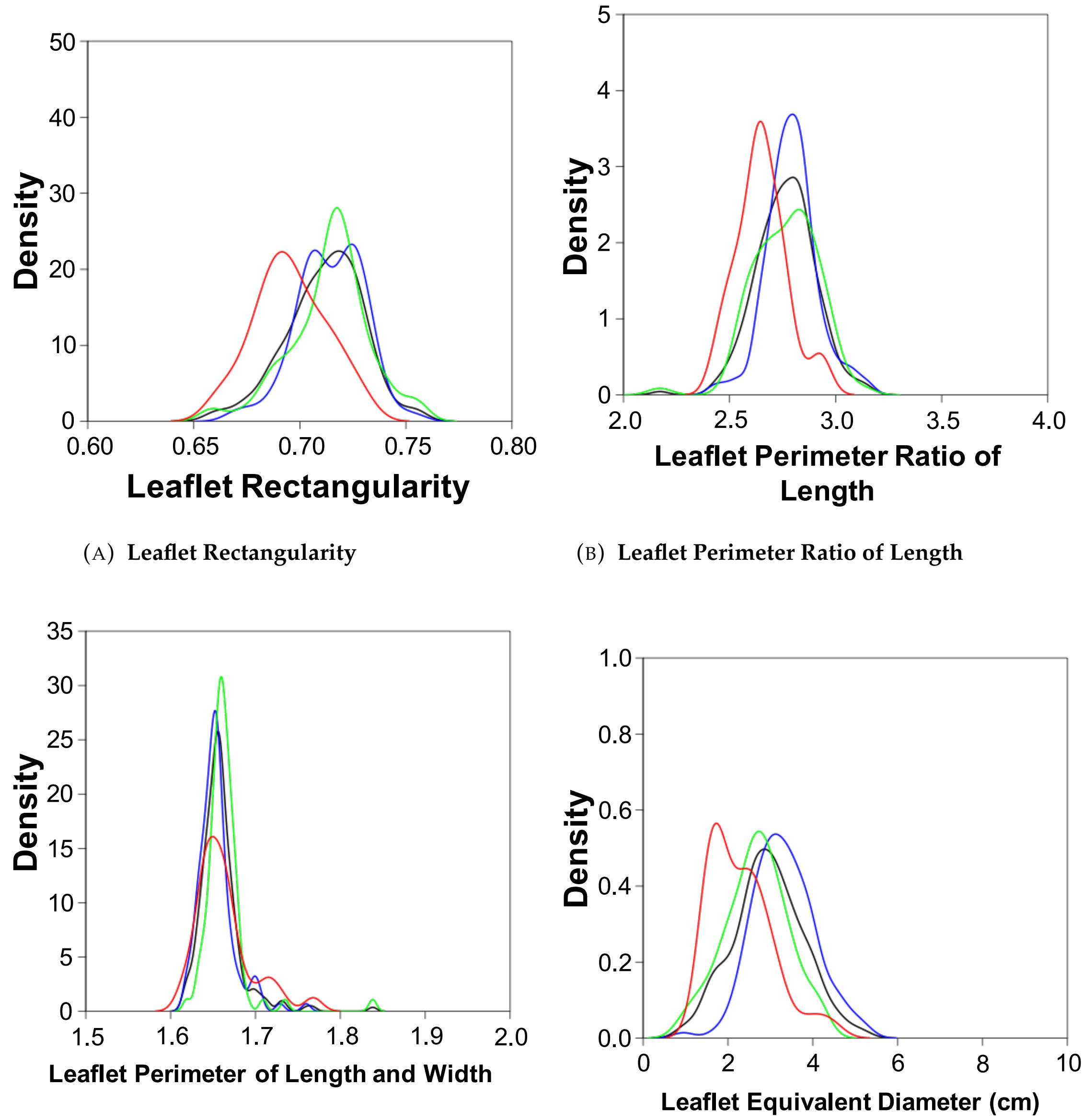
(A) **Leaflet Teeth**

(B) **Leaflet Aspect Ratio**

(C) **Leaflet Roundness**

(D) **Leaflet Compactness**

FIGURE 3.17: **Kernel Density Plot for leaflet SMSDs of peas grown in a glasshouse environment.** The core collection (black) was also separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). The SMSDs shown are (A) Leaflet Teeth, (B) Leaflet Aspect Ratio, (C) Leaflet Roundness and (D) Leaflet Compactness. Original in colour.



(A) **Leaflet Rectangularity**

(B) **Leaflet Perimeter Ratio of Length**

(C) **Leaflet Perimeter Ratio of Length and Width**

(D) **Leaflet Equivalent Diameter**

FIGURE 3.18: **Kernel Density Plot for leaflet SMSDs of peas grown in a glasshouse environment.** The core collection (black) was also separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). The SMSDs shown are (A) Leaflet Rectangularity, (B) Leaflet Perimeter Ratio of Length, (C) Leaflet Perimeter Ratio of Length and Width and (D) Leaflet Equivalent Diameter. Original in colour.

Comparison of groups for leaflet

Leaflet SMSDs were statistically tested using the Kruskal-Wallis test to determine if the groups differed. The null hypothesis tested was that there was no difference between the groups. For all leaflet SMSDs, except leaflet teeth, there was no statistical support for the null hypothesis, indicating that there were differences between groups for all SMSDs bar leaflet teeth (Appendix E).

To determine which of these groups differed a post hoc Nemenyi test was used. For leaflet width, roundness and compactness - all groups differed to one other (Appendix F). Leaflet perimeter ratio of length and width (PRLW) differed between landraces and cultivars (Appendix F). For length, equivalent diameter, area and perimeter landraces differed to the other two groups (Appendix F) but cultivars and wild appear to be similar. For aspect ratio, rectangularity and perimeter ratio of length (PRL) - the wild varieties were statistically different to the other groups. Leaflet teeth show no difference between groups (Appendix F).

The descriptive statistics (Appendix G) show that for several leaflet SMSDs, landrace had the largest mean, followed by the cultivars and the wild varieties had the smallest mean; these SMSDs were length, width, perimeter, area, teeth, roundness, rectangularity and PRL.

The SMSDs which were exceptions to this observation are aspect ratio and compactness where wild varieties had the largest mean, followed by cultivars and landraces had the smallest mean (Appendix G). Perimeter ratio of length and width is also an exception with the largest mean in cultivar material, followed by wild and the smallest means in landrace, however, the range of these results is small (Appendix G). The mean for the equivalent diameter is largest in landrace, cultivar and wild in descending order (Appendix G).

Analysis of variance for leaflet

The variability of leaflet SMSDs was analysed with the null hypothesis for each SMSD is that there is no difference between replicates. The ANOVA results for leaflet suggest that

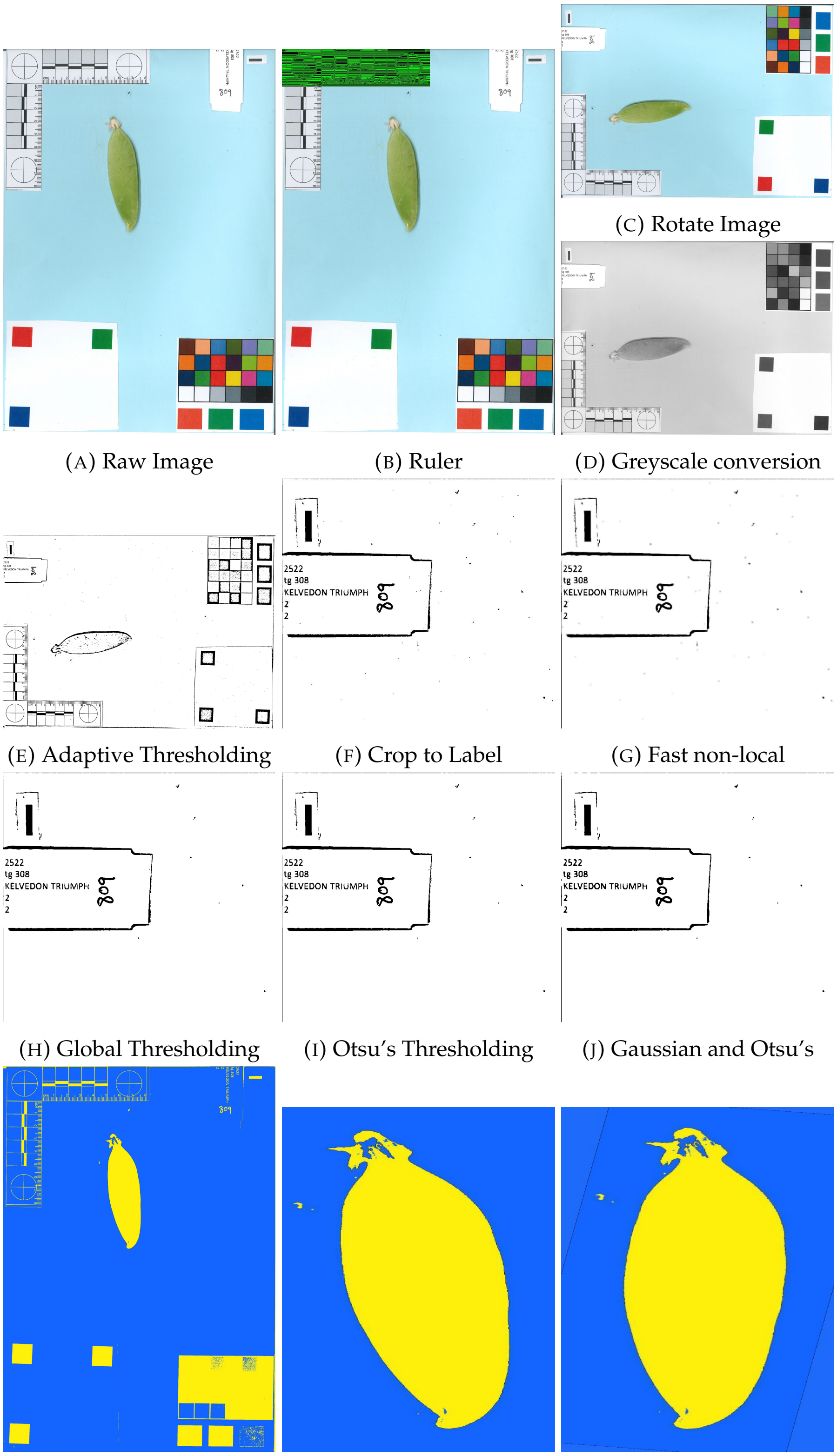
TABLE 3.4: **ANOVA for Leaflet SMSDs.** This denotes degrees of freedom (Df), sum of squares (SS), mean square (MS), F value test statistic and the corresponding p-value.

Phenotype		Df	SS	MS	F value	p-value
Length	Replicates	2	57	28.494	7.436	0.00064
	Residuals	666	2552	3.832		
Width	Replicates	2	26.4	13.181	7.131	0.000863
	Residuals	666	1231.1	1.848		
Perimeter	Replicates	2	448	224.00	7.54	0.000578
	Residuals	666	19787	29.71		
Area	Replicates	2	320	159.88	6.098	0.00238
	Residuals	666	17462	26.22		
Teeth	Replicates	2	0.9	0.4499	0.305	0.737
	Residuals	666	983.4	1.4766		
Aspect Ratio	Replicates	2	7.1	3.538	6.793	0.0012
	Residuals	666	346.9	0.521		
Roundness	Replicates	2	1.77	0.8846	7.25	0.000767
	Residuals	666	81.26	0.1220		
Compactness	Replicates	2	797	398.5	7.387	0.000672
	Residuals	666	35927	53.9		
Rectangularity	Replicates	2	1.54	0.7712	7.65	0.000519
	Residuals	666	67.14	0.1008		
PRL	Replicates	2	23.2	11.597	7.555	0.00057
	Residuals	666	1022.3	1.535		
PRLW	Replicates	2	8.3	4.137	7.579	0.000556
	Residuals	666	363.5	0.546		
Equivalent Diameter	Replicates	2	35.5	17.729	7.465	0.000622
	Residuals	666	1581.7	2.375		

there are differences between the replicates for leaflet teeth (p-value = 0.737) and not for any other leaflet SMSD (Table 3.4). This suggests there maybe genetic variability for leaflet teeth.

3.3.5 MktStall- Pod

MktStall was used to detect pod length, width, area, perimeter, aspect ratio, rectangularity and equivalent diameter. An example of the intermediary files is shown in Figure 3.19.



(A) Raw Image (B) Ruler (C) Rotate Image (D) Greyscale conversion

(E) Adaptive Thresholding (F) Crop to Label (G) Fast non-local

(H) Global Thresholding (I) Otsu's Thresholding (J) Gaussian and Otsu's

(K) HSV conversion (L) Crop to Pod (M) Orientate

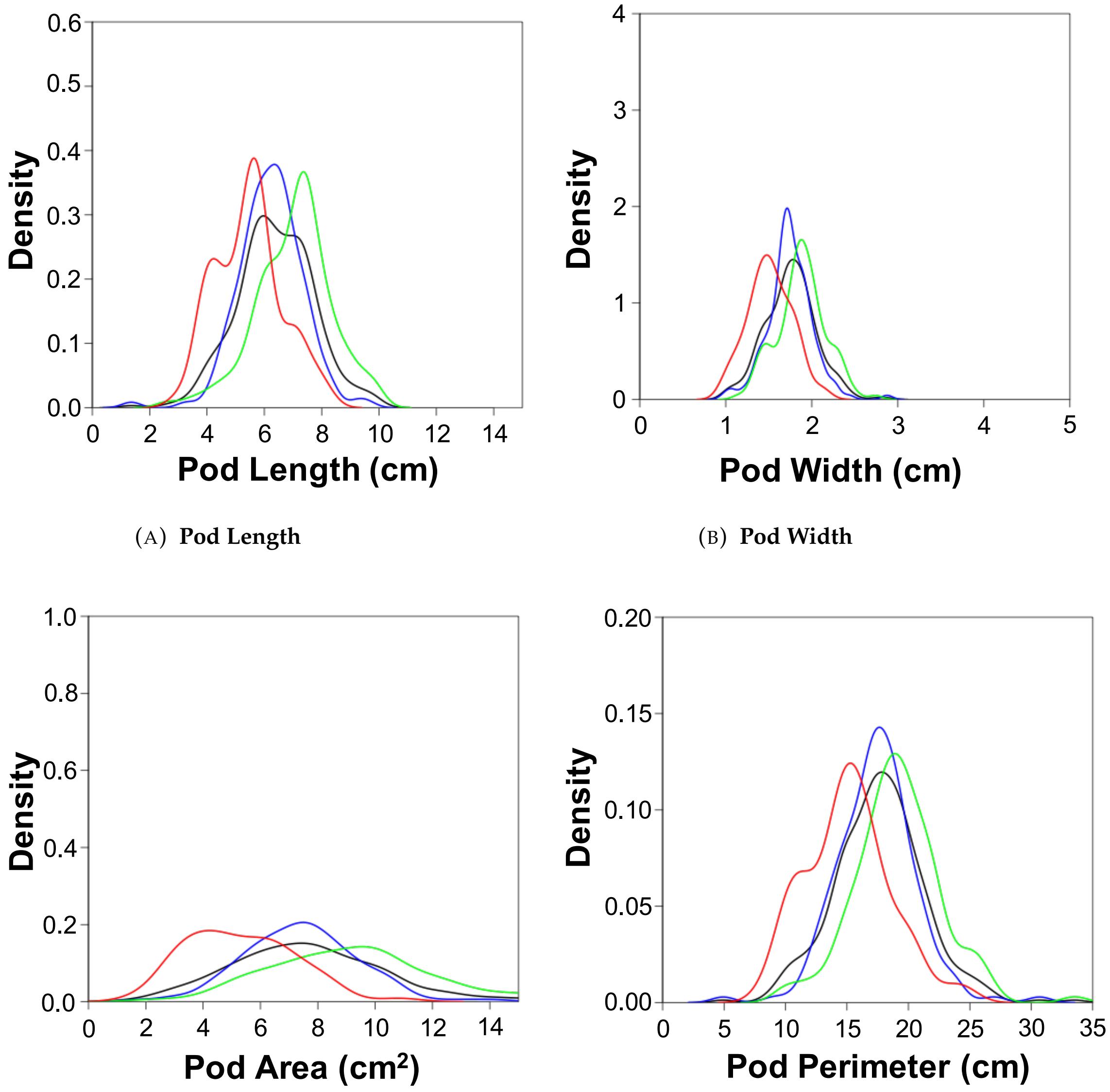
FIGURE 3.19: **MktStall Pod Module.** An example of the steps taken by MktStall Pod Module where each subfigure is an intermediary file produced. **(A)** Raw, **(B)** Ruler, **(C)** Rotate, **(D)** Greyscale, **(E)** Adaptive Thresholding, **(F)** Crop, **(G)** Fast Non-local Means Thresholding, **(H)** Global Thresholding, **(I)** Otsu's Thresholding, **(J)**Gaussian and Otsu's Thresholding, **(K)** HSV, **(L)** Crop, **(M)** Orientate. Original in colour.

Normality testing of Pods

Pod length, width, area, perimeter, aspect ratio, rectangularity and equivalent diameter for peas grown in the glasshouse were statistically tested for normality using the Kolmogorov-Smirnov test with the null hypothesis being there is no difference between the pod SMSDs of the core collection and a normal distribution. None of the pod SMSDs of this core collection were normally distributed (Appendix C).

To visually explore the distribution of each pod SMSD, a kernel density plot was drawn (Figures 3.20 and 3.21). For each phenotype the core collection does not appear to be normally distributed, however, out of the groups, the landraces (blue) appear the most normally distributed group with some irregularity in the tails. Cultivars appear to the right of the landraces indicating their larger values whereas wild appears to the left of the landraces and hold more irregularities.

Furthermore, Quartile-Quartile (Q-Q) plots for each seed phenotype were drawn to further explore the normality of the seed SMSDs. As described in Section 3.3.2, this can provide information on the tails of the data. Pod length appears to follow the reference line with a long tail indicated in the left-hand tail in the core collection (Appendix D.20). Pod width, area and aspect ratio indicate long tails on both sides of the distribution for the core collection (Appendices D.21, D.22 and D.24). Whereas, pod perimeter has a long right-hand tail for the core collection (Appendix D.23). Pod rectangularity displays a curve in the core collection as well as landrace and cultivars (Groups 1 and 2) indicating the data is skewed to the left (Appendix D.25). Pod equivalent diameter closely follows the normal reference line (Appendix D.26) despite the Kolmogorov-Smirnov test indicating it is not normally distributed (Appendix C).



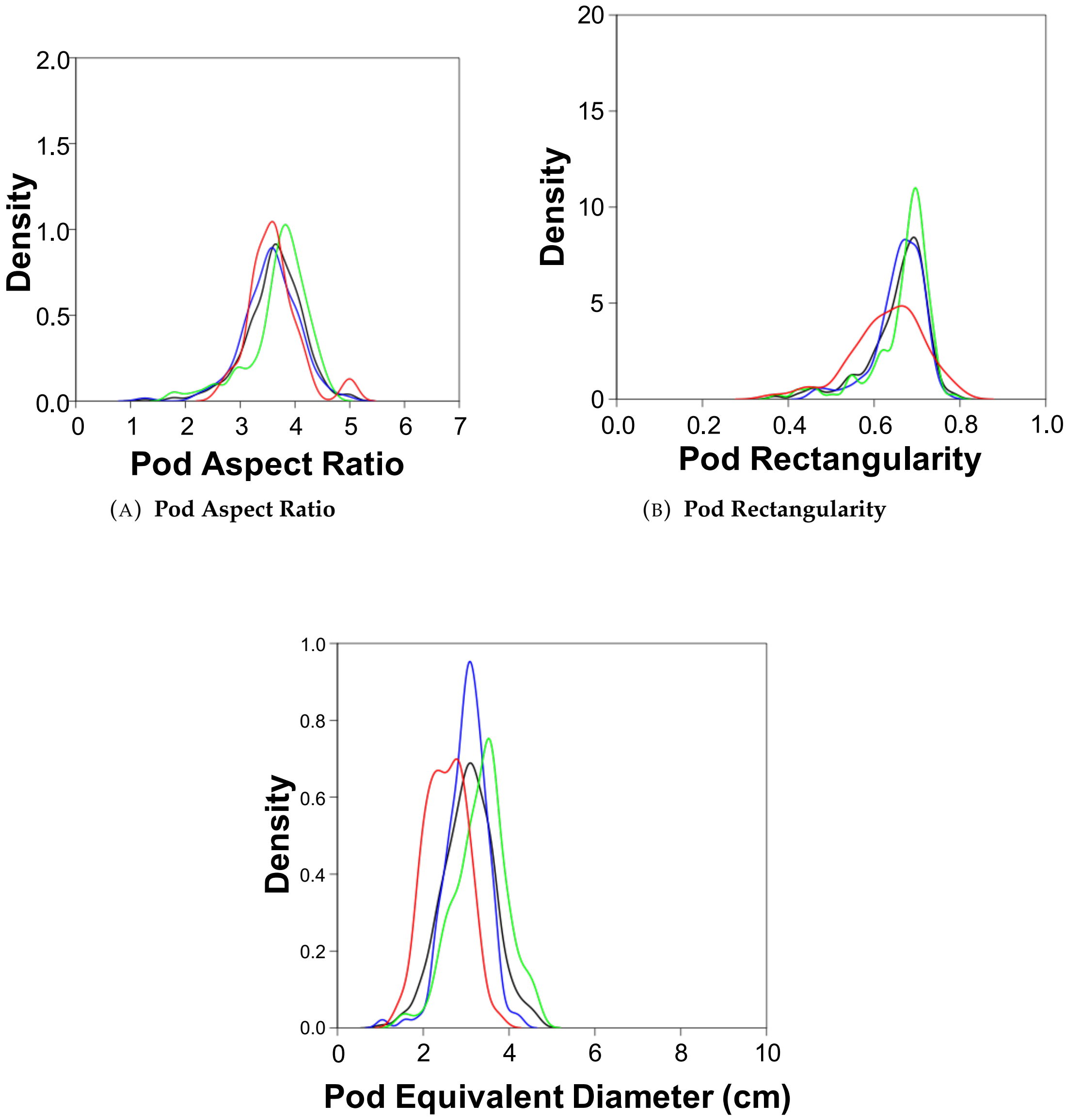
(A) **Pod Length**

(B) **Pod Width**

(C) **Pod Area**

(D) **Pod Perimeter**

FIGURE 3.20: **Kernel Density Plot for pod SMSDs of peas grown in a glasshouse environment.** The core collection (black) was also separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). The SMSDs shown are (A) Pod Length, (B) Pod Width, (C) Pod Area and (D) Pod Perimeter. Original in colour.



(A) **Pod Aspect Ratio**

(B) **Pod Rectangularity**

(C) **Pod Equivalent Diameter**

FIGURE 3.21: **Kernel Density Plot for pod SMSDs of peas grown in a glasshouse environment.** The core collection (black) was also separated into groups: group 1 (landrace in blue), group 2 (cultivar in green), group 4 (wild in red). The pod SMSDs shown are (A) Pod Aspect Ratio, (B) Pod Rectangularity and (C) Pod Equivalent Diameter. Original in colour.

Comparison of groups for pods

Pod SMSDs were statistically tested using the Kruskal-Wallis test to determine if the groups differed. The null hypothesis tested was that there was no difference between the groups. For all pod SMSDs tested, the Kruskal-Wallis test did not support the null hypothesis (Appendix E). This indicates all groups differed.

To identify which groups statistically differed, a post hoc Nemenyi test was used. For all pod SMSDs, except aspect ratio and rectangularity, all groups differed to one another (Appendix F). Landrace and wild material held a statistically significant p-value for pod aspect ratio indicating these are not statistically different (Appendix F). For pod rectangularity, landrace was not found to be statistically different to wild or cultivar (Appendix F). The descriptive statistics also indicate a tendency for cultivars to hold a higher pod SMSD value than landrace and landrace holding a higher pod SMSD than wild. This indicates that domesticated varieties of pea have larger pods (Appendix G).

Analysis of variance for pod

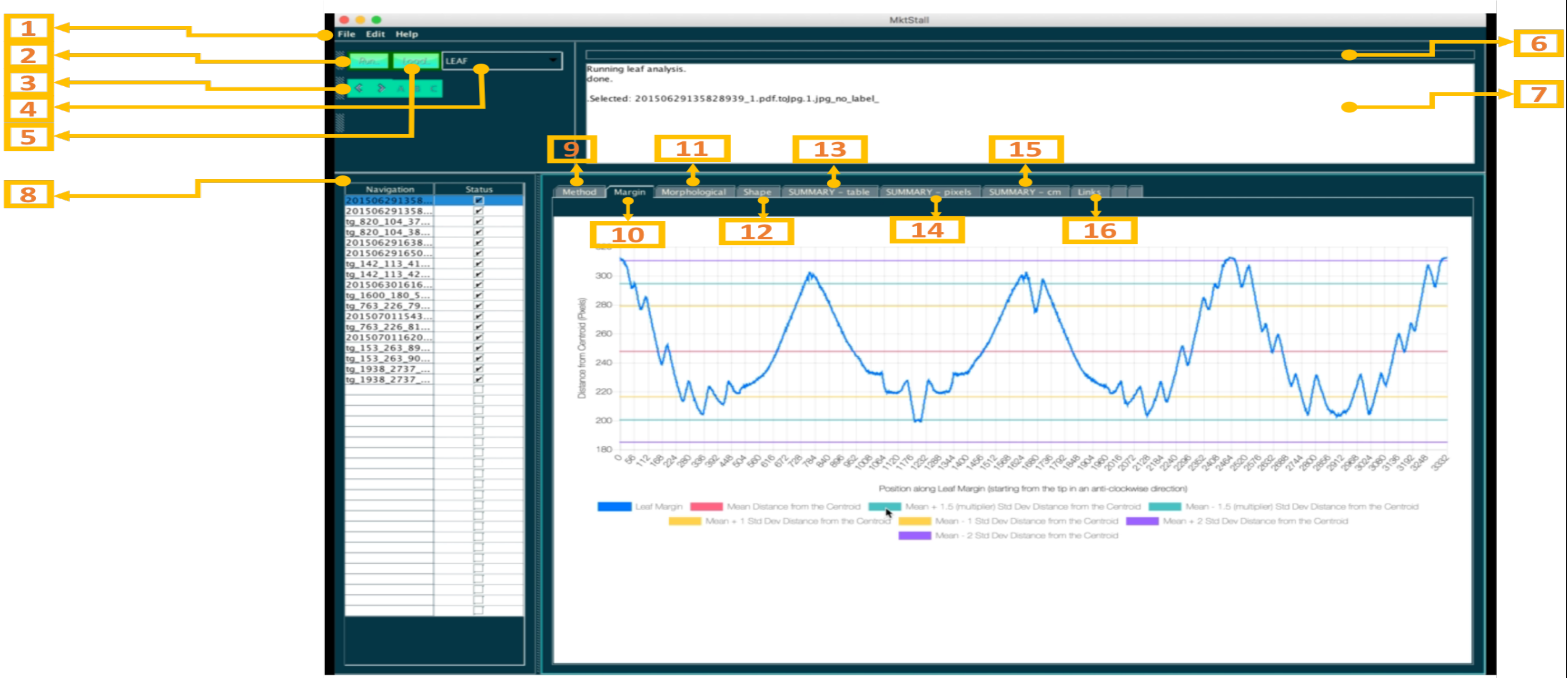
The variability of pod SMSDs was analysed with the null hypothesis for each SMSD being that there is no difference between replicates. The ANOVA results for pod suggest that for all pod phenotypes observed there are differences between the replicates (p-values range between 0.163- 0.819). This suggests that there may be genetic variability for pod phenotypes.

TABLE 3.5: **ANOVA for Pod SMSDs.** This denotes degrees of freedom (Df), sum of squares (SS), mean square (MS), F value test statistic and the corresponding p-value.

Phenotype		Df	SS	MS	F value	p-value
Length	Replicates	2	6	3.206	0.429	0.651
	Residuals	981	7328	7.470		
Width	Replicates	2	0.9	0.4485	0.886	0.413
	Residuals	981	496.7	0.5063		
Area	Replicates	2	6	3.205	0.199	0.819
	Residuals	981	15781	16.087		
Perimeter	Replicates	2	37	18.41	0.317	0.728
	Residuals	981	56996	58.10		
Aspect Ratio	Replicates	2	4.2	2.102	1.007	0.366
	Residuals	981	2047.4	2.087		
Rectangularity	Replicates	2	0.23	0.11440	1.818	0.163
	Residuals	981	61.72	0.06291		
Equivalent Diameter	Replicates	2	2.4	1.220	0.744	0.476
	Residuals	981	1609.4	1.641		

3.3.6 MktStall GUI

The above results can also be visualised in the MktStall GUI shown in Figure 3.22.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
MktStall
File Edit Help
Run...
Load...
LEAF
A B C
Method
Margin
Morphological
Shape
SUMMARY - table
SUMMARY - pixels
SUMMARY - cm
Links
Distance from Centroid (Pixels)
300
280
260
240
220
200
180
Position along Leaf Margin (starting from the tip in an anti-clockwise direction)
Leaf Margin
Mean Distance from the Centroid
Mean + 1.5 (multiplier) Std Dev Distance from the Centroid
Mean - 1.5 (multiplier) Std Dev Distance from the Centroid
Mean + 1 Std Dev Distance from the Centroid
Mean - 1 Std Dev Distance from the Centroid
Mean + 2 Std Dev Distance from the Centroid
Mean - 2 Std Dev Distance from the Centroid

Running leaf analysis.
done.

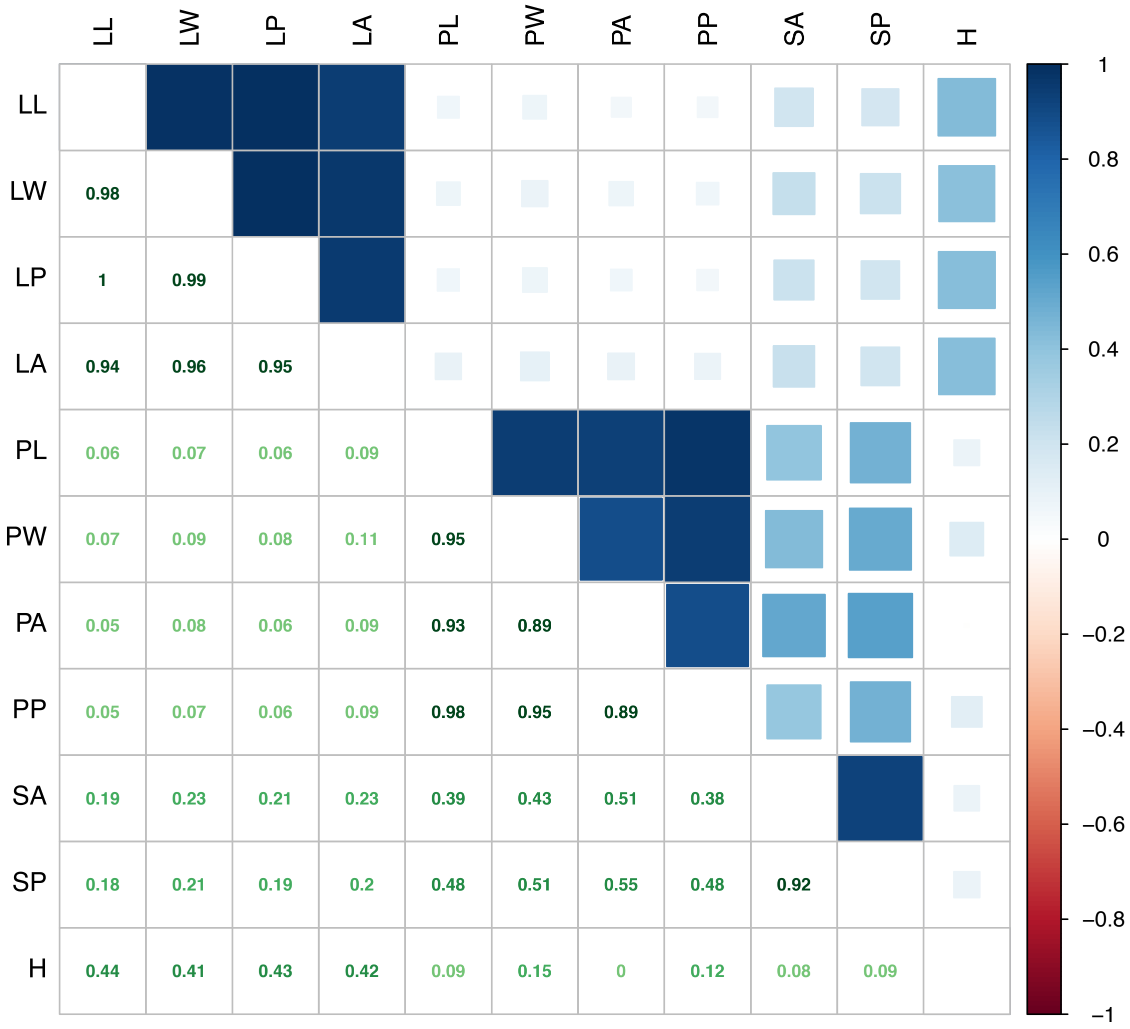
.Selected: 20150629135828939_1.pdf.toJpg.1.jpg_no_label_

Navigation	Status
201506291358...	☑
201506291358...	☑
tg_820_104_37...	☑
tg_820_104_38...	☑
201506291638...	☑
201506291650...	☑
tg_142_113_41...	☑
tg_142_113_42...	☑
201506301616...	☑
tg_1600_180_5...	☑
tg_763_226_79...	☑
201507011543...	☑
tg_763_226_81...	☑
201507011620...	☑
tg_153_263_89...	☑
tg_153_263_90...	☑
tg_1938_2737_...	☑
tg_1938_2737_...	☑

FIGURE 3.22: **MktStall Graphical User Interface (GUI).** An annotated diagram of the MktStall GUI shows **(1)** Tool Bar, **(2)** Run, **(3)** Toggle Buttons, **(4)** Organ Selection, **(5)** Load Button, **(6)** Progress Bar, **(7)** Progress Message, **(8)** Navigation Bar of Accessions, **(9)** Method Tab, **(10)** Margin Tab, **(11)** Morphological Tab, **(12)** Shape Descriptor Tab, **(13)** Summary Table of all Accessions Tab, **(14)** Summary Plots of all Accessions in pixels Tab, **(15)** Summary Plots of all Accessions in centimetres Tab, **(16)** Links to Interactive Plots. Original in colour.

3.3.7 Exploring the trait space

A correlation plot in Figure 3.23 of the key phenotypes observed indicates that plant height and leaf traits are highly correlated with one another (p-values range from 0.41 to 0.44), likewise, pod and seed traits are also highly correlated to one another (p-values range from 0.39 to 0.55). To a lesser extent, leaf and seed traits are also correlated to one other (p-values range from 0.18 to 0.23). Organ traits are highly correlated with other traits from the same organ. This can be seen by the dark blue clusters in Figure 3.23.



	LL	LW	LP	LA	PL	PW	PA	PP	SA	SP	H
LL											
LW	0.98										
LP	1	0.99									
LA	0.94	0.96	0.95								
PL	0.06	0.07	0.06	0.09							
PW	0.07	0.09	0.08	0.11	0.95						
PA	0.05	0.08	0.06	0.09	0.93	0.89					
PP	0.05	0.07	0.06	0.09	0.98	0.95	0.89				
SA	0.19	0.23	0.21	0.23	0.39	0.43	0.51	0.38			
SP	0.18	0.21	0.19	0.2	0.48	0.51	0.55	0.48	0.92		
H	0.44	0.41	0.43	0.42	0.09	0.15	0	0.12	0.08	0.09	

FIGURE 3.23: **Correlation matrix of phenotypes (top) and corresponding p-values (bottom).** Phenotypes include leaflet length (LL), width (LW), perimeter (LP) and area (LA); pod length (PL), width (PW), area (PA) and perimeter (PP), as well as, seed area (SA), seed perimeter (SP) and plant height (PH). Original in colour.

The biplot in Figure 3.24 corresponds closely with the correlation matrix in Figure 3.23, as the biplot also indicates plant height and leaflet traits being positively correlated, and likewise, seed and pods being positively correlated due to the loading values clustering together. Although all groups overlap somewhat, landraces in group 1 appear to be pointing in an opposite direction to cultivars in groups 2 and wild varieties in group 4.

Supplementary data which contributed to the biplot is included in Appendix H. Appendix H.1 and H.2 shows the contribution of variables to dimension 1 and 2 respectively. Pods are the highest contribution to the first dimension and leaflet is the highest contributor to the second dimension. Appendix H.3 is a scree plot that illustrates the contribution of each dimension. A principle components analysis of all individuals is found in Appendix H.4 which shows distinct clustering in the landraces (group 1). The cultivars (group 2) are found close to the landraces but a sub-cluster found close to wild (group 4) relatives too.

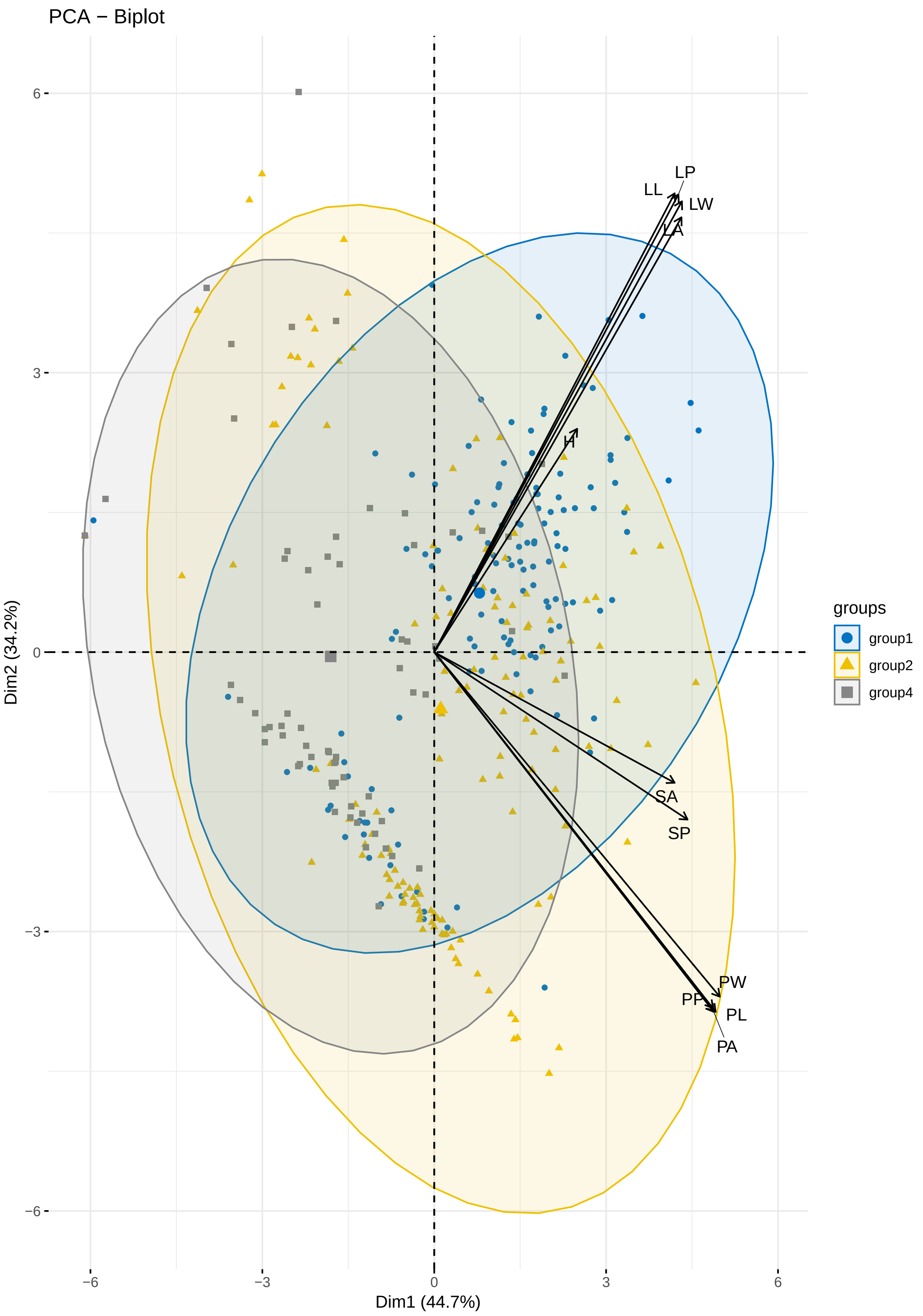


FIGURE 3.24: **Biplot of phenotypes.** Groups 1 (landrace), 2 (cultivar) and 4 (wild) are shown for phenotypes, leaflet length (LL), width (LW), perimeter (LP) and area (LA); pod length (PL), width (PW), area (PA) and perimeter (PP), as well as, seed area (SA), seed perimeter (SP) and plant height (H). Original in colour.

3.4 Discussion

3.4.1 Major Findings

MktStall is a novel multi-organ image analysis tool which requires no prior knowledge of computing. It measures phenotypes in accordance with the Plant Trait Ontology, as well as providing human readable data for the reading of labels and for the measurement of phenotypes in centimetres instead of pixels. This tool is a high-throughput fully-automated tool designed to improve on reproducibility in comparison to other general tools. The phenotypes obtained from this tool and from manual measurements are described below.

Our findings show the core collection is not normally distributed for all phenotypes; this is an important finding since all subsequent statistical methods should use non-parametric testing. Chapter 2 found the JIC domain was normally distributed, and although a representative collection was formed based on seed weight, some accessions were not available in the seed bank, resulting in substitutions of preferred accessions, this is the likely cause of the core collection not being normally distributed.

Seed

The results of the seed weight of our collection is as expected; field seed weight is smaller than glasshouse seed weight which may be due to environmental pressures. The cultivars and landraces have larger seed weight compared to wild material (Figures 3.10 and 3.11), also as expected. Furthermore, the seed weight is statistically different to the JIC collection and whilst this maybe due to the environmental pressures for plants grown in the field, it does not explain the changes in glasshouse, where plants are largely shielded from such environmental pressures. The results also found no statistical difference between groups.

All seed phenotypes (seed weight, seed area, seed perimeter, seed equivalent diameter and seed eccentricity) were found to differ statistically between all groups. This indicates a low level of similarity between groups for the above seed traits. However, the seed traits measured in this chapter all display similar trends with Group 4 (wild) holding smallest value, Group 1 being larger (landrace) and Group 2 (cultivar) holding the highest value. This indicates a trend towards larger seeds through as peas become more domesticated. This observation is well-known (Weeden, 2007).

Plant height

The results found that plant height is not normally distributed. The cultivated group differed statistically significantly in plant height to landrace and wild groups, however, landrace and wild groups were not statistically significantly different. This shows that Group 2 differed whereas Group 1 and 4 groups were similar and may be due to domestication preferences with respect to plant height. Plant height appears to be a domesticated trait towards smaller plants, possibly due to lodging resistance (Tar'an et al., 2003).

Leaflet

For leaflet length, width, perimeter and area, landraces were found to differ statistically significantly in comparison to cultivated or wild material. This indicates that in early breeding of pea, leaflet length, width, perimeter and area may have been a desirable trait. The potential reasons for these traits being selected as a trait of interest could include increasing photosynthetic capacity and transpiration rates. However, in all cases, cultivars hold lower values for these traits, suggesting that these traits may not have been of interest to modern breeding practices. Landraces need to be hardier than cultivars since early agricultural technology did not nurture the plant as well as modern practices, and therefore, increases in leaflet phenotypes that increase size could help the plant be a better competitor than its wild relatives. In cultivated material, the values are reduced compared to landraces but larger than the wild, suggesting these phenotypes are of some value to the plant but not deliberately selected for by breeders.

For leaflet aspect ratio, rectangularity and perimeter ratio of length, wild material was found to differ statistically significantly in comparison to domesticated material in landraces and cultivars. This suggests that these traits may have been selected against, either through deliberate selection or through inheritance alongside deliberately selected traits. These traits provide information on the shape of the leaflet: high aspect ratio shows the leaf is longer than it is wider, high rectangularity shows how much more of the leaf fits a rectangle drawn around it and high perimeter ratio of length shows much larger the perimeter is to its length. These results suggest that domesticated material have been bred to have lower aspect ratio and higher rectangularity, perhaps because these descriptors can affect the area of the leaf and, perhaps, affect photosynthetic surface area.

Pod

For pod length, width, area, perimeter and equivalent diameter - all groups statistically differ, suggesting that there is a difference in pods as they become more domesticated. Reasons for this could include the need for larger pods to accommodate more seed in order to increase yield. Furthermore, as seen above, seeds begin to increase in size as they become more domesticated and this could also influence pod size.

Pod aspect ratio showed that group 1 (landrace) and 4 (wild) were similar, meaning that they hold similar length to width ratios. This further indicates that Group 2 (cultivars) have larger pods and that this could be due to domestication.

Correlation of phenotypes

The findings also indicate correlation between phenotypes, suggesting that they may be inherited together. Figures 3.23 and 3.24 suggest that plant height and leaf phenotypes are correlated and also suggest pod phenotypes and seed phenotypes are also correlated. Potential reasons for this maybe that as the leaf gets bigger, more nutrients can be expended on growth or that larger pods provide more space for seeds to grow. Additionally, the genes for these traits might be in linkage disequilibrium with one another, and therefore inherited together. Whilst this does not explain the positive correlation between these traits, it does suggest that perhaps if one trait was selected it would not be unreasonable to expect its correlated phenotype to be exhibited in the plant too.

3.4.2 Application to wider contexts

Domestication

It is known that domesticated peas have shortened plant height (Weeden, 2007) and the results corroborate this (Figure 3.9, Appendix E and F). Domesticated peas are also known to have large seeds and the results (Figures 3.10 and 3.11) suggest the same, as seed weight is markedly smaller in wild varieties than in landrace or cultivars.

Phenotypic Plasticity

Figure 3.12 suggests that seed weight can sometimes change under environmental conditions, suggesting possible **phenotypic plasticity**. Most varieties grown in field environments exhibited smaller seed weights, suggesting the plant is adapting to the environment. Environmental pressures can occur in the field, such as salt stress which can decrease seed weight (Okçu, Kaya, and Atak, 2005), water availability during reproductive development can increase seed weight (Fougereux et al., 1997) and high temperature can reduce seed weight (McDonald and Paulsen, 1997). This may possibly be caused by diverting nutrients away from the seed during reproductive development, and towards the roots to improve survival. One study observed changes in environmental conditions corresponded with a change the pea's proteome, particularly in proteins involved in seed storage such as vicilin and therefore changes in seed weight (Bourgeois et al., 2009). This study suggest abiotic stresses can cause phenotypic plasticity and that this could also impact on pea nutritional content.

3.4.3 Limitations

There are some limitations to this study, caused by sample drop out, in other words, the loss of phenotypic data for some replicates. Furthermore, not all phenotypes were observed in field. The reduction in sample size will affect statistical power as the **central limit theorem** dictates that as sample size increases, sample mean and standard deviation gets closer to the actual population mean and standard deviation and become normally distributed (Ennos, 2007). In some cases, the Kolmogorov-Smirnov test is prone to Type II errors (Ennos, 2007), the failure to detect differences that are statistically significant. This problem is resolved by using kernel density plots and Q-Q plots to explore the shape and the fit to normality. Additionally, non-parametric tests were used which do not assume normal distribution. Another key limitation is the experimental design, described in detail in the section below.

3.4.4 Evaluation of experimental design

This experimental design is limited in some areas. Robust experimental design considers Blocking, Replicating and Randomising the samples (Fisher, 1960; Ryan and Morgan, 2007).

Homogeneous peas could have been arranged into a block. There are two ways this could have been done based on existing *a priori* knowledge: first by arranging seed weight and second by the Structure order found by Jing et al. (2010). The blocking arrangement could include guard rows that could be discarded prior to phenotyping (Hammer and Hopper, 1997). The function of the guard row would be to reduce edge-effects in the experiment which can affect variances and introduce biases into the phenotyping experiment (Vanclay, 2006).

Blocking usually occurs in a randomised fashion such as the Randomized Controlled Block Design. Blocking controls variation and the randomisation within the block ensure that all plants get the same treatment (Hammer and Hopper, 1997) by ensuring a reduction in variation of the micro-climate that could affect the plants, such as light, temperature and humidity as well as other factors such as competition from neighbouring plants (Brien et al., 2013). Additionally, whilst this experiment did use three replicates per plant, additional replicates could have been used, additional replication could have been performed over different spaces and replication could have been performed over different time points (Brien et al., 2013). Whilst this experiment grew peas in glasshouse and field, it must be noted that seed weight was the only phenotype to be observed in both environments.

Secondly, the organ harvesting from imaging was the healthiest available in leaves given that the plants were at full maturity and senescence had set in and random pods were selected for imaging. Issues caused by this are that they are not from the same growing point on the plant and may cause some issue in the measurements of this phenotype, however, all plants were phenotyped at full maturity.

In summary, the limitations of this study's design must be considered when analysing downstream results. The variance of the phenotypic data may have been affected by the experimental design. Subsequently, the GWAS analysis performed in Chapter 4 will have limitations, such as, spurious associations whereby associations have been made that do not exist or true associations that do exist are not identified.

3.4.5 Future work

Future work could include additional datasets derived from field, since this study lacked in field data for some phenotypes, this would provide additional comparisons and provide greater insight into phenotypic plasticity for some phenotypes. MktStall could be further developed to include new organs for subsequent version releases - fitting our analogy of a market stall for every trait. Understanding the genetic architecture for the traits evaluated will provide insights into key loci that could be involved in causing a particular morphological trait, this work is embarked upon in the next chapter.

3.4.6 Conclusion

In short, this chapter has explored the development of a novel image analysis tool for the purpose of exploring Simple and Morphological Shape Descriptors in a variety of organs - leaflet, seed and pod. This chapter has combined this dataset with manual measurements of plant height and seed weight. The phenotypic dataset was obtained to explore variation between groups to better understand pea domestication and between conditions to provide insights into phenotypic plasticity. The results show statistically significant differences between domesticated and wild material in plant height, leaflet length, width, perimeter and area as well as in some shape descriptors: leaflet aspect ratio, rectangularity and perimeter ratio of length, pod aspect ratio, pod rectangularity. Furthermore, phenotypic plasticity has been observed between glasshouse and field environments in seed weight. These results help better understand how morphology has evolutionarily changed in pea through the course of domestication as a result of human **artificial selection** and pea's natural adaptive response to changes in environments.

Chapter 4

Results II: Genotyping

4.1 Introduction

Charles Darwin was greatly interested in domestication. He wrote a book on domestication in “The variation of animals and plants under domestication” (Darwin, 1868) and also mentions domestication in “On the origin of the species by means of natural selection” (Darwin, 1859).

Darwin defined evolution but evolution is distinct from domestication. Domestication can be considered to be a type of plant-animal co-evolution since humans have a major influence on the domestication of plants. Humans often artificially select for beneficial traits leading to domestication of elite cultivars. Domestication is a process where species change on an evolutionary level leading to morphological traits where domesticated accessions differ to their wild progenitors (Purugganan and Fuller, 2009).

Domestication does not come without issues. Often, domestication can lead to genetic erosion, which loses a lot of genetic diversity (Wouw et al., 2010). Understanding domestication involves the understanding of genomic signatures of adaptation (Ross-Ibarra, Morrell, and Gaut, 2007).

Chapter 1 explains the current state of pea genomics and explained the challenges behind assembly of pea genome. Here, we introduce the key concepts and the justification for the method used to explore domestication in this chapter and explain how the information produced can to better understand the genetic changes that drive domestication in pea.

4.1.1 Marker Information for Peas

The type of marker obtained from sequencing data provides different information on the genetic diversity. Simple Sequence Repeats (SSR) or microsatellites are markers which make it easy to distinguish how closely related accessions are (Jing et al., 2007). This is because SSRs evolve quickly. Conversely, Single Nucleotide Polymorphisms (SNPs) evolve more slowly and retrotransposon insertion-based markers evolve at a more intermediate pace (Jing et al., 2007).

One study (Jing et al., 2010) used Retrotransposon Based Insertion Polymorphisms to look for genetic diversity in the pea genome with the aim of understanding the diversity between wild and cultivated pea in the John Innes' Germplasm Collection of *Pisum sativum* and performed diversity structure analysis which has been previously outlined in Figures 1.6 and 1.7. They found that there were 3 main groups of pea, which further sub-divided into a total of 14 sub-groups. It is this study, which gives us our existing knowledge of genetic diversity and evolution of the pea and provides us with the sub-division structure, which will be used in this chapter.

Although previous work has been performed to identify SNP markers in pea from alternative core collections (Deulvot et al., 2010), this chapter will generate SNPs albeit in this core collection derived from the John Innes *Pisum* collection with the aim of highlighting genomic regions associated with a trait.

4.1.2 GWAS

The markers obtained from the genotypic data can be associated to a phenotype using a Genome-Wide Association Study. This will help us look at domestication traits to find associated loci. Here, we explore key concepts and explore other well known GWAS studies in pea.

Linkage Disequilibrium (LD) in pea

Studies in calculating linkage disequilibrium (LD) in different pea core collections have been performed. Here, LD in pea is explored. Cheng et al., 2014 showed that in a study of 384 landrace and cultivated peas in the USDA collection, average LD decay across 5-10cM using 203 SNPs was below the critical value (0.17) and found that average r^2 was 0.0169. This shows that SNPs are not very well correlated and are not likely to be observed together which conveys the need for greater number of markers. Jing et al., 2007 showed that in a study of 48 individuals in the JIC *Pisum* germplasm, there is high LD in cultivars and landraces as well as some evidence of LD decay in wild peas. This information suggests that whilst LD averaged at r^2 0.0169 in landrace and cultivars, this is a low value and indicates a need for greater marker density (Cheng et al., 2014), furthermore, wild varieties exhibit LD decay (Jing et al., 2007) suggesting an even greater number of markers should be considered for wild varieties in comparison to domesticated varieties.

Rapid LD decline across a small distance results in greater resolution but more markers will be required (Smýkal et al., 2012). Here, marker density for use in pea GWAS is explored. In pea, LD decay rate has been compared with rice and maize and is said to be similar (approx. 1kb), therefore it is suggested an excess of 100,000 markers should be used in GWAS analysis (Smýkal et al., 2012). However, it must be noted that pea is not closely related to rice or maize. Investigating these claims further, Ching et al. (2002) found no decay in LD in maize using only 114 SNPs, and Huang et al. (2010) found LD decay of 100-200kb using 3,625,000 SNPs. A study on pea by Jing et al. (2007) using 39 markers suggests LD decay in pea is rapid and hence this suggests a larger number of markers should be used for pea GWAS. However, two pea GWAS have been performed in pea both using GBS *de novo*, using 66,591 markers to find 25 statistically associated loci for flower colour (Holdsworth et al., 2017) and 671 markers to find 26 statistically associated markers for flowering onset, 26 statistically associated markers for seed yield and 21 statistically associated markers for adjusted seed yield (Annicchiarico et al., 2017). This evidence suggests that the notion of over 100,000 for a pea GWAS suggested by Smýkal et al. (2012) is idealistic and that a marker density in the thousands (between 671 and 66,571) should be sufficient.

Key Genome-wide Association Studies in pea

Annicchiarico et al. (2017) used the Elshire GBS protocol (Elshire et al., 2011) with ApeKI as the restriction enzyme in 3 RIL populations of pea. This paper is an important study for pea as it explored how marker density changes in the organism when changing parameters. Depending on the level of coverage and amount of missing data, the paper obtained 100 (8x coverage per locus, 10% missing genotype data) to 7500 markers (4x coverage per locus, 50% missing genotype data). This paper decided on using 6x coverage per locus, 20% missing data, 0.01 minor allele frequency (MAF) with inflation factor correction to provide 617 markers for GWAS and this produced up to 26 statistically associated loci. Arguably, the use of RIL populations in the study means that all samples are highly homozygous and not very diverse, which might justify use of such low coverage, however, 617 markers across a 4.5Gb genome is a sparse distribution of SNPs. This study used *Medicago truncatula* as a reference genome. Given that only 15% of their reads aligned to the reference when using Bowtie2 on the "verysensitivelocal" preset, it provides evidence for the argument of *Medicago truncatula* not being an ideal reference for pea.

Holdsworth et al. (2017) also used the Elshire GBS protocol (Elshire et al., 2011) using ApeKI as a restriction enzyme on a USDA core collection, before performing a GWAS on flower colour. Holdsworth et al. (2017) states that in their preliminary analysis, using *Medicago truncatula* as a reference yielded 50% fewer SNPs than *de novo* GBS analysis alone and cites the 25 million year divergence between the two organisms as the reason for this - further substantiating the claim that *Medicago truncatula* is not a good reference for pea. This paper uses Stacks and UNEAK to identify SNPs and took the union of SNPs found by these two tools as the final dataset. For the GWAS analysis, they obtained 66,591 markers, with minor allele frequency 0.01 with 20% missing data, of which 25 associated with flower colour at a corrected Bonferroni threshold of 5%. It explored collinearity of this loci with *Medicago truncatula* by aligning the GWAS associated loci to *Medicago truncatula* using BLASTN and found that it lay in a *M.truncatula* homolog.

Table 4.1 compares the two studies, showing that Annicchiarico et al. (2017) used more stringent parameters for the GWAS in comparison to Holdsworth et al. (2017) and consequently the number of markers were found to be lower. However, Holdsworth et al. (2017) used more sophisticated association tests to account for population structure. Both studies resulted in similar number of statistically associated loci.

TABLE 4.1: **Comparison of Pea Genome-wide Association Studies.** Studies by Annicchiarico et al. (2017) and Holdsworth et al. (2017) were compared for the methods used.

Method	Annicchiarico	Holdsworth
Coverage per Locus	6	Not selected
Missing Data (%)	20	20
Minor Allele Frequency	0.025	0.01
Total SNP number for GWAS	617	66,591
Association Test 1: Single Locus Analysis	GenABEL - Generalized linear model	No
Association Test 2: Population Stratification	Inflation factor	GAPIT - Mixed Linear Model with Kinship Matrix, PCA
Association Test 3: Corrections for Multiple Testing	No	Bonferroni 0.05
Association Test 4: Multi-locus Analysis	No	No
Phenotype (statistically associated loci)	Flowering onset (26), Grain yield (26), Adjusted grain yield (21)	Flower colour (25)

These key studies show that the Elshire protocol (Elshire et al., 2011) provides a suitable number of markers and that *Medicago truncatula* is not an appropriate reference, however, collinearity with this organism can help order and orientate markers in pea.

4.1.3 Justification of method

Based on methodologies used in other GWAS studies previously mentioned, here, a justification of methods is provided for the methodology employed in this chapter.

Assembly

Whole Genome Shotgun (WGS) sequencing is the form of sequencing that this study will use to create this assembly. An alternative form of sequencing would be a targeted form like exome capture. This is however, a difficult, time-consuming and expensive approach and does not sequence within intergenic regions and will require a reference for designing probes to capture the exons (Mamanova et al., 2010). A pilot study within the 1000 Genomes Project (Consortium, 2010) performed exome-targeted sequencing on 697 samples at almost 56x coverage. In the 1000 Genomes Project, this method found rare SNP calls which may not be suited to a GWAS study and is an expensive method.

Therefore, this chapter intends to make a reference genome using Whole Genome Shotgun Sequencing on a cultivated variety, Gradus No.2 (JI 1153). This accession was chosen due to its presence in the USDA collection (PI 210639) and due to its key phenotypes: Tall, Early Flowering, Long Pods, Large Seed Weight, mixture of Resistance and Susceptibility in Fusarium Wilt Race 1 (*SeedStor JI1153*; *US National Plant Germplasm System PI 210639*).

The resulting reads from the WGS sequencing will then be assembled using assembly tools. These tools are either De Bruijn Graph Assemblers or Overlap-Layout-Consensus Assemblers, which have previously been mentioned in section 1.5.6. A key consideration when assembling pea is the size of the genome (4.45Gb) and repetitiveness. Conventional assemblers have not been designed for assembly of such genomes. However, a recent de Bruijn Graph Assembly tool called w2rap (Wheat/Whole-genome Robust Assembly Pipeline) (Clavijo et al., 2017) has been used for this purpose in wheat, *Triticum aestivum*, a 17Gb hexaploid organism, and therefore a larger and more complex genome. This tool has also been tested in *Homo sapiens* and *Arabidopsis thaliana* and results compared against assemblies produced by other de Bruijn Graph Assemblers (Clavijo et al., 2017). The results obtained by Clavijo et al. (2017) showed w2rap obtained 6x more contiguity in the *Arabidopsis thaliana* assembly compared to Abyss v2.0 (Jackman et al., 2017) and 16x more contiguity compared to SOAPdenovo2 (Luo et al., 2012). In human, Clavijo et al. (2017) obtained 3x the contiguity with w2rap compared to Abyss v2.0 (Jackman et al., 2017) and 30x the contiguity compared to SOAPdenovo2 (Luo et al., 2012). Thus, proving that it works well on large, repetitive such as wheat but also it works on diploid organisms such as *Arabidopsis thaliana*. Therefore, this tool is appropriate for a large, repetitive diploid organism such as pea.

Genotyping by Sequencing

As discussed in detail in Chapter 1 and previously above, here the reasons for choosing Genotyping by Sequencing (GBS) are briefly re-iterated. GBS is a sequencing method to produce a high-density marker panel of SNPs at low cost (Davey et al., 2011). This method can be performed *de novo* or with the use of a reference. GBS is reduced representation sequencing method meaning it uses a restriction enzyme to cut the DNA and sequences around the cut site (Davey et al., 2011). Therefore, not all of the genome is sequenced. The choice of restriction enzyme is an option that can be exploited for number of cuts based on cutter length (Andrews et al., 2016) and for avoiding repetitive regions based on using a methylation-sensitive cutter (Davey et al., 2011). This type of sequencing possesses an easier library preparation stage compared to other reduced-representation sequencing methods such as RAD-seq (Davey et al., 2011). The two pea GWAS discussed above both used GBS (Holdsworth et al., 2017; Annicchiarico et al., 2017) using a methylation-sensitive restriction enzyme, ApeKI, and both followed the Elshire protocol (Elshire et al., 2011), therefore, this method has been shown to work in pea and to produce a sufficient number of markers to perform GWAS in pea.

Types of GBS assembly tools and their respective pros and cons have been discussed in Chapter 1. Annicchiarico et al. (2017) used UNEAK and Holdsworth et al. (2017) used both UNEAK and Stacks, indicating that these two tools have been shown to work well with pea GBS data. However, these GBS assembly tools hold their own advantages and disadvantages. Whilst Stacks is capable of using a reference, large numbers of samples can introduce errors into the "catalog" of loci built by the tool (*Stacks denovo map*). UNEAK is known for its capacity to work with high and low coverage of sequencing data, however, it trims the sequencing reads to 64bp (*TASSEL 3.0 Universal Network Enabled Analysis Kit (UNEAK) pipeline documentation*), essentially discarding a third of the data before assembly. Hence, to avoid any biases from the assembly tools, this study has chosen not to assemble using GBS assembly tools, but instead, choses to directly align the GBS reads to the assembly.

Alignment, Variant calling and Imputation

Alignment of reads to the genome assembly can be performed using alignment tools such as BWA (Li and Durbin, 2009) or Bowtie2 (Langmead and Salzberg, 2012). Of these Bowtie2 is faster and more sensitive than BWA (Langmead and Salzberg, 2012). Therefore, given this information, Bowtie2 is an appropriate alignment tool for aligning GBS reads of each pea sample to the pea assembly.

There are different types of alignment: global and local alignment. Global alignment means that the total length of the read aligns to the reference, whereas local alignment, using tools such as BLAST (Altschul et al., 1990), means that part of the read aligns to the reference (Kagale et al., 2016). Local alignment removes unaligned bases from the ends of the reads to increase the alignment score but this can result in multi-mapping of reads, in other words, reads that align to several places in the reference (Kagale et al., 2016). Using global rather than local alignment means that the entirety of the read aligns and this is a more accurate albeit slower way to align (Kagale et al., 2016). The resulting Binary Alignment Map (BAM) file will help to call variants using a variant calling tool (Kagale et al., 2016) such as VarScan (Koboldt et al., 2012), FreeBayes (Garrison and Marth, 2012) or GATK (McKenna et al., 2010). VarScan scores and sorts alignments removing unmapped or ambiguously mapped reads and therefore leaves uniquely mapped reads to detect variants such as SNPs or indels (Koboldt et al., 2012). FreeBayes is a haplotype based variant detector that uses a Bayesian statistical method to allow for the detection of multi-allelic loci and identify SNPs or indels (Garrison and Marth, 2012).

The SNPs obtained from these tools will be filtered to reduce false positives. Missing genotype calls and Minor Allele Frequency (the frequency of the least common allele) are two common methods to filter SNPs (Kagale et al., 2016). Minor Allele Frequency detects rare variants, which can be discarded since GWAS is suited to looking for common variants. Most GWAS recommend a MAF cut-off of 0.05, this is because the lower the MAF the more likely it is to be an error in the variant call (Coleman et al., 2015). This is especially true of smaller sample sizes (< 10,000). However, both Holdsworth et al. (2017) and Annicchiarico et al. (2017) papers are proof of concept studies in pea, using MAFs of 0.01 and 0.025 respectively, to show that GWAS can work with such low MAFs. Missing genotype calls can affect population diversity statistics (Arnold et al., 2013). Missing data is a metric that provides an overview of how complete the dataset is and if the quality of the genotype calls is affected since missing data tends to be non-random (Laurie et al., 2010). Missing data can be removed prior to downstream analysis or predicted through imputation.

Some genome-wide association studies make the decision to impute any missing genotype data using imputation tools such as: Impute (Marchini and Howie, 2010), MACH (Li et al., 2010) and Beagle (Browning and Browning, 2009). Imputation predicts genotypes *in silico* when these genotypes have not been detected though variant calling by using a statistical model called the **hidden Markov Model** (HMM) (Marchini and Howie, 2010), a probabilistic method that assumes that the observations are hidden from the observer, follow the Markov design that the current state is independent of the previous states and that the value is discrete (Ghahramani, 2001). This can improve the statistical power of a GWAS study by calling SNPs that were previously missing and reduce the likelihood of false positives by correcting incorrect variant calls (Marchini and Howie, 2010). A review on missing genotype imputation found that Beagle had the quickest run-time of all tools reviewed with the least amount of maximum memory allocation, although it provides less accurate data (Ellinghaus et al., 2009). Impute is more accurate than Beagle, and faster than MACH, so could be the best compromise (Ellinghaus et al., 2009). However, important considerations must be made, such as, will imputing the genotype actually produce correct genotypes? Imputation requires a high-density SNP reference panel of haplotypes and its accuracy is improved when MAF decreases and trio information of mother-father-child is used (Marchini and Howie, 2010). Since, the pea reference that this thesis produces will be in draft format, and a high-density marker panel of SNPs cannot be guaranteed nor trio information obtained, the decision not to impute has been made due to the potential to introduce incorrect genotypes.

Population Genetics Analysis

Structure (Pritchard, Stephens, and Donnelly, 2000) can be used to identify **population stratification** from genotypic data. It identifies subpopulations of the entire population based on the allele frequency at each locus using a Bayesian clustering approach. In GWAS this can be useful to determine if an association is a "true" association caused by a genetic locus or a spurious one caused by the population structure (Cardon and Palmer, 2003).

In this study, the use of **phylogenetics** aims to understand pea's evolutionary history and give an understanding of domestication over time. Phylogenetic trees can be drawn from the analysis of DNA sequences and a variety of tools can do this such as the "phylo" module of VCF-kit (Cook and Andersen, 2017) or Phylip (Felsenstein, 1989). **Principle Components Analysis (PCA)** visualises variation in a simplistic manner. It can be used to correct for population stratification causing incorrect associations between markers and genes (Price et al., 2006) through the identification of samples which should be removed.

The combination of population stratification, phylogenetics and Principle Components Analysis can identify individuals which can be removed from the dataset to reduce confounding biases.

Genome-wide Association Study

As previously discussed in Chapter 1, there are several association tests available for Genome-wide Association Studies. This section outlines which tools and tests were chosen for this study with the aim of preventing spurious associations. The first consideration made was the use of Generalised or Mixed Linear Models (MLM). This study has chosen Mixed Linear Models due to its ability to discern confounding factors that could cause spurious associations by taking in population structure and relatedness into account. GAPIT is a GWAS tool that works using a compressed Mixed Linear Model, a computationally efficient MLM method (Lipka et al., 2012) that uses kinship matrix to account for **cryptic relatedness** - individuals that are related that was not known by at the start of the study (Voight and Pritchard, 2005). This study has opted for MLM, over tools based on Generalised Linear Model such as GenAbel so as to avoid spurious associations (Aulchenko et al., 2007).

For correction of multiple testing, a Bonferroni correction threshold should be used as performed previously in another pea GWAS (Holdsworth et al., 2017), where the Bonferonni correction threshold (α) is set to 0.05 divided by the total number of tests (the total number of SNPs).

$$\text{Bonferonni correction threshold} = \frac{0.05}{\text{Number of SNPs}} \tag{4.1}$$

Whilst the Bonferroni correction is a more conservative approach because the p-value is adjusted to a lower value, those which do pass this stringent correction are unlikely to be false positives (Type I errors). The Bonferroni correction can provide some confidence behind SNPs found to be statistically associated (Johnson et al., 2010). Whilst the False Discovery Rates method corrects for false positives and provides an indication of how many true associations are present in the data (Bush and Moore, 2012), it is not as conservative as the Bonferroni correction which attempts to remove all false positives. Therefore, in this study, Bonferonni correction will be performed on the GWAS dataset.

Visualisation of GWAS data provides a way to easily observe statistical association between the genotype and phenotype. Manhattan plots are a plot where chromosomes are plotted against the negative logarithm of p-values for the association between the SNP and phenotype. The observation of large peaks above a statistically significant threshold indicate the presence of a statistically significant genotype-phenotype association. These may be either a causal variant, a tag marker in linkage disequilibrium with the causal variant or a confounding SNP.

One way to detect whether these SNPs are confounding is through the use of visualising the data with Quantile-Quantile (Q-Q) plots. Here, the negative logarithms of p-values of the SNP association are plotted against their expected value, where the null hypothesis is defined as there is no association between the genotype and the phenotype (Lipka et al., 2012; Turner, 2014). Therefore, most unassociated genotype-phenotype data lie along the reference line but the associated data points deviate from the reference line on the right-hand side of the plot (Turner, 2014). Any deviations not on the right-hand side of the plot are likely to be spurious associations from confounding factors such as population stratification (Turner, 2014).

4.1.4 Aims

The aim of this project is to identify genomic loci or regions of interest that are statistically associated with domesticated traits in order to help inform breeding practices and develop improved cultivars. From the largely representative core collection of wild, landrace and cultivated lines from the JIC *Pisum* germplasm that we have obtained in Chapter 2 and the phenotypes obtained in Chapter 3, this chapter will genotype each accession using **Genotyping by Sequencing (GBS)**, create the first publicly available pea genome reference and use these to find genotype-to-phenotype associations through a GWAS in order to better inform breeders of breeding lines based on their phenotype of interest (Figure 4.1).

[Redacted]

FIGURE 4.1: **Overview of Project Strategies.** Figure redacted.

This project aims to study the phenotypes and traits in order to provide some insight into key markers - single nucleotide polymorphisms (SNPs) - associated with agriculturally beneficial traits, and potentially identify the genomic loci which may be directly or indirectly associated with this trait.

The phenotypes investigated in this study include: plant height, seed weight, leaflet teeth as well as SMSDs in leaflet, seed and pod. The arguments for the relevance of these phenotypes in agriculture are made in Chapter 3.

4.2 Materials and Methods

4.2.1 DNA extraction for GBS

Leaf material was harvested (22-24th June 2015) from the peas grown in the glasshouse (sown: April 2015). In most cases, leaf material was harvested but in the cases of *Afila* (leafless mutants) the stipules were taken. For each accession, two leaves were taken (to serve as duplicates) and placed into 96 well plates.

The leaf tissue samples were freeze-dried over a 48 hour period. Ball bearings of size 4mm were added to each sample and 500μl of Extraction buffer (0.1M Tris-HCl pH 7.5, 0.05 EDTA pH 8.0, 1.25% SDS, warmed to 65°C) was added. The plant material was ground down and shaken using the GenoGrinder Spex Sample Prep 2010 (*Geno/Grinder® - Automated Tissue Homogenizer and Cell Lyser*) for two minutes at 1750 rpm. The samples were briefly centrifuged to bring material to the bottom of the plates and then incubated for an hour at 65°C. 250μl of cold 6M Ammonium Acetate (4°C) was added to each sample and held on ice for 20 minutes. The liquid lysate was transferred into new plate using a robot and this lysate was centrifuged at max speed for 30 mins to produce a pellet of DNA. This was then stored in the fridge. The DNA was extracted using beads from the GenFind Blood Kit (*Agencourt GenFind V2*), 300μl of beads were added to the supernatant, shaken at 1000 rpm for a minute then incubated at room temperature for 5 minutes. This was placed on the magnet for 12 minutes. The supernatant was removed and the beads washed in 500μl of GenFind Wash 1 and incubated on the magnet. This step was repeated twice. The supernatant was removed and the beads washed in 250μl of GenFind Wash 2 and incubated on the magnet. This step was repeated twice. The plates were stored overnight at 4°C and then the DNA was eluted in EB.

4.2.2 Sequencing

Whole Genome Shotgun Sequencing

The accession, Gradus No 2 (John Innes Collection Number: JI 1153; USDA collection number: PI 210639) was selected as the chosen cultivar for Whole Genome Shotgun sequencing based on its key phenotypes: Tall, Early Flowering, Long Pods, Large Seed Weight and mixture of Resistance and Susceptibility in Fusarium Wilt Race 1 (*SeedStor JI1153*; *US National Plant Germplasm System PI 210639*) as well as its presence in the John Innes and USDA collections. DNA was extracted by a service provider (Earlham Institute) and PCR-free 250bp Paired End libraries sequenced on 4 lanes of Illumina HiSeq2500 using Rapid Run mode and v2 chemistry (*HiSeq SBS Kit V2*).

GBS Sequencing of the core collection

After DNA extraction, Genotyping-by-Sequencing was performed by a service provider (Cornell University) to sequence 350 samples from the core collection using ApeKI as a restriction enzyme to create 100bp Single End 96-plex GBS libraries following the Elshire protocol (Elshire et al., 2011). These were sequenced across six lanes of Illumina HiSeq2500 on High Output Mode with v4 chemistry (*HiSeq SBS Kit V4*).

4.2.3 Bioinformatic Analysis

De novo assembly

Whole Genome Shotgun reads were assessed for quality using FASTQC v0.11.4 (*FastQC: a quality control tool for high throughput sequence data*). The k-mer coverage was estimated using KAT v2.3.4 (Mapleson et al., 2016) to prevent k-mers from erroneous reads being included in the assembly. Whole Genome Shotgun reads were assembled *de novo* using w2rap version a43f5a0 (Clavijo et al., 2017). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession PUCA000000000. The version described in this paper is version PUCA010000000. The contiguity of the assembly was assessed using Abyss-fac v2.0.2 (Simpson et al., 2009). Completeness of the assembly was assessed using BUSCO v3.0 (Simão et al., 2015) in genome mode against the plant database "embryophyta_odb9".

Pseudo-chromosomes

Pea contigs were ordered with respect to *Medicago truncatula* to create pseudo-chromosomes. The raw reference prior to NCBI upload and aligned to the *Medicago truncatula* reference genome v4.0 (Tang et al., 2014) downloaded from J.Craig Venter Institute FTP server using blastn from BLAST v2.6.0 (Altschul et al., 1990). The top contig hit was taken using a custom Perl script written by Laura-Jayne Gardiner (LJG) and ordered by *M.truncatula* chromosome number and start point in Unix. Another custom Perl script, written by LJG and adapted by the author, was used to order the pea pseudo-chromosomes with respect to *M.truncatula* creating 8 pseudo-chromosomes and then places remaining contigs into pseudo-chromosomes at the end (remaining 13 pseudochromosomes).

GBS pre-processing

The GBS reads were quality checked by FastQC v0.11.4 analysis (*FastQC: a quality control tool for high throughput sequence data*). The reads were demultiplexed using GBSX v1.3 (Herten et al., 2015) allowing for one mismatch in the barcode. For each sample, all 6 lanes were concatenated and blank wells removed and samples with a file size <200M removed. This left 319 out of 350 samples after losses included in DNA extraction stages.

Alignment of GBS reads of the core collection to the assembly

GBS reads for the core collection were aligned to the pea pseudo-chromosomes using Bowtie v2.2.9 (Langmead and Salzberg, 2012) and were filtered to remove non-aligned reads using awk (Aho, Kernighan, and Weinberger, 1979). Alignments with a mapping quality (probability of the read being misplaced) of less than 10 were filtered using samtools v0.1.18 (Li et al., 2009) using parameter -q 10. SNPs were called ensuring there was a minimum coverage of 5x for each SNP called using VarScan v2.3.7 (Koboldt et al., 2012) with parameter –min-coverage 5 to provide a Variant Call Format (VCF) file of all SNPs.

Alignment post-processing

Changes in number of sites across minor allele frequency (0 to 0.05 in increments of 0.01) was performed using vcftools v0.1.13 (Danecek et al., 2011). Missing data for each sample and depth for each site was calculated for MAF 0.01 and MAF 0.05 using vcftools v0.1.13 and samples removed where depth was less than or equal to 20x. For MAF of 0.01, sites were excluded with a proportion of missing data from 0.1 to 0.9 in increments of 0.1 using vcftools v0.1.13. For proportion of missing data 0.2 and MAF 0.01, VCF was converted to HapMap format using TASSEL v5.2.41 (Bradbury et al., 2007).

Structure

The filtered VCF file (missing data 0.2 and MAF 0.01) was converted for structure input using PGDSpider v2.1.1.3 (Lischer and Excoffier, 2011). Structure v2.3.4 (Pritchard, Stephens, and Donnelly, 2000) was used with default parameters except: BURNIN 10000, NUMREPS 10000, NUMINDS 297, NUMLOCI 13058. The number of populations (K) iterated from 3 to 21 and was replicated for 10 runs of each K. The average of these 10 runs were taken and difference between the average of K and the average of the previous K calculated (Evanno, Regnaut, and Goudet, 2005). The highest K selected was 14 and highest run selected (run 2) for all subsequent analysis.

Phylogenetics

A Neighbour-Joining phylogenetic tree was drawn from the VCF using vcfkit v0.1.6 (Cook and Andersen, 2017) with dependency MUSCLE v3.8.31 (Edgar, 2004) and visualised in Figtree v1.4.3 (Rambaut, 2007).

Discriminant Analysis of Principal Components (DAPC)

Discriminant Analysis of Principal Components (DAPC) was performed on the VCF file in R version 3.4.3 (R Core Team, 2013) with the libraries: vcfR (Knaus and Grünwald, 2017) and adegenet (Jombart, 2008). The analysis was performed by retaining 250 Principal Components and using 2 Discriminant functions.

Genome-Wide Association Study

The HapMap file and the phenotypes obtained in Chapter 3 were used to perform the Genome-wide Association Analysis. The analysis was performed with compressed mixed linear model (Zhang et al., 2010) implemented in the GAPIT R package (Lipka et al., 2012) with pre-requisites installed: Bioclite, scatterplot3d, gplots, genetics, multtest, gplots, LDheatmap, genetics, ape, EMMREML and compiler using R version 3.4.3 (R Core Team, 2013). Default parameters were used with the exception of the "cutOff" parameter set to 0.05. Manhattan plots were then visualised using the "qqman" package (Turner, 2014).

For each statistically associated SNP, genomic co-ordinates of the closest marker on the left and the right-hand side was obtained to form a flanking region either side of this SNP. The flanking region was aligned to *Medicago truncatula* using BLASTN 2.8.1+ (Altschul et al., 1990) in order to find the genomic co-ordinates of the region in *Medicago truncatula*. The co-ordinates were searched in MedicMine (Krishnakumar et al., 2014) to find the gene. The gene from *Medicago truncatula* was aligned against a model organism *Arabidopsis thaliana* to identify orthologs and their function in *Arabidopsis thaliana*. The alignment was performed using BLASTX 2.2.8 (*The Arabidopsis Information Resource (TAIR)*) on default settings to align the *Medicago truncatula* gene to The Arabidopsis Information Resource (TAIR) protein database.

4.3 Results

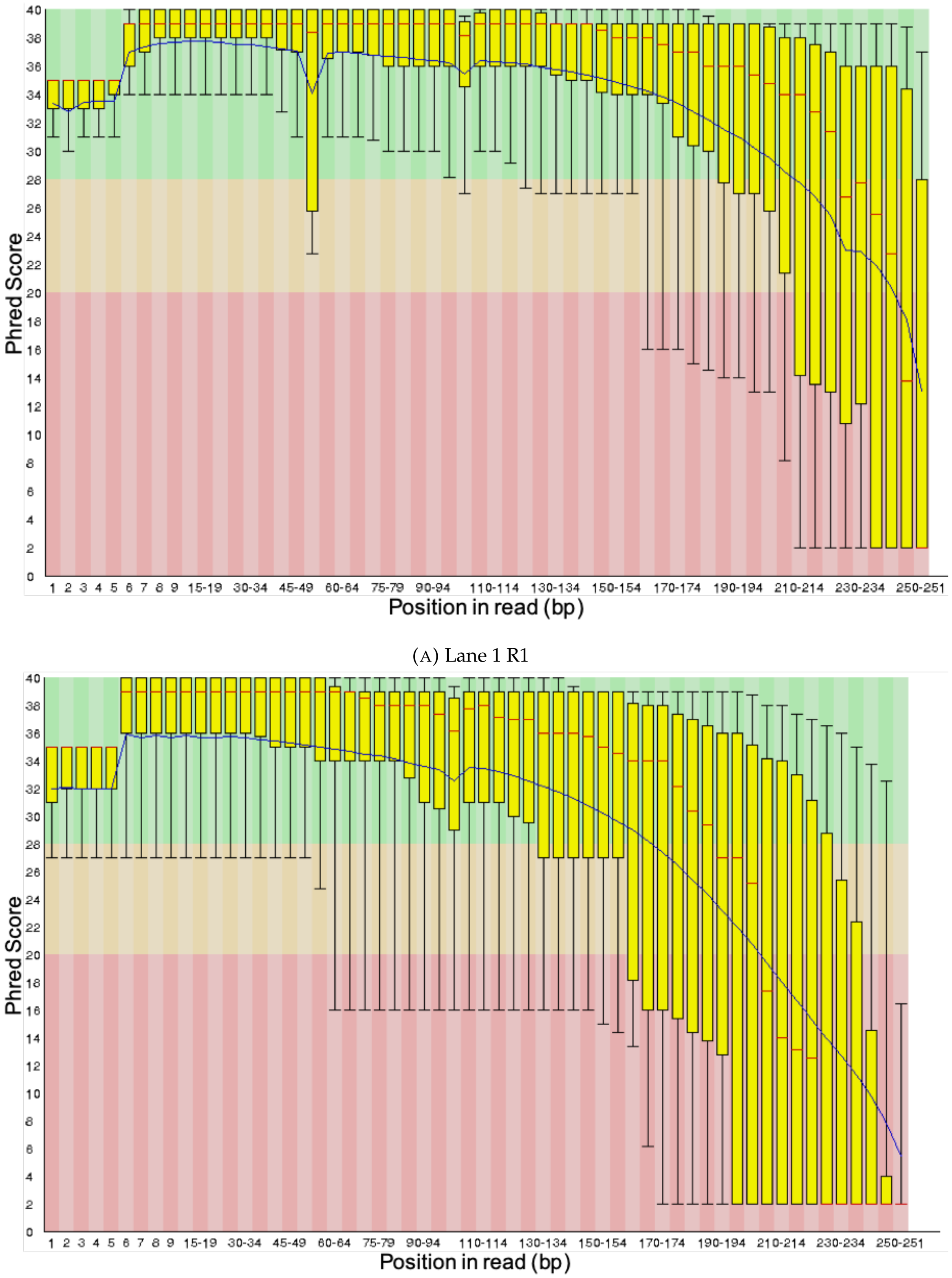
4.3.1 Assembly

The 250bp Paired-End PCR-free sequencing was run on 4 lanes providing 1,539,718,400 sequences in total. The output of reads across 4 lanes of sequencing is shown in Table 4.2. The results show the sequencing data had a GC content of between 36-37%.

TABLE 4.2: **Whole Genome Sequencing Reads.** Total number of sequences, sequence length and GC content are compared for each lane of Whole Genome Sequencing reads.

Lane	Read	Total Sequences	Sequence Length	%GC
Lane 1	Read 1	194,604,528	251	36
	Read 2	194,604,528	251	37
Lane 2	Read 1	194,551,620	251	36
	Read 2	194,551,620	251	37
Lane 3	Read 1	190,625,724	251	36
	Read 2	190,625,724	251	37
Lane 4	Read 1	190,077,328	251	36
	Read 2	190,077,328	251	37

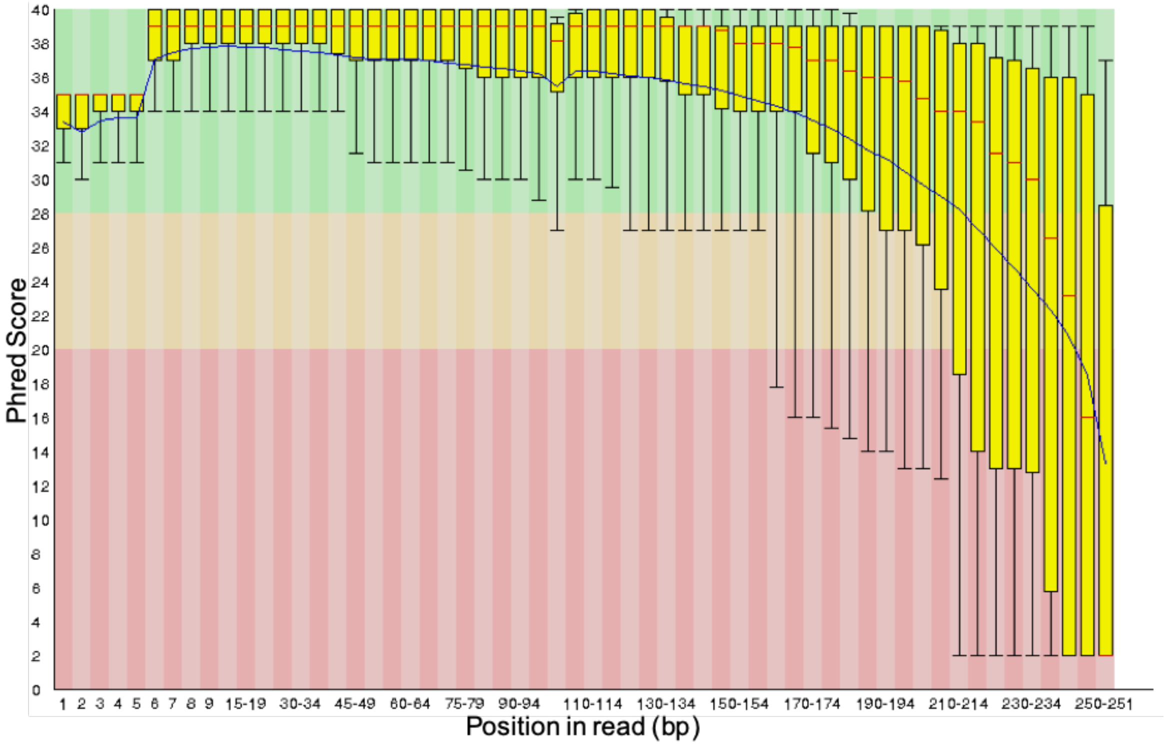
A FASTQC analysis of the per base sequence quality of the FASTQ reads is shown in Figures 4.2, 4.3, 4.4 and 4.5. The Phred Score of these reads shows that the error probability of these reads has 99.9-99.99% accuracy (Phred Score 30-40) up to 190bp before dropping to as low as 37% (Phred Score 2) for Read 1 and for Read 2 has 99.9-99.99% accuracy (Phred Score 30-40) up to 130bp before dropping to as low as 37% (Phred Score 2).



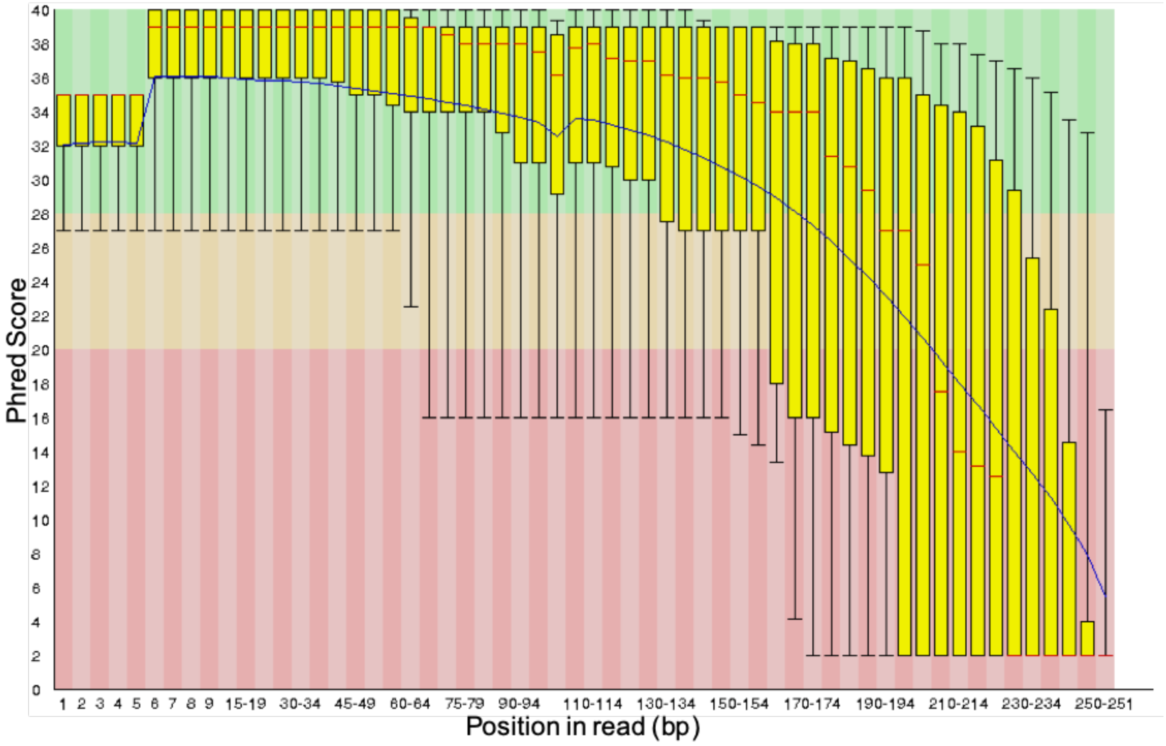
(A) Lane 1 R1

(B) Lane 1 R2

FIGURE 4.2: **Per Base Sequence FASTQC analysis of lane 1 of Whole Genome Sequencing (WGS).** On the x-axis is position in base pair, on the y-axis the Phred Score and graph background illustrates good quality (green), medium quality (orange) and poor quality (red) base calls. Original in colour.

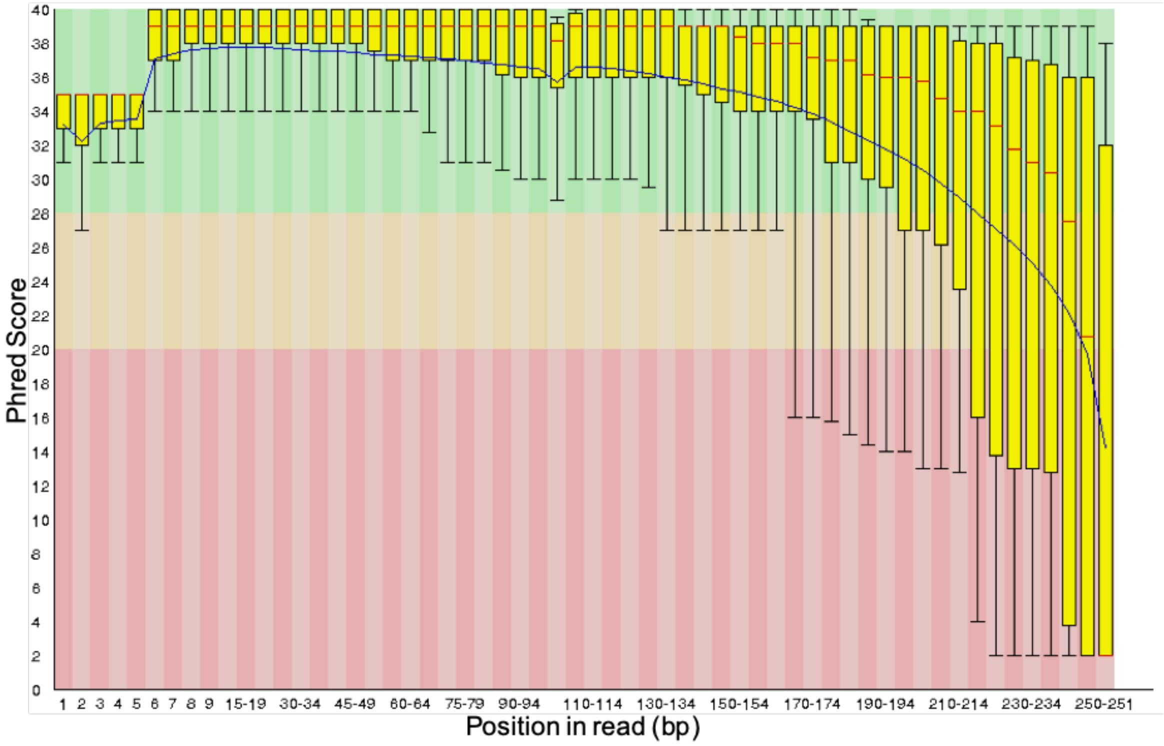


(A) Lane 2 R1

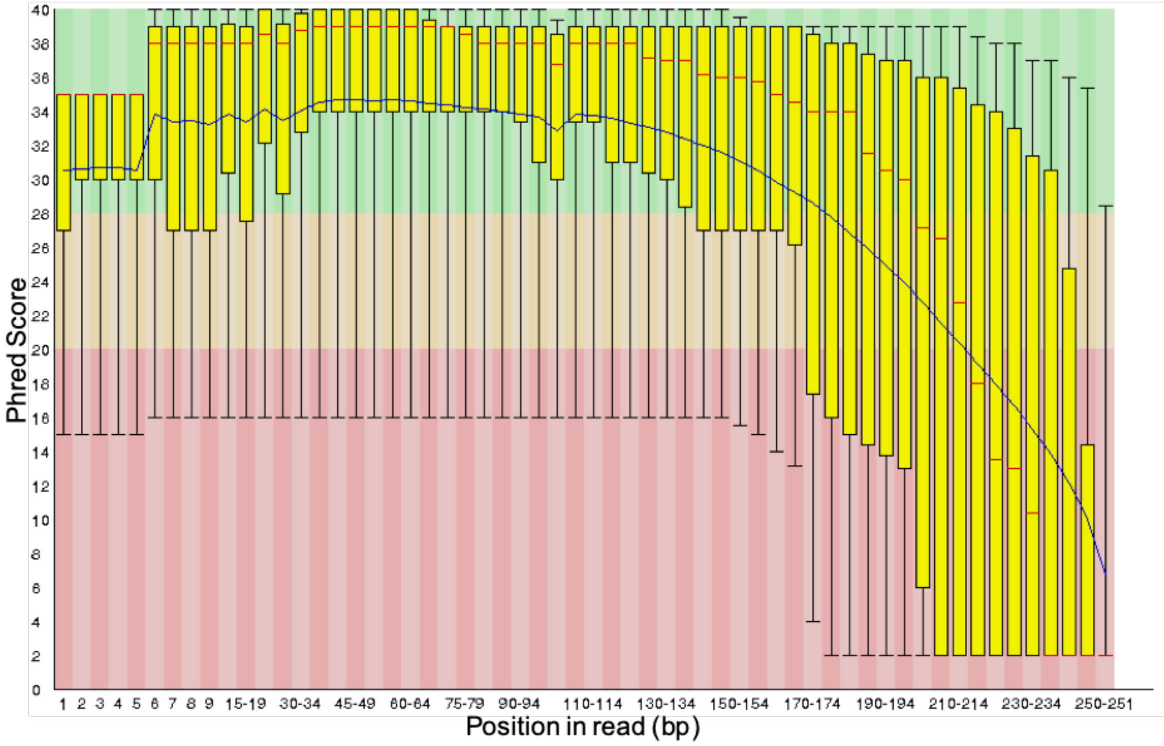


(B) Lane 2 R2

FIGURE 4.3: **Per Base Sequence FASTQC analysis of Lane 2 of Whole Genome Sequencing (WGS).** On the x-axis is position in base pair, on the y-axis the Phred Score and graph background illustrates good quality (green), medium quality (orange) and poor quality (red) base calls. Original in colour.

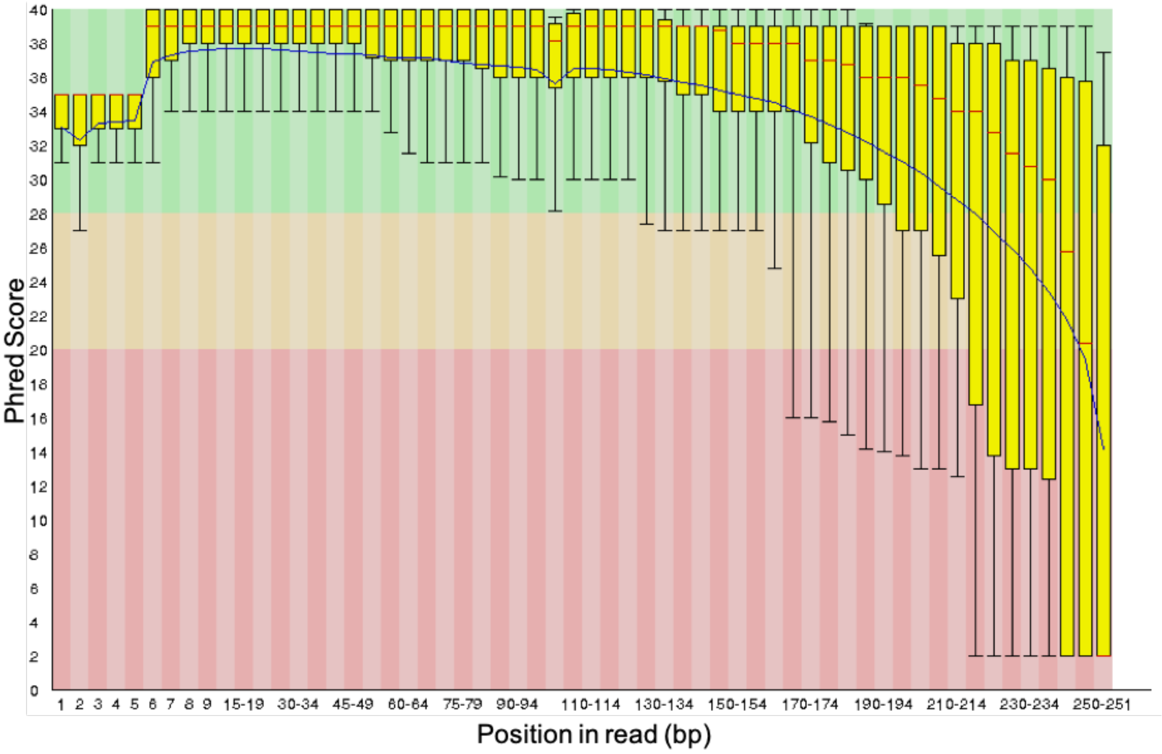


(A) Lane 3 R1

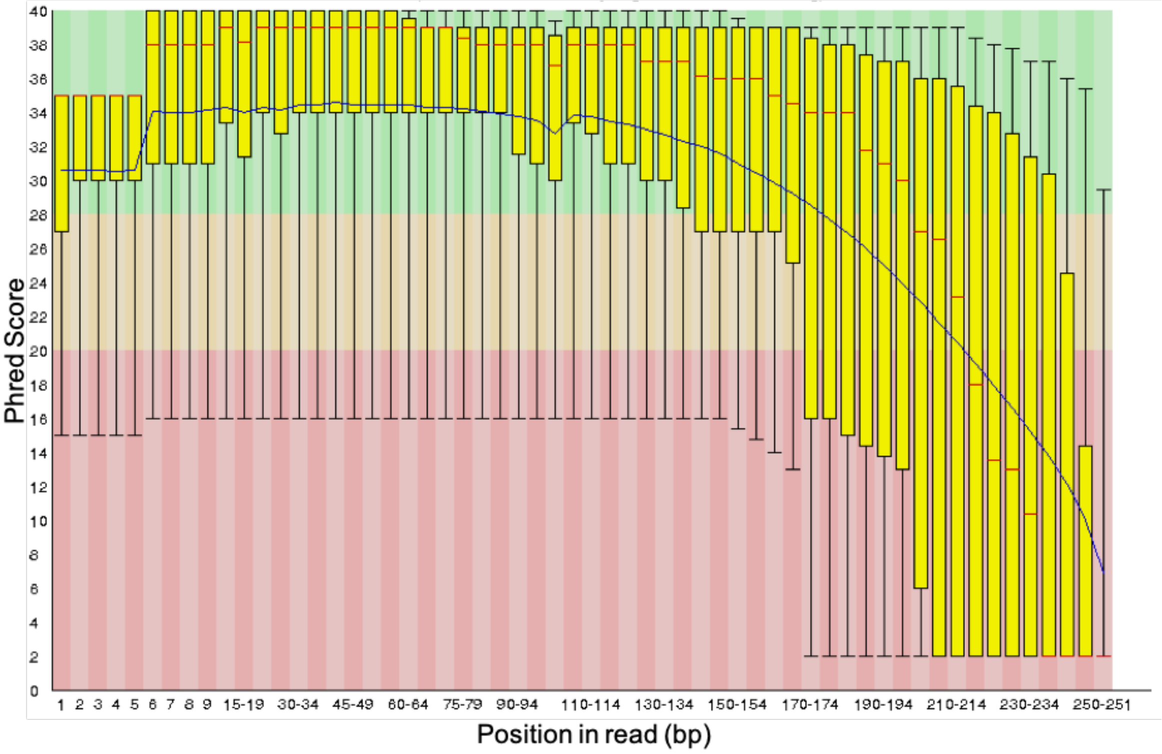


(B) Lane 3 R2

FIGURE 4.4: **Per Base Sequence FASTQC analysis of Lane 3 of Whole Genome Sequencing (WGS).** On the x-axis is position in base pair, on the y-axis the Phred Score and graph background illustrates good quality (green), medium quality (orange) and poor quality (red) base calls. Original in colour.



(A) Lane 4 R1



(B) Lane 4 R2

FIGURE 4.5: **Per Base Sequence FASTQC analysis of Lane 4 of Whole Genome Sequencing (WGS).** On the x-axis is position in base pair, on the y-axis the Phred Score and graph background illustrates good quality (green), medium quality (orange) and poor quality (red) base calls. Original in colour.

Table 4.3 shows the contiguity metrics of the *de novo* assembly. The assembly holds an N50 of 12,597. There are 58,993 sequences greater than or equal to this N50. The smallest contig is of length 500bp, of which there were 938,163 sequences this length and the largest contig is 217,288bp in length.

TABLE 4.3: **Assembly Metrics.** Where n is total number of sequences, n:500 is number of sequences at least 500bp in length, L50 is number of sequences that are N50 or greater in length, min is the minimum size of contig, N80 is , N50 is, N20 is , E-size is sum of square of sequence length divided by assembly length (*ABySS - Stats*).

Abyss Statistic	Value
n	5,450,201
n:500	938,163
L50	58,993
min	500
N80	2,063
N50	12,597
N20	33,775
E-size	19,431
max	217,288

The Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis was performed on the assembly. Table 4.4 shows the 87.2% of single copy orthologs are complete, with few fragmented or missing (3.8% and 9% respectively). The total number of BUSCOs found were 1440.

TABLE 4.4: **Benchmarking Universal Single-Copy Orthologs (BUSCO) found in the assembly.**

	BUSCOs found	Percentage
Complete BUSCOs (C)	1256	87.2
Complete and single-copy BUSCOs (S)	1178	81.8
Complete and duplicated BUSCOs (D)	78	5.4
Fragmented BUSCOs (F)	54	3.8
Missing BUSCOs (M)	130	9
Total BUSCO groups searched	1440	

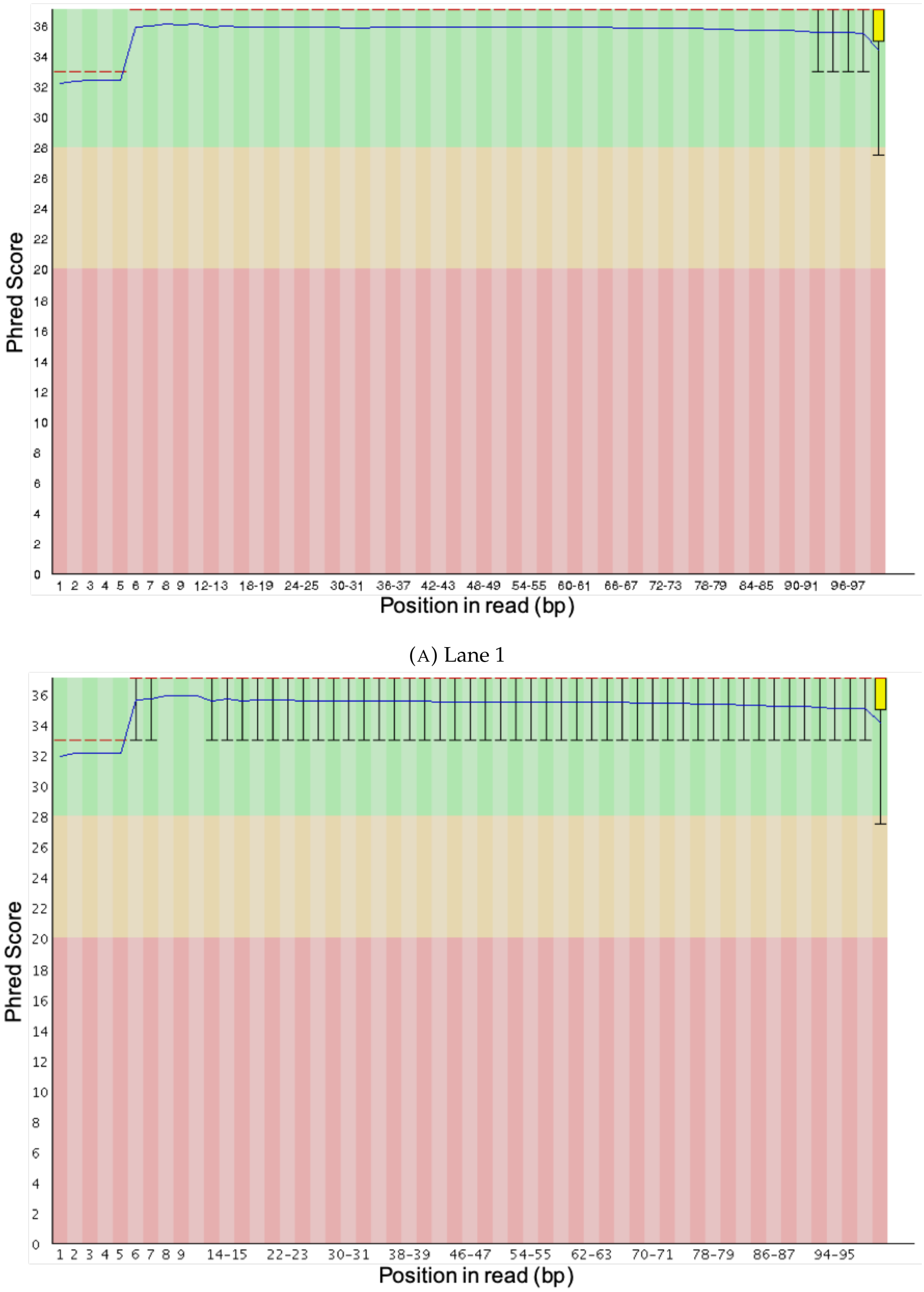
4.3.2 GBS pre-processing

The GBS data was sequenced and demultiplexed. Table 4.5 shows a total of 233,492,104 reads were sequenced across 6 lanes and percentage of sequence discarded from the demultiplexing stage ranged from 2.32 to 3.23%.

TABLE 4.5: **Total number of sequences, sequence length and GC content for all 6 lanes of Genotyping by Sequencing (GBS) data and number of sequences unassigned, number of sequencing remaining after demultiplexing and percentage of sequences lost.**

GBS lane	Sequences	Length (bp)	%GC	Unassigned sequences	Sequences After Demultiplexing	% Lost
1	233,492,104	101	47	5,409,413	228,082,691	2.32
2	296,101,094	101	47	7,003,136	289,097,958	2.37
3	279,499,705	101	48	7,873,790	271,625,915	2.82
4	291,544,297	101	47	8,661,959	282,882,338	2.97
5	178,187,326	101	47	5,750,731	172,436,595	3.23
6	315,320,668	101	48	9,497,845	305,822,823	3.01

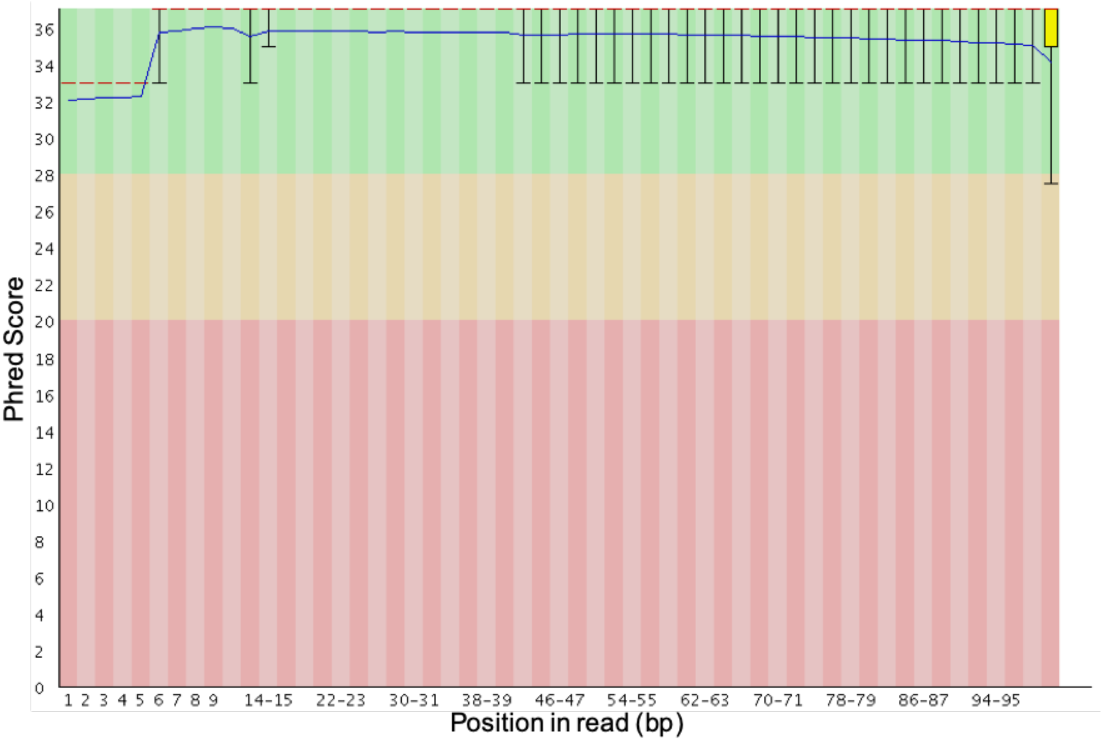
A FASTQC analysis of the per base sequence quality of the demultiplexed FASTQ reads of the GBS dataset. These reads had a GC content of 47-48%. Figures 4.6, 4.7 and 4.8 shows the Phred Sequencing Quality score for each lane of sequencing, each of these lanes show the Illumina reduction in the 3′ end, however the majority of the read lengths remain around a Phred Score of 30, suggesting trimming is not needed. However, lane 6 appears to have more variability across read length in comparison to the other lanes of sequencing.



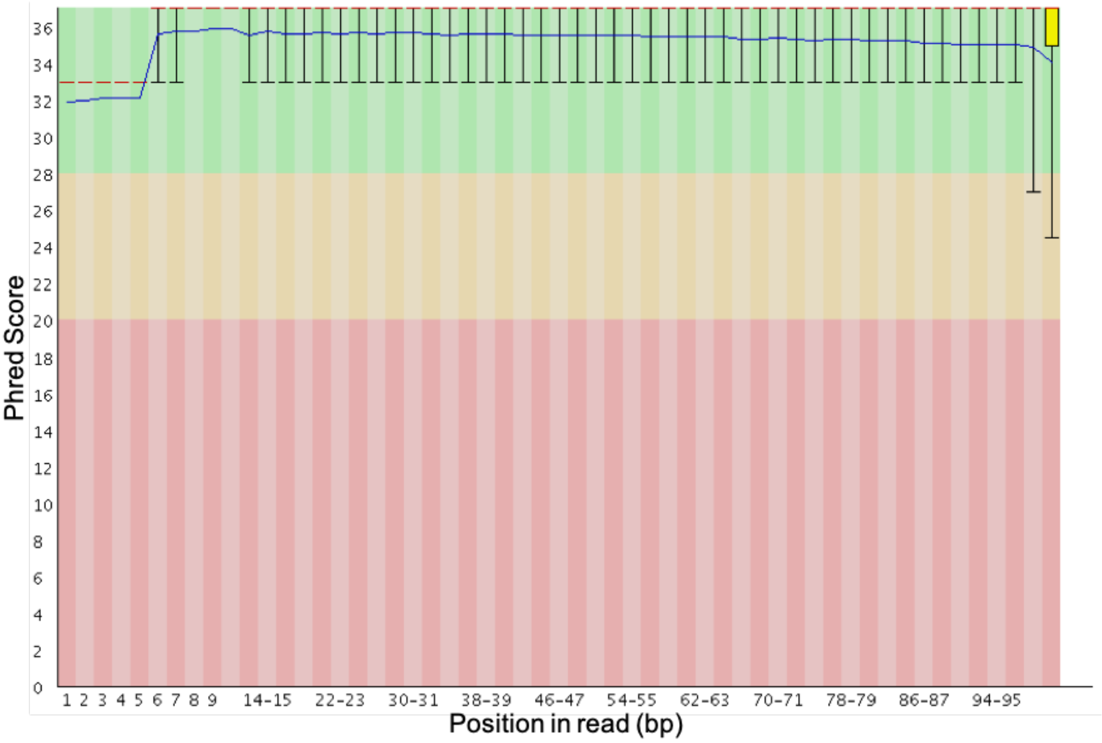
(A) Lane 1

(B) Lane 2

FIGURE 4.6: **Per Base Sequence FASTQC analysis of Lane 1 and 2 of Genotyping By Sequencing (GBS).** On the x-axis is position in base pair, on the y-axis the Phred Score and graph background illustrates good quality (green), medium quality (orange) and poor quality (red) base calls. Original in colour.

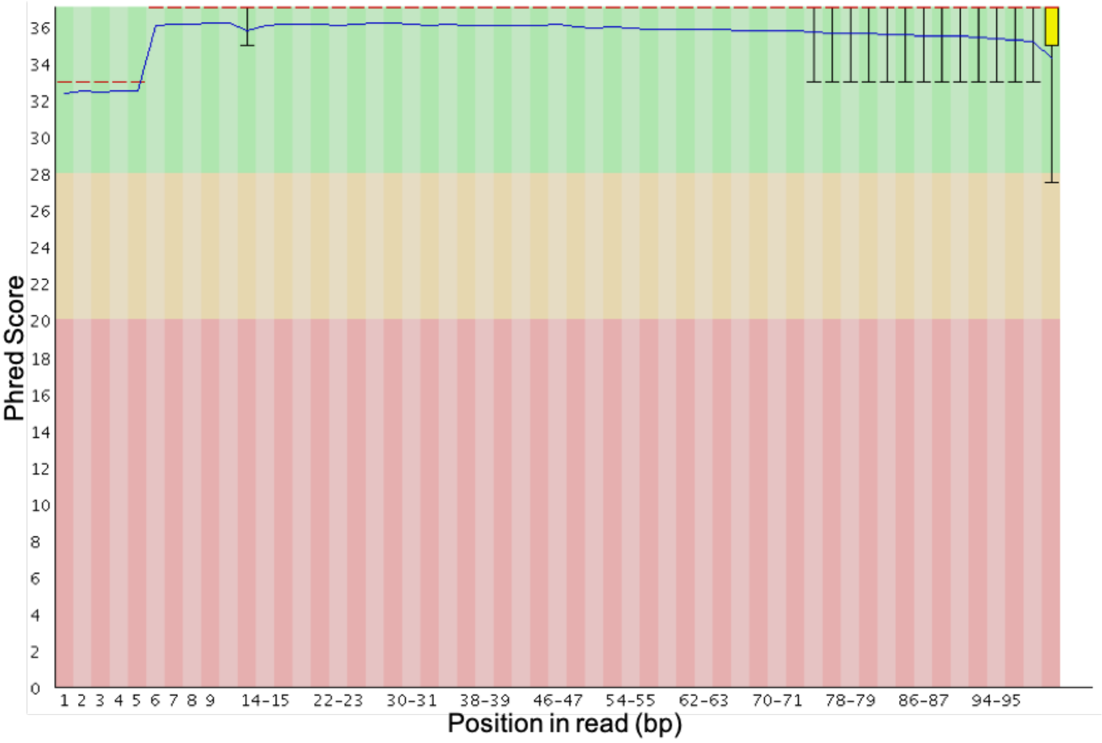


(A) Lane 3

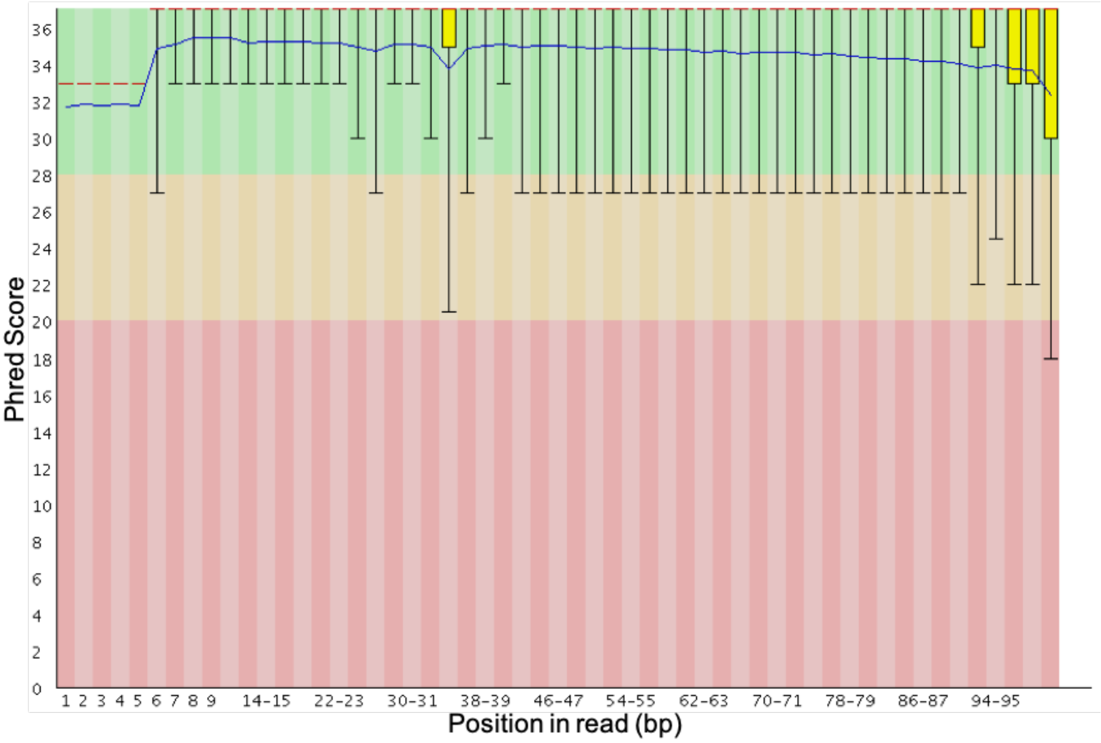


(B) Lane 4

FIGURE 4.7: **Per Base Sequence FASTQC analysis of Lane 3 and 4 of Genotyping By Sequencing (GBS).** On the x-axis is position in base pair, on the y-axis the Phred Score and graph background illustrates good quality (green), medium quality (orange) and poor quality (red) base calls. Original in colour.



(A) Lane 5



(B) Lane 6

FIGURE 4.8: **Per Base Sequence FASTQC analysis of Lane 5 and 6 of Genotyping By Sequencing (GBS).** On the x-axis is position in base pair, on the y-axis the Phred Score and graph background illustrates good quality (green), medium quality (orange) and poor quality (red) base calls. Original in colour.

4.3.3 Alignment post-processing

Minor Allele Frequency is a common filtration step for GWAS and Figure 4.9 explores the effect of MAF on number of sites. The largest change in number of sites is from 0 to 0.01, and then the total site number decreases in a linear fashion as MAF increases. This figure suggests MAF is an effective filtration step for reducing number of sites.

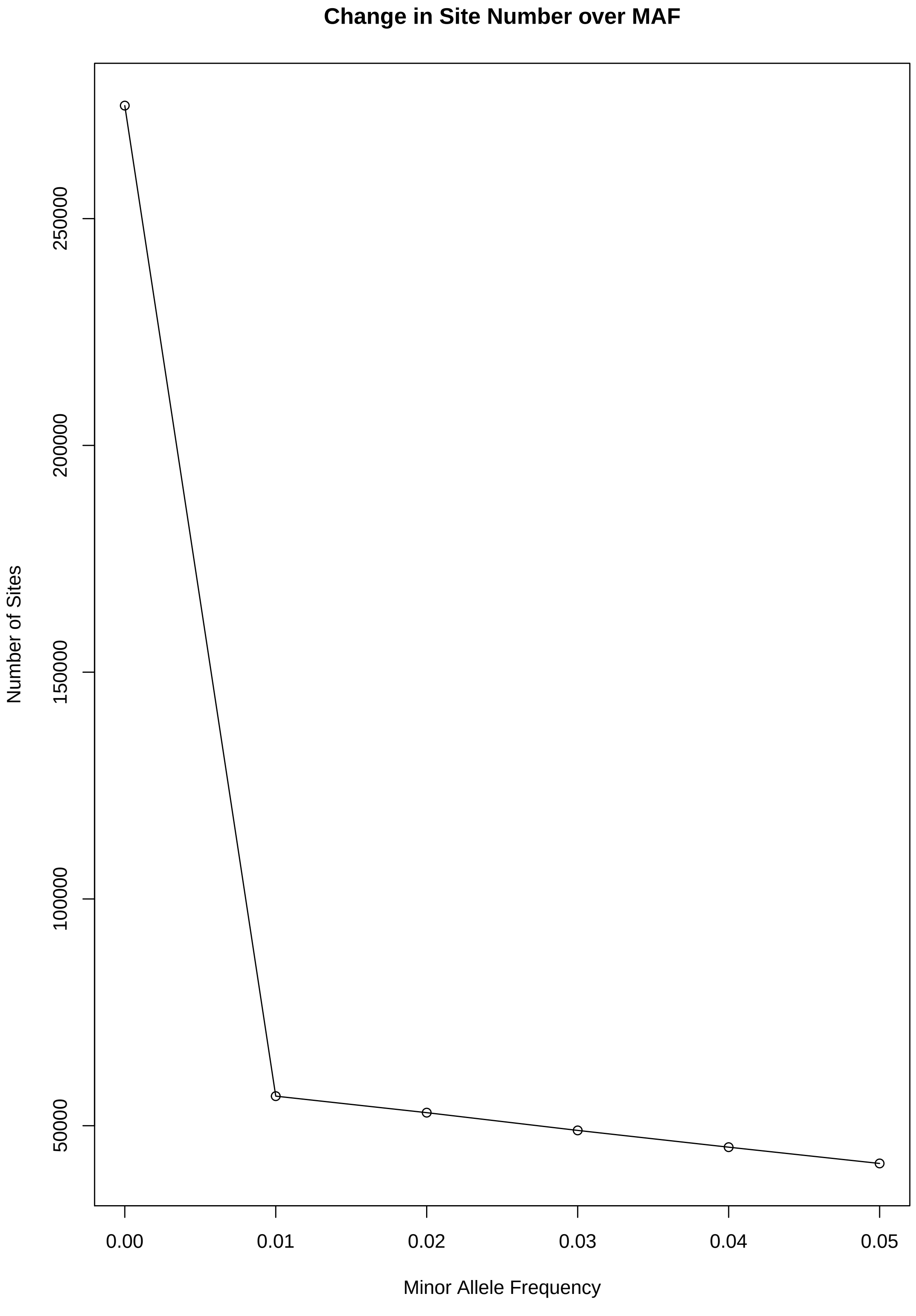
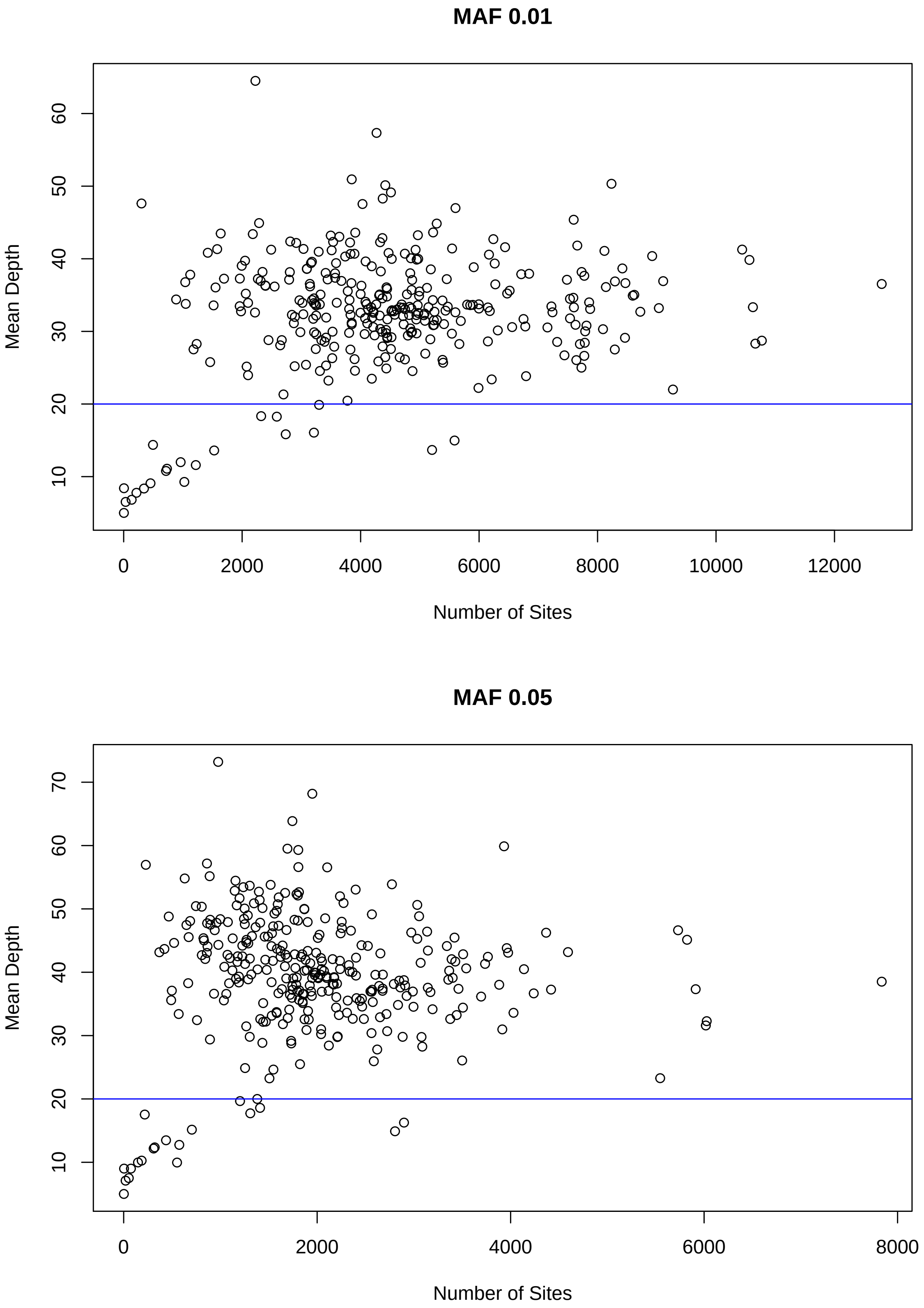


FIGURE 4.9: **Change in number of sites across changes in Minor Allele Frequency.**

Figure 4.10 explores mean depth of each site for MAF of 0.01 and 0.05. It can be observed that whilst MAF of 0.01 has samples with over 12,000 sites, samples tend to cluster around the 2,000-6,000 region for number of sites and it can also be observed that the mean depth of some samples can be as high as 60x. Whereas, with MAF 0.05, the sample with the greatest number of sites is approximately 8,000, yet most samples cluster at approximately 2,000 for number of sites, and mean depth of some samples can reach over 70x. This shows that with a higher MAF of 0.05, samples tend to have fewer sites, but in some cases, depth can increase.


MAF 0.01

MAF 0.05

FIGURE 4.10: **The number of sites and the mean depth for MAF 0.01 and 0.05.**Blue line represents mean depth of 20x. Sites with less than or equal to 20x depth were removed from downstream analysis.

Figure 4.11 explores the effect of MAF on the amount of missing data. Whilst missing data remains high in both cases, above 0.8, the MAF of 0.01 produces more samples with less missing data.

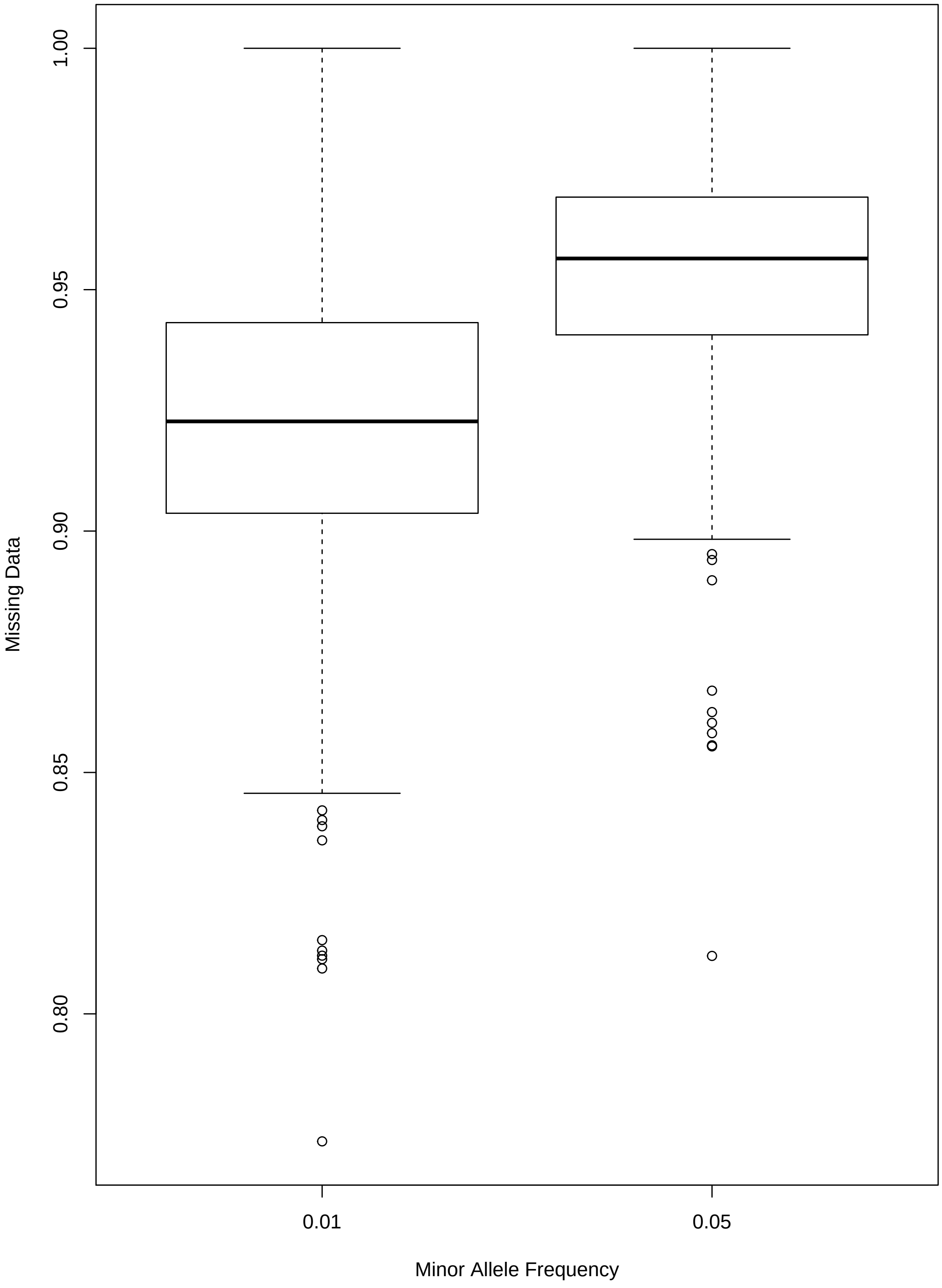


FIGURE 4.11: **Missing data of each site for Minor Allele Frequency of 0.01 and 0.05**

For MAF of 0.01, the effect of missing data on number of sites was explored (Figure 4.12). It can be observed that as missing data increases, the number of sites decreases, and the curve appears to exponentially decay. This shows that large amounts of missing data have a large effect on the reduction of number of sites.

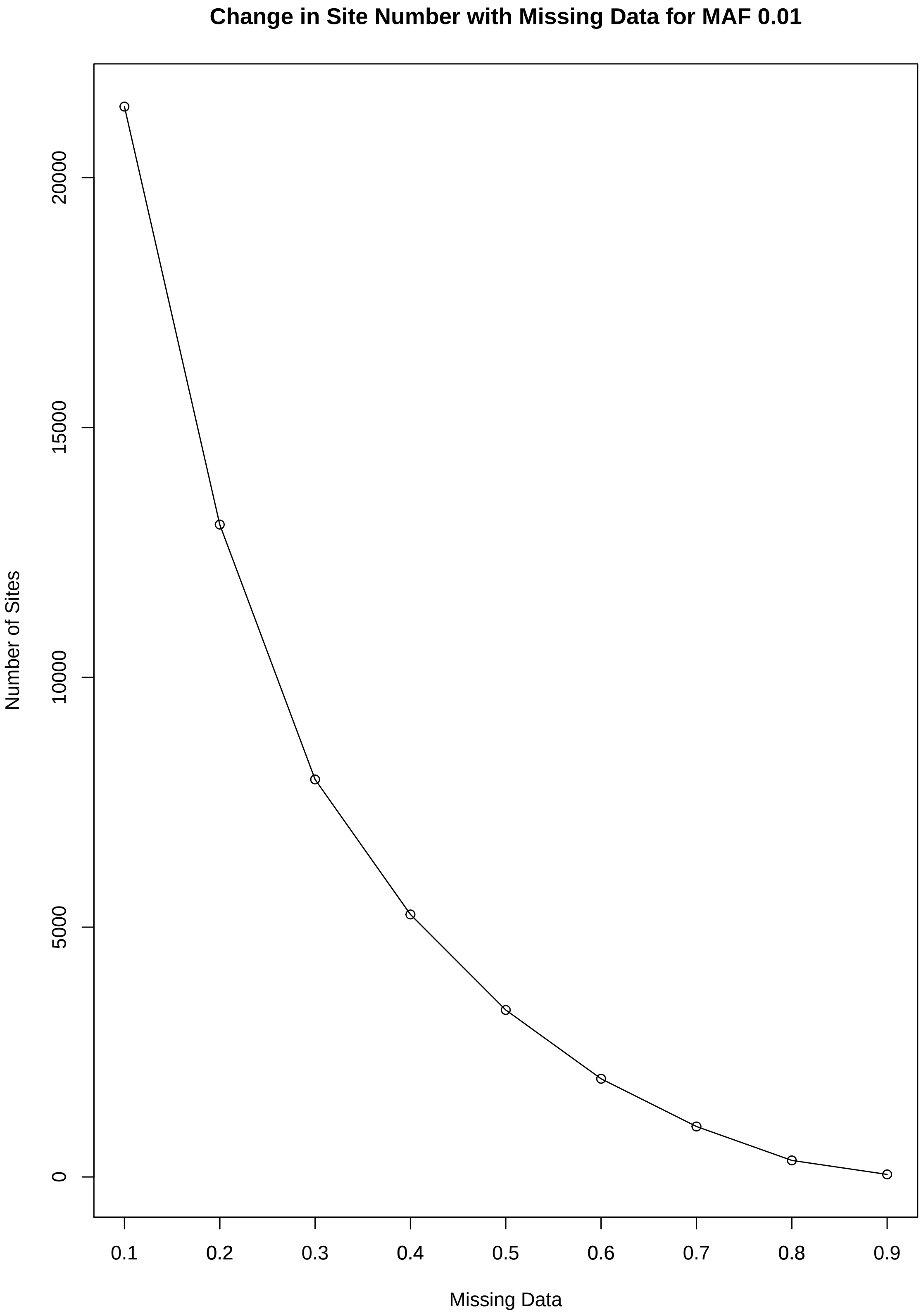


FIGURE 4.12: **Number of sites remaining as missing data increases for Minor Allele Frequency 0.01.**

4.3.4 Population analysis

To select the best number of populations (K) for Structure analysis, 10 runs were performed for each K (3 to 21) and the average of these runs were plotted, to select the best K (Figure 4.13). This figure shows that the best K to use would be 14, since it holds the highest change in probability for K.

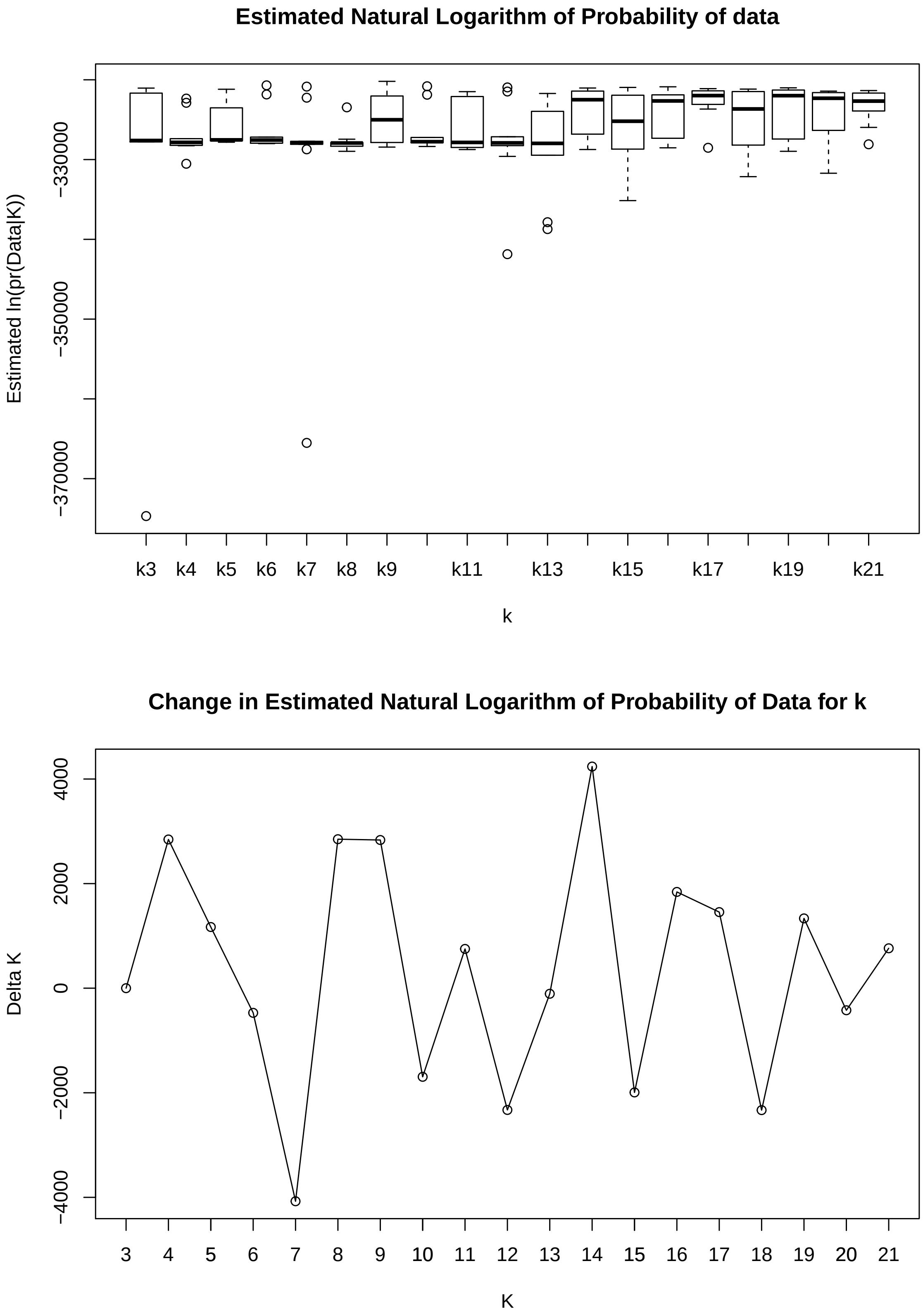


FIGURE 4.13: **Selecting the number of populations (K) for subsequent Structure Analysis.** Estimated Natural Logarithm for each number of populations (K) across ten runs (top graph) and Delta K of the average of these 10 runs (bottom graph). The selected number of populations is 14.

Based on K = 14, a structure analysis was performed (Figure 4.14).

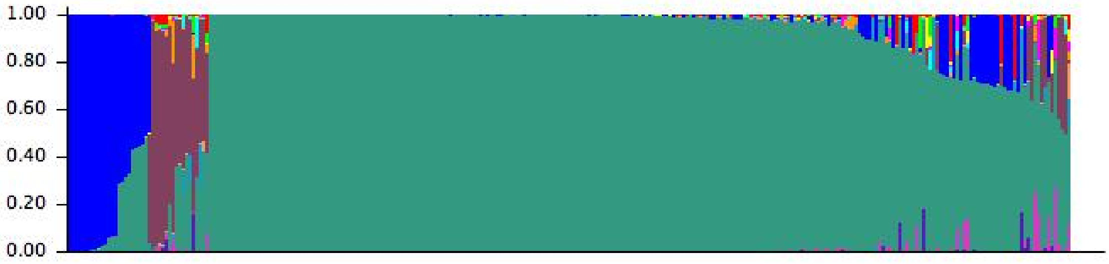


FIGURE 4.14: **Structure.** The structure analysis using the selected number of populations (K) = 14. Original in colour.

Figure 4.15 shows a phylogenetic tree analysis for these samples, with group 1 (landrace) in blue, group 2 (cultivars) in green, group 4 (wild) in red. This figure shows that wild varieties have clustering, and that most samples show the divergence between wild and domesticated material. Within the domesticated material, there is less clustering.

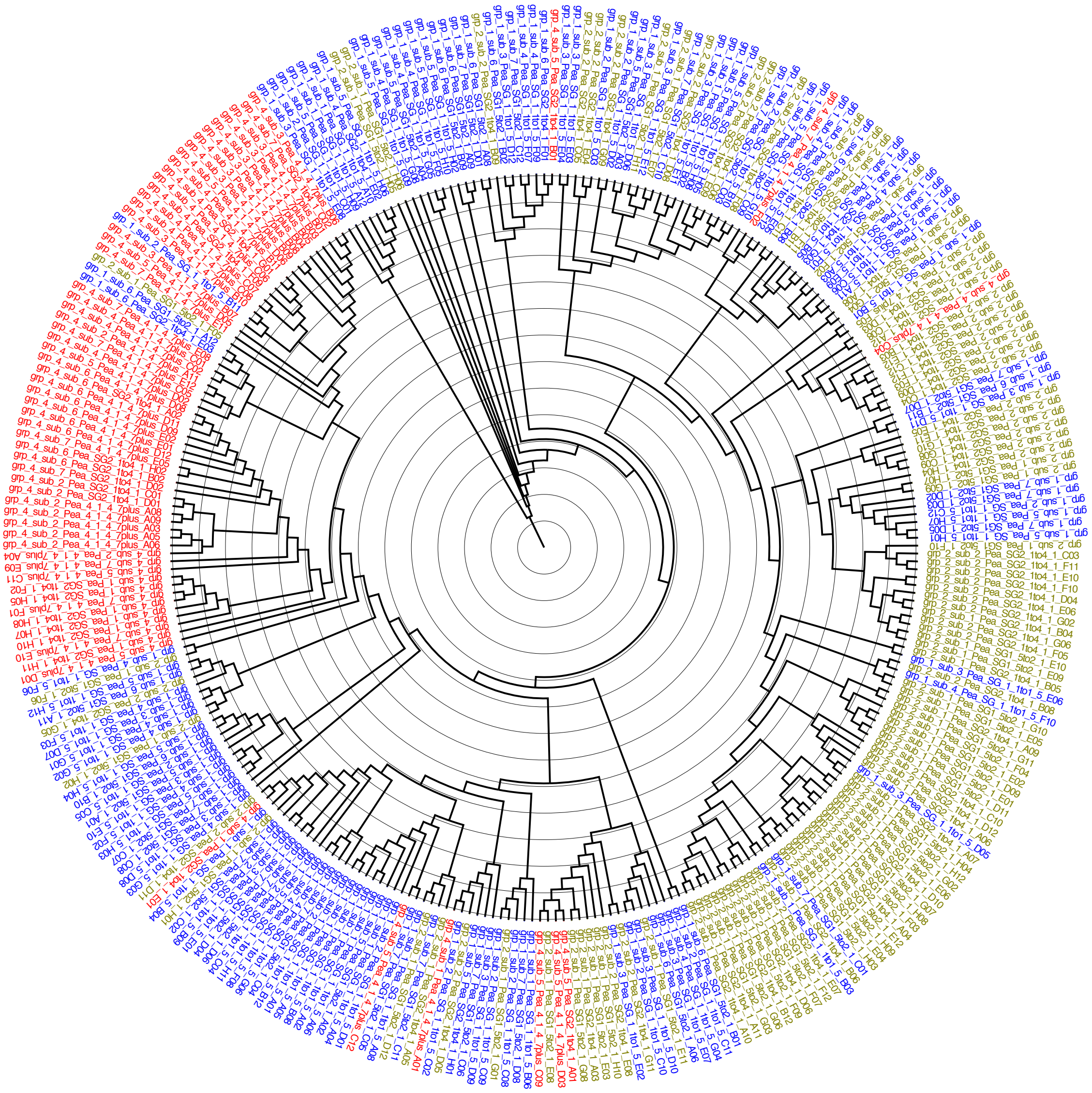


FIGURE 4.15: **Phylogenetic tree of the core collection formed from the SNPs produced.** Group 1 (landraces) shown in blue, Group 2 (cultivars) in green and Group 4 (wild) in red. Original in colour.

Discriminant Analysis of Principal Components (DAPC) based on the SNPs called shows no clear outliers, suggesting that each group is distinct and does not suggest errors in sampling techniques such as incorrect labelling (Figure 4.16). Group 4 appears further apart from the domesticated material across the horizontal eigenvector, interestingly across the vertical eigenvector it sits slightly closer to group 1 (landraces) than it does to group 2 (cultivated). This suggests that there may be a genomic site responsible for a large domestication bottleneck observed in the horizontal eigenvector, but also another site responsible for the change from landrace to cultivar such as a potential smaller modernisation bottleneck.

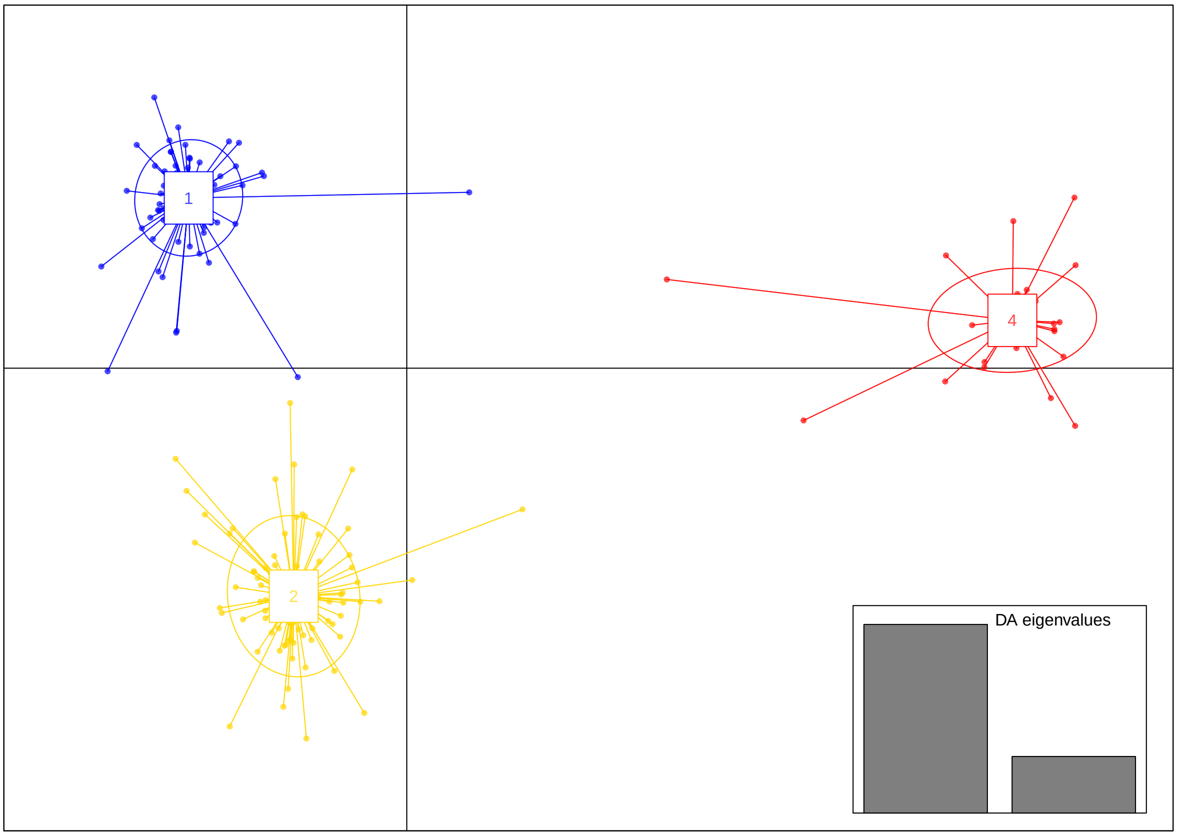


FIGURE 4.16: **Discriminant Analysis of Principal Components (DAPC) based on the SNPs produced.** Group 1 (landraces) shown in blue, Group 2 (cultivars) in green and Group 4 (wild) in red. Original in colour.

4.3.5 GWAS Analysis

Plant Height

Genome-wide association studies were conducted for plant height grown in the glasshouse. Manhattan plots were drawn for each pea pseudo-chromosome and the probability of each SNP being associated with the phenotype alongside Q-Q plots for the probability of each SNP association for plant height. No SNPs were found to be statistically associated with plant height (Figure 4.17). The largest peak was found in pseudo-chromosome 2 and the next largest found in pseudo-chromosome 1. The Q-Q plot shows some distension away from the reference line which is expected on the right-hand side but these were not statistically significant (Figure 4.18).

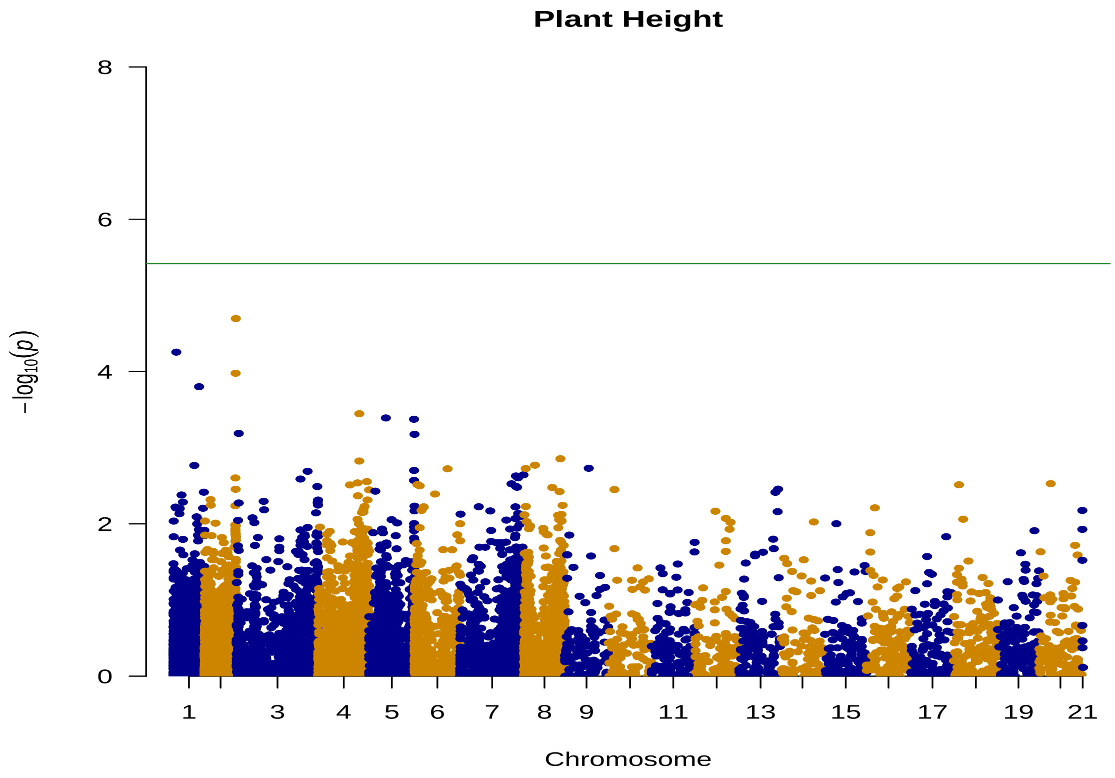


FIGURE 4.17: **Genome-wide Association Study for Plant Height.** Manhattan Plot for plant height of the core collection grown in the glasshouse. Original in colour.

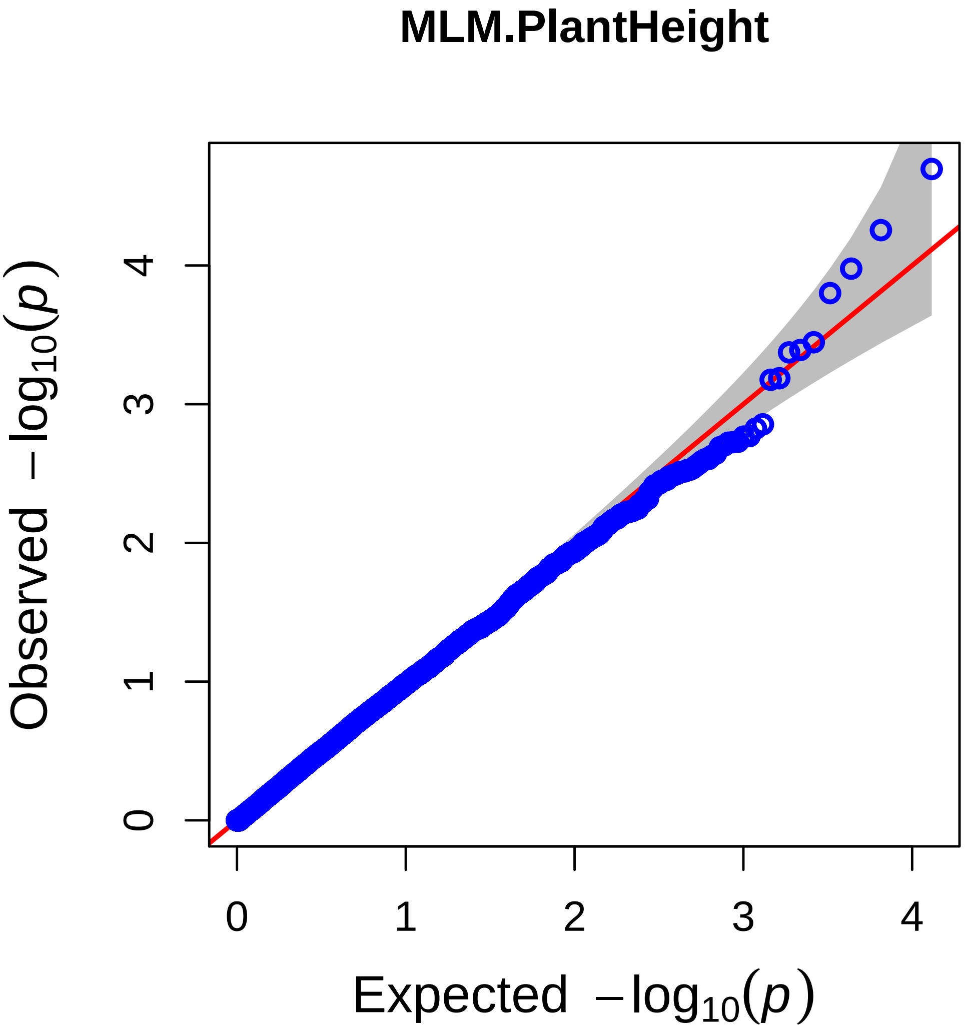


FIGURE 4.18: **Genome-wide Association Study for Plant Height.** Quantile-Quantile plot for plant height of the core collection grown in the glasshouse. Original in colour.

Seed

Genome-wide association studies were conducted for seed weight grown in the glasshouse. Manhattan plots were drawn for each pea pseudo-chromosome and the probability of each SNP being associated with the phenotype alongside Q-Q plots for the probability of each SNP association for seed weight. No SNPs were found to be statistically associated with seed weight (Figure 4.19). The highest peak was found to be in pseudo-chromosome 5, with the next highest peaks found in pseudo-chromosomes 1, 8 and 19. The Q-Q plot shows that most SNPs follow the reference line, indicating that there is no association between these SNPs (Figure 4.20).

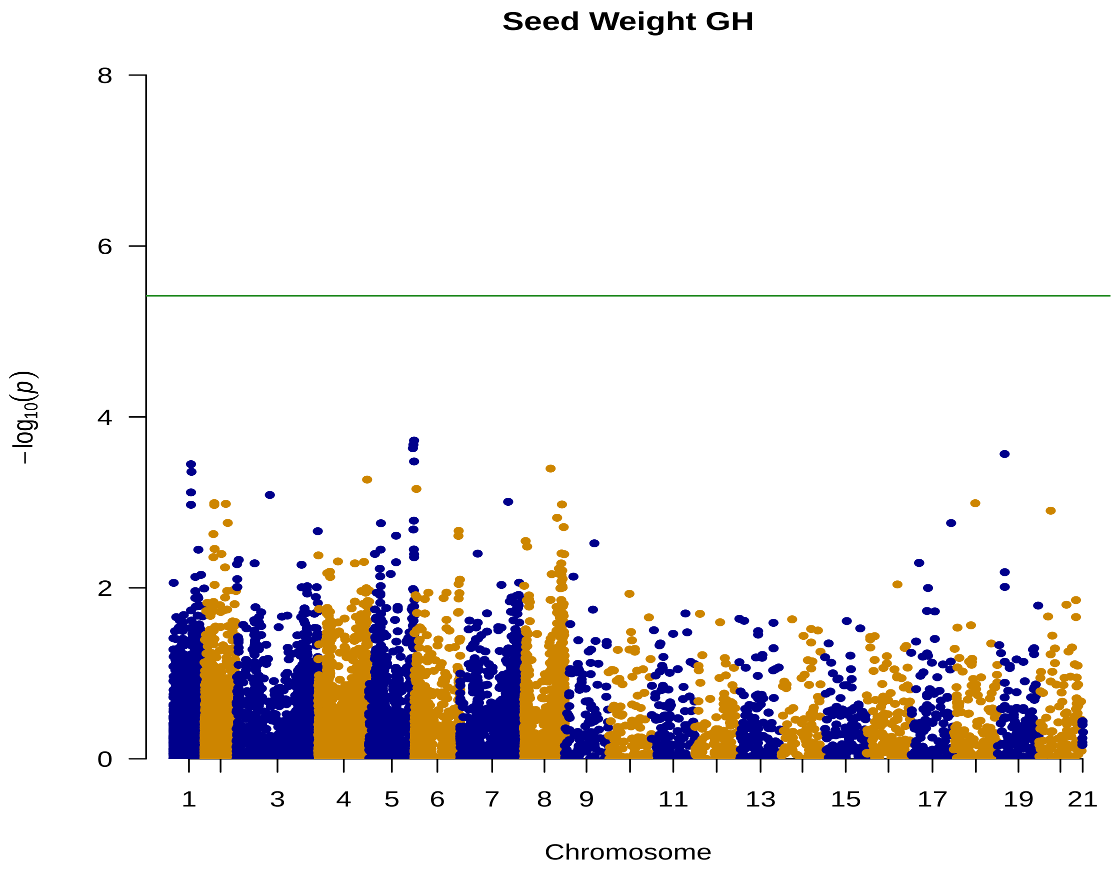


FIGURE 4.19: **Genome-wide Association Study for Seed Weight.** Manhattan Plot for seed weight of the core collection grown in the glasshouse. Original in colour.

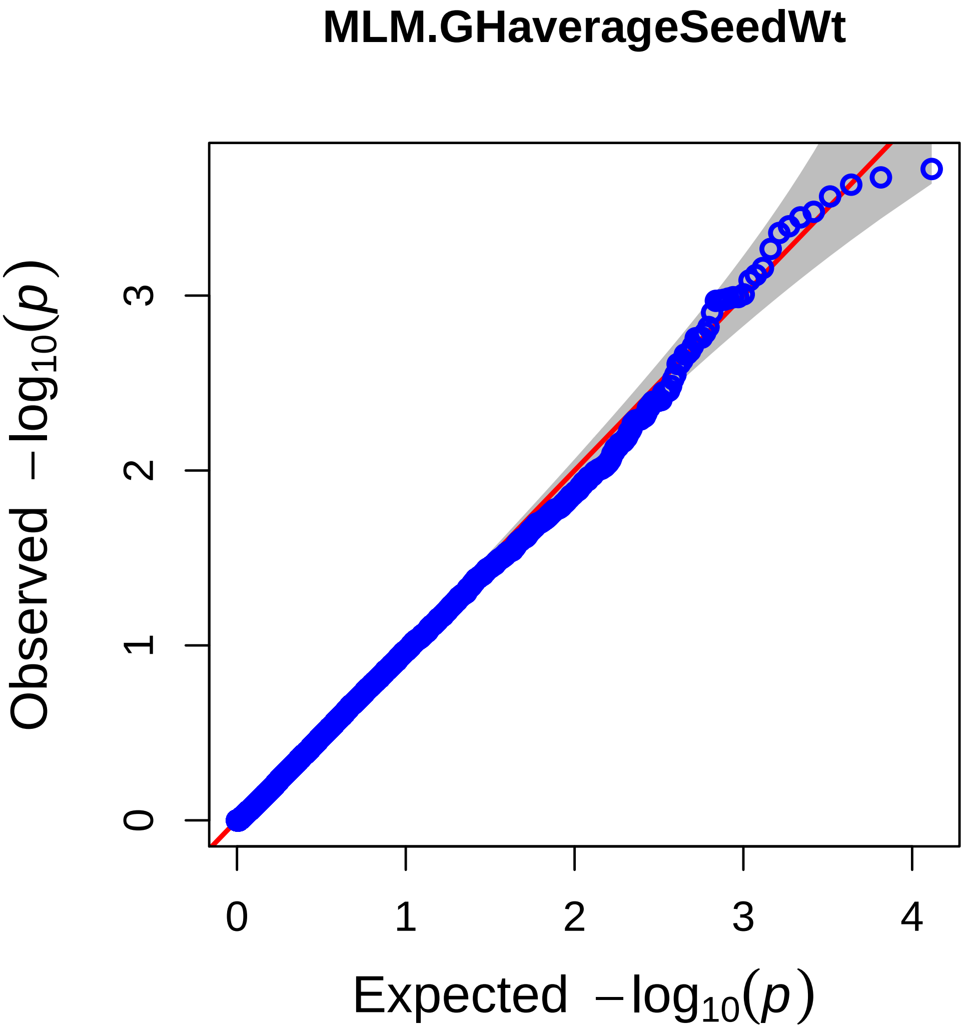


FIGURE 4.20: **Genome-wide Association Study for Seed Weight.** Quantile-Quantile plot for seed weight of the core collection grown in the glasshouse. Original in colour.

Seed area of peas grown in the glasshouse also showed no SNPs were found to be statistically associated with seed area. The highest peak was found to be in pseudo-chromosome 6 (Figure 4.21). The Q-Q plot shows the samples following the reference line with little movement way from this line, which indicates that there are no statistically associated loci (Figure 4.22).

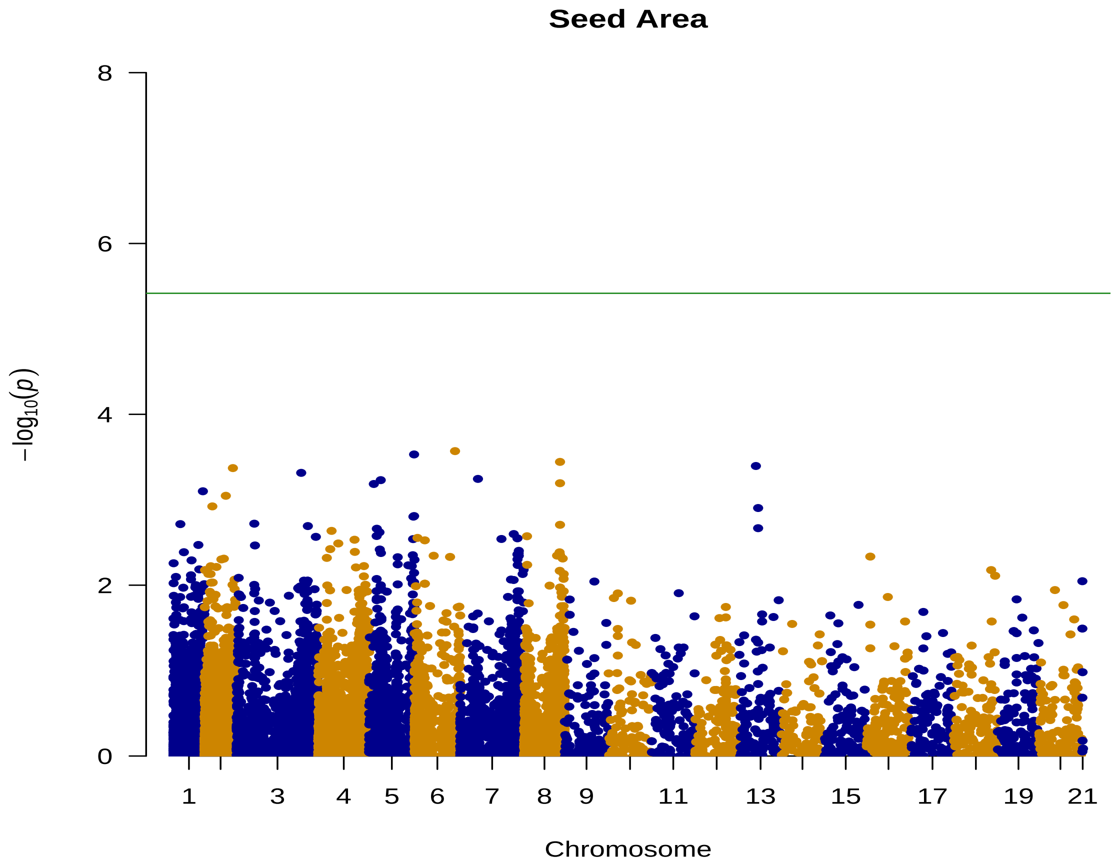


FIGURE 4.21: **Genome-wide Association Study for Seed Area.** Manhattan Plot for seed area of the core collection grown in the glasshouse. Original in colour.

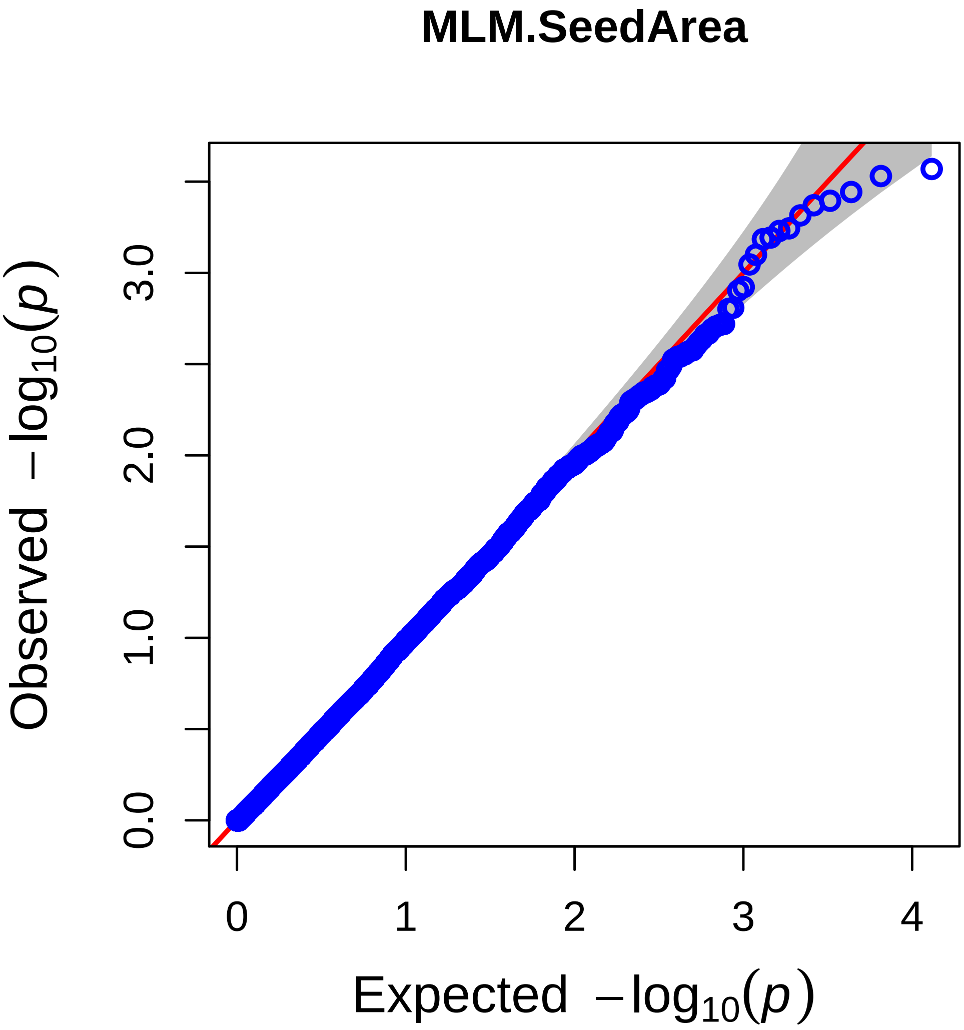


FIGURE 4.22: **Genome-wide Association Study for Seed Area.** Quantile-Quantile plot for seed area of the core collection grown in the glasshouse. Original in colour.

Seed Equivalent Diameter of peas grown in the glasshouse also showed no SNPs were found to be statistically associated with seed equivalent diameter (Figure 4.23). The highest peak was found to be in pseudo-chromosome 7. The Q-Q plot shows some distension above the reference line and this may indicate spurious associations (Figure 4.24).

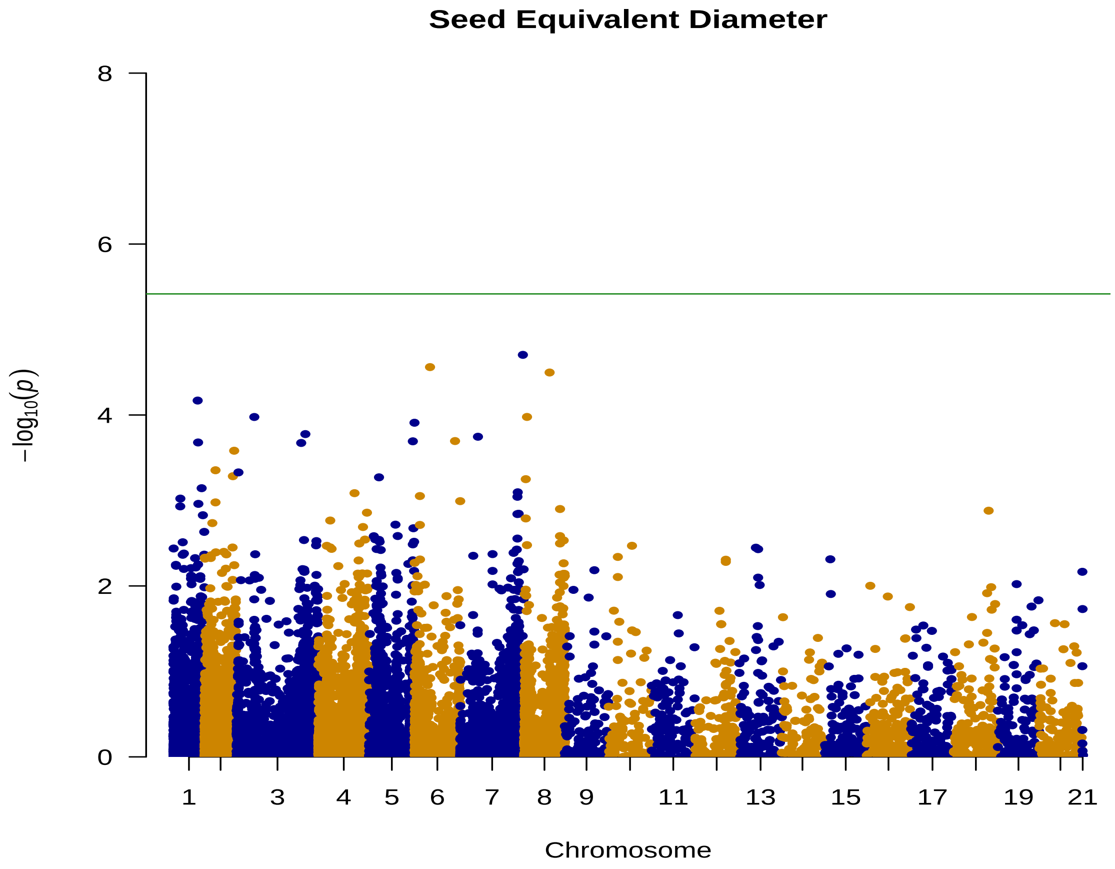


FIGURE 4.23: **Genome-wide Association Study for Seed Equivalent Diameter.** Manhattan Plot for seed equivalent diameter of the core collection grown in the glasshouse. Original in colour.

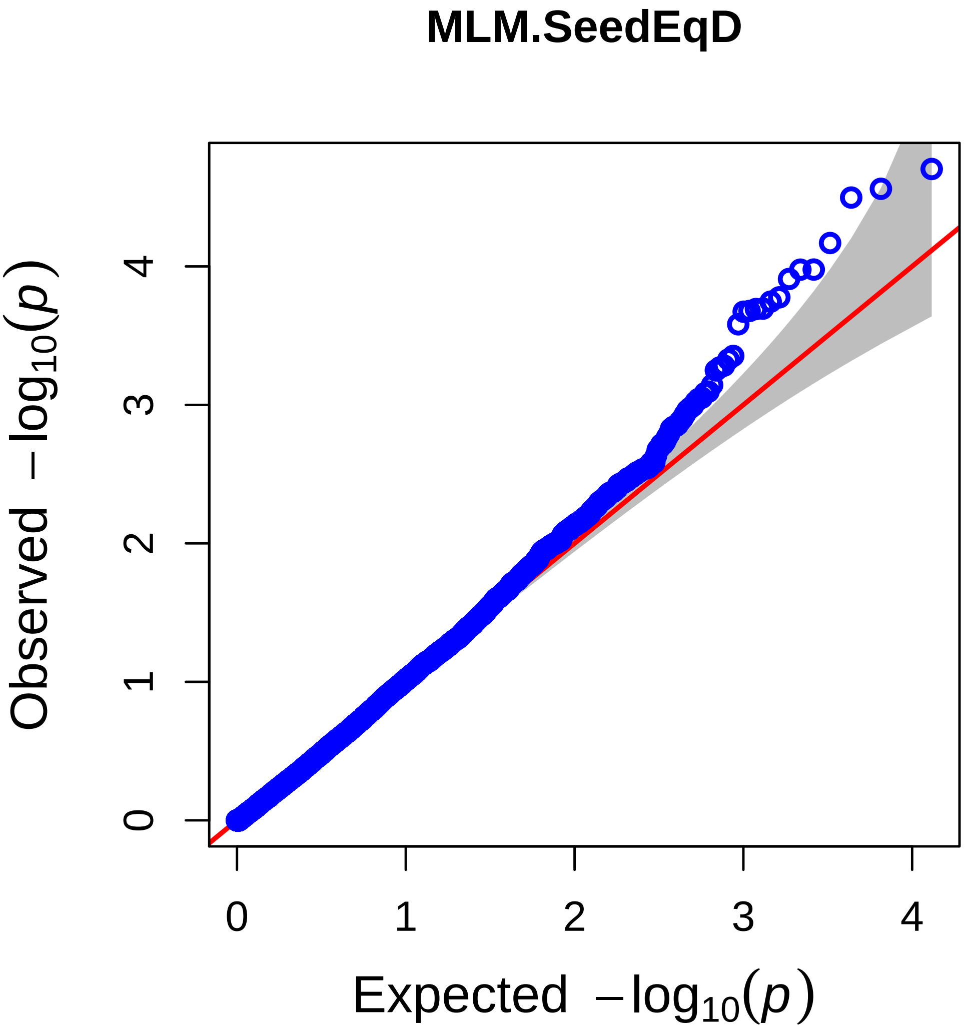


FIGURE 4.24: **Genome-wide Association Study for Seed Equivalent Diameter.** Quantile-Quantile plot for seed equivalent diameter of the core collection grown in the glasshouse. Original in colour.

Seed Perimeter of those grown in the glasshouse has a statistically significant SNP in Chromosome 7 (Figure 4.25). The Q-Q plot shows some large distension above the curve indicating that the higher data point is the statistically significant SNP and the remaining points that distend from the curve are spurious associations (Figure 4.26). Figure 4.27 shows this association is towards the end of chromosome 7 and lies above the cut off threshold.

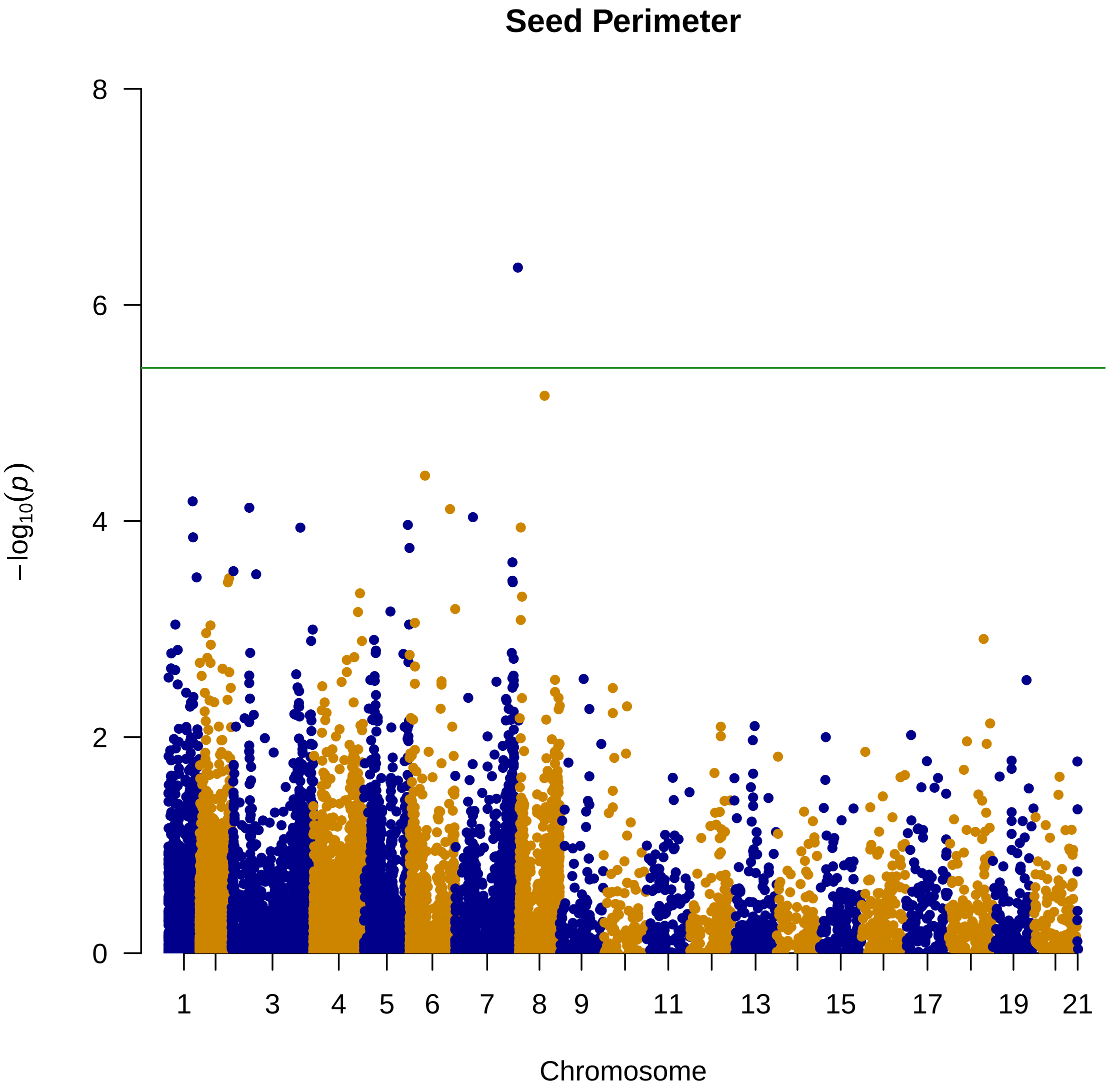


FIGURE 4.25: **Genome-wide Association Study for Seed Perimeter.** Manhattan plot for seed perimeter of the core collection grown in the glasshouse. The Manhattan plot displays the statistically significant threshold as a green line, SNPs above this line hold a statistically significant association with seed perimeter. Original in colour.

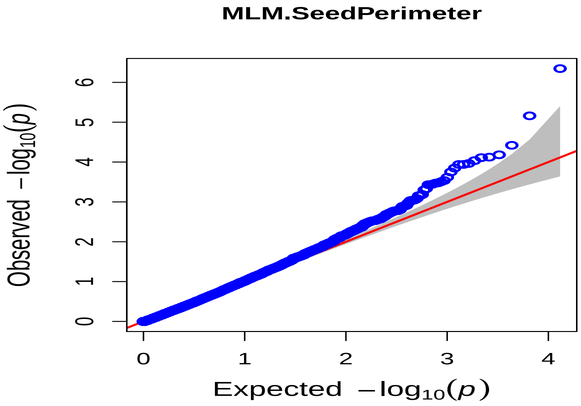


FIGURE 4.26: **Genome-wide Association Study for Seed Perimeter.** Quantile-Quantile plot for seed perimeter of the core collection grown in the glasshouse. Original in colour.

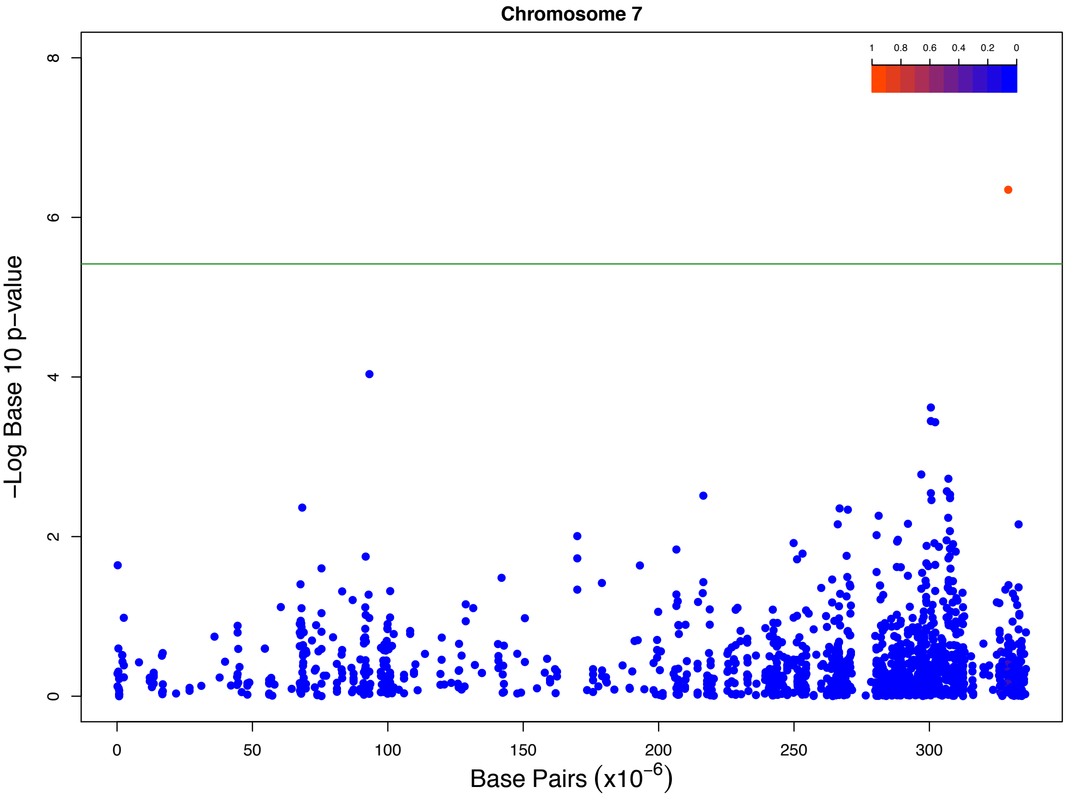


FIGURE 4.27: **Manhattan plot of Chromosome 7 from the Genome-wide Association Study for Seed Perimeter.** The Manhattan plot displays the statistically significant threshold as a green line, SNPs above this line hold a statistically significant association with seed perimeter. Original in colour.

Seed Eccentricity of those grown in the glasshouse has statistically significant SNPs in Chromosomes 3 and 8 (Figure 4.28). The Q-Q plot exhibits larges distension from the curve indicating that the highest two data points are the statistically significant SNPs and the remainder of data points that distend from the curve are spurious (Figure 4.29).

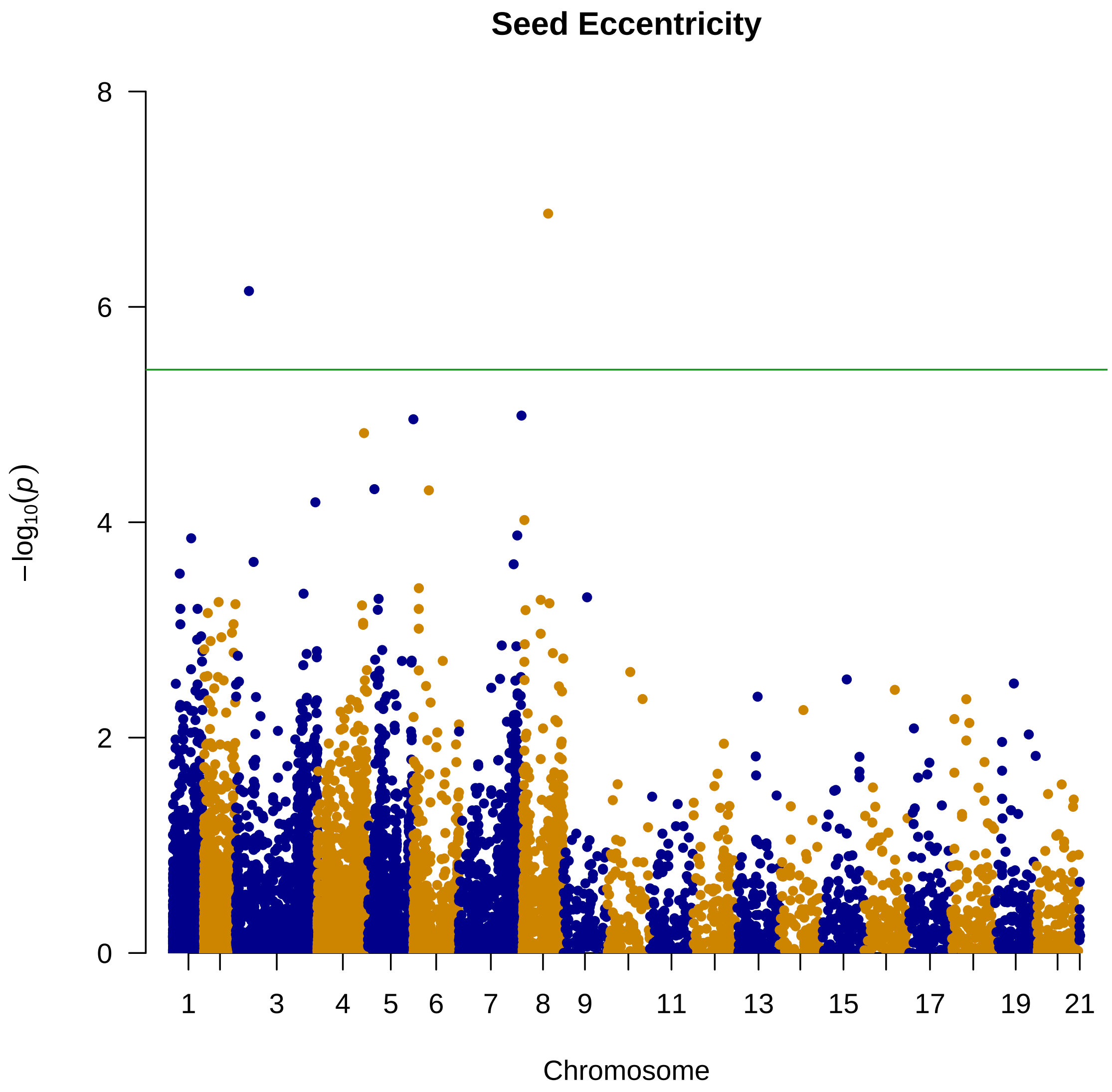


FIGURE 4.28: **Genome-wide Association Study for Seed Eccentricity.** Manhattan plot for seed perimeter of the core collection grown in the glasshouse. The Manhattan plot displays the statistically significant threshold as a green line, SNPs above this line hold a statistically significant association with seed eccentricity. Original in colour.

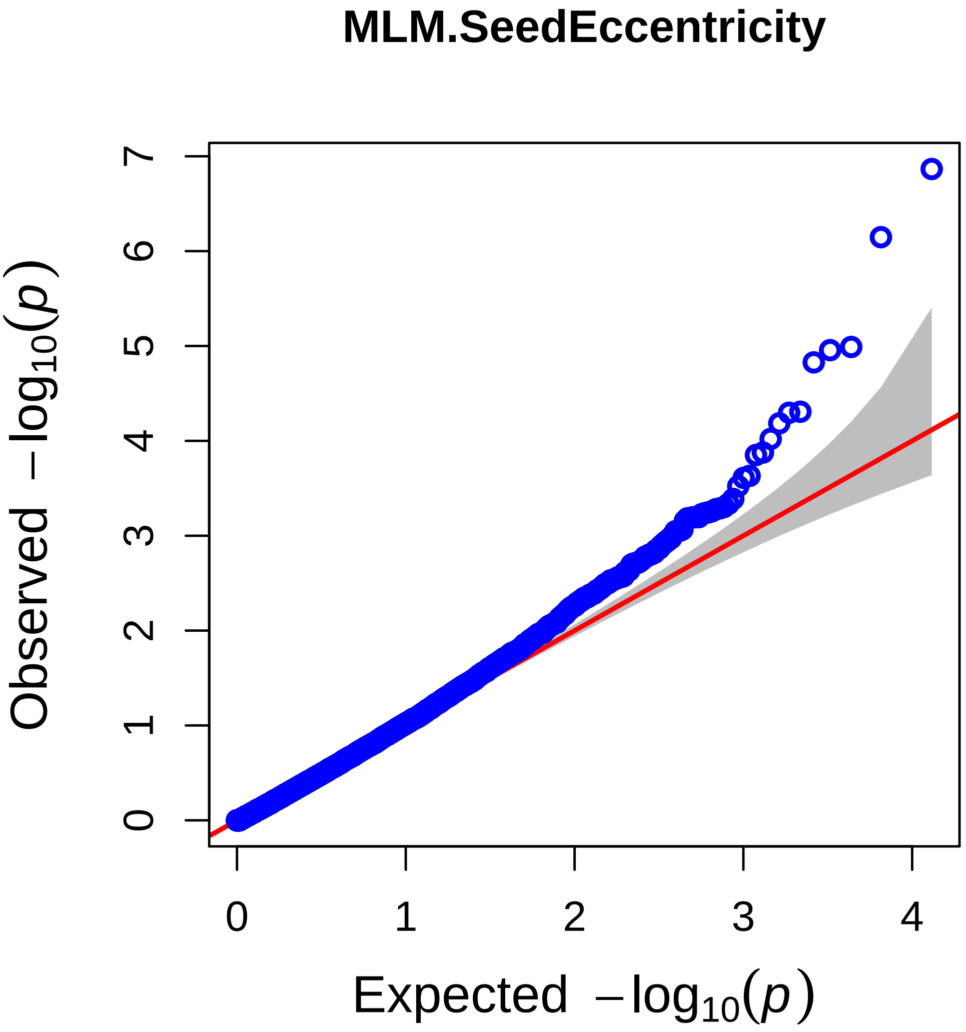
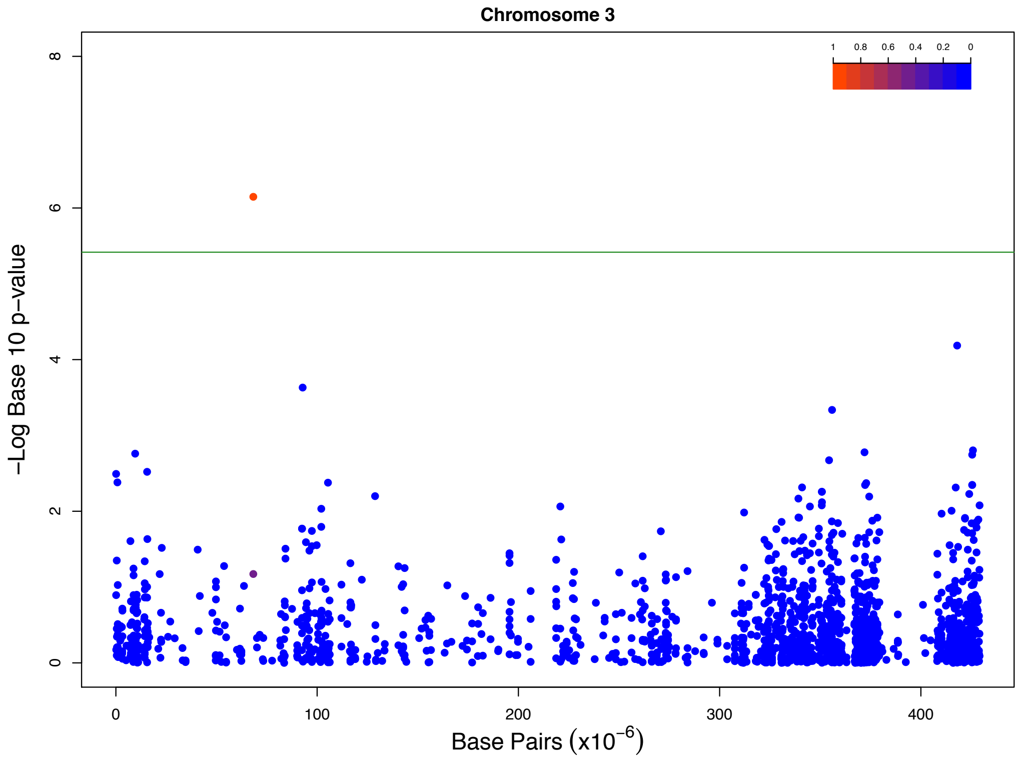


FIGURE 4.29: **Genome-wide Association Study for Seed Eccentricity.** Quantile-Quantile plot for seed perimeter of the core collection grown in the glasshouse. Original in colour.

Figure 4.30 shows that the SNP in chromosome 3 lies towards the start of the chromosome whereas the SNP in chromosome 8 lies towards the middle of the chromosome. The figure also shows that the p-value for the SNP in Chromosome 8 is higher than that of Chromosome 3.



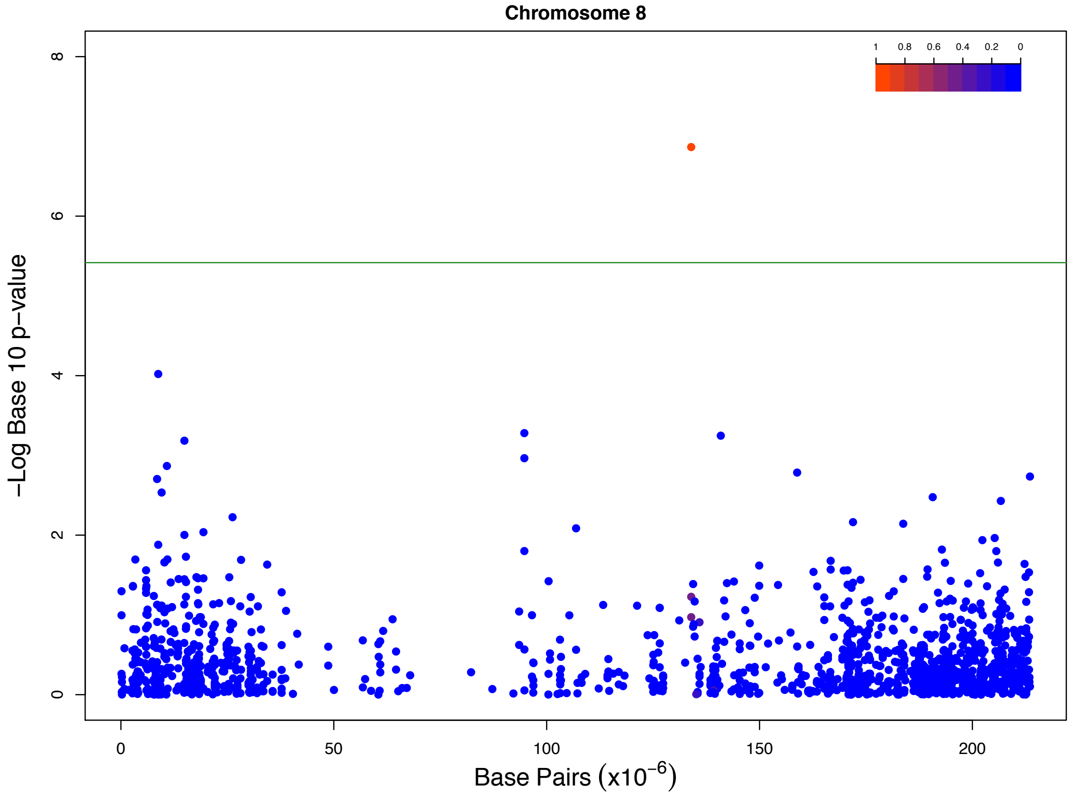
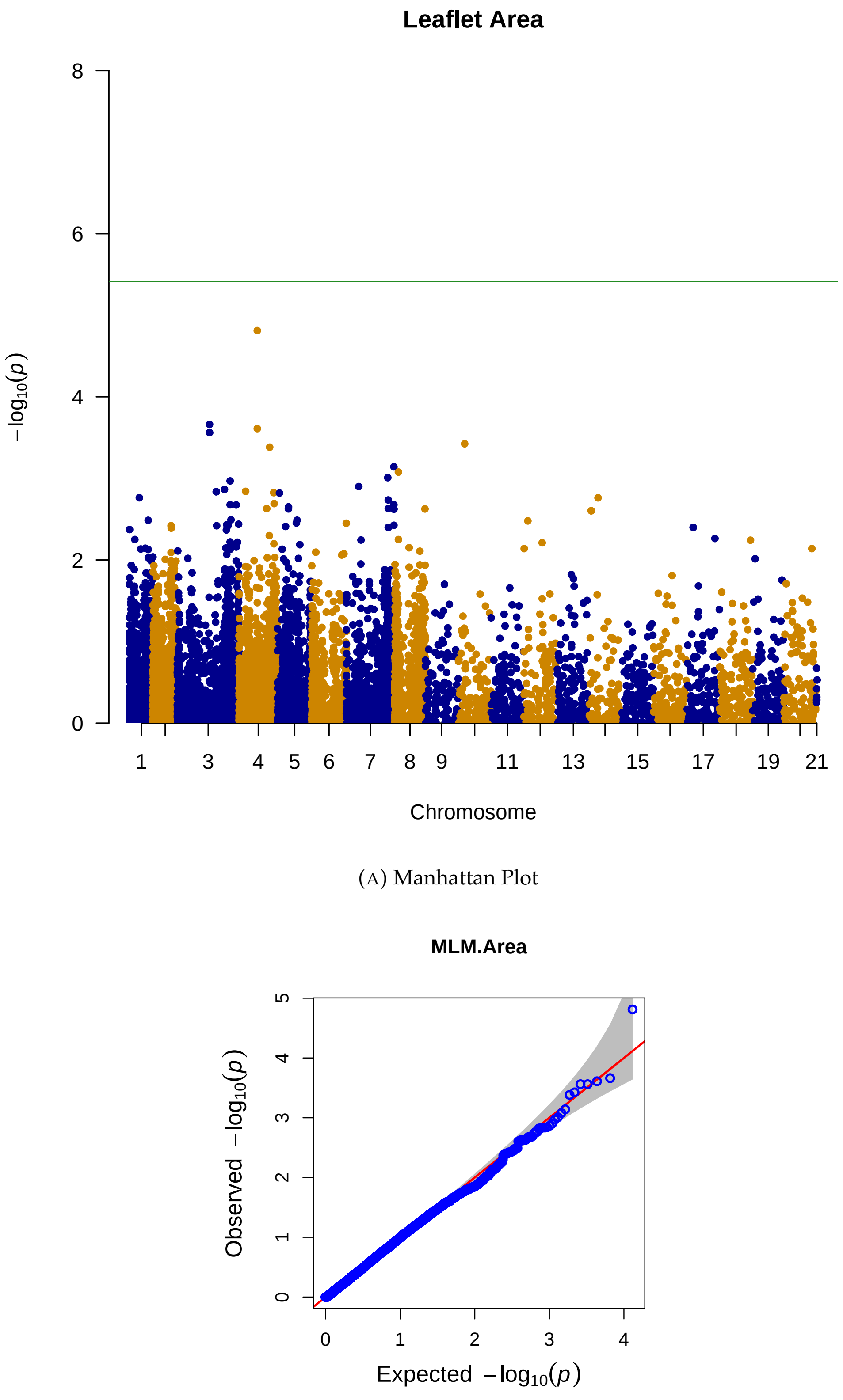


FIGURE 4.30: **Manhattan plot of Chromosome 3 and 8 from the Genome-wide Association Study for Seed Eccentricity.** The Manhattan plot displays the statistically significant threshold as a green line, SNPs above this line hold a statistically significant association with seed eccentricity. Original in colour.

Leaflet

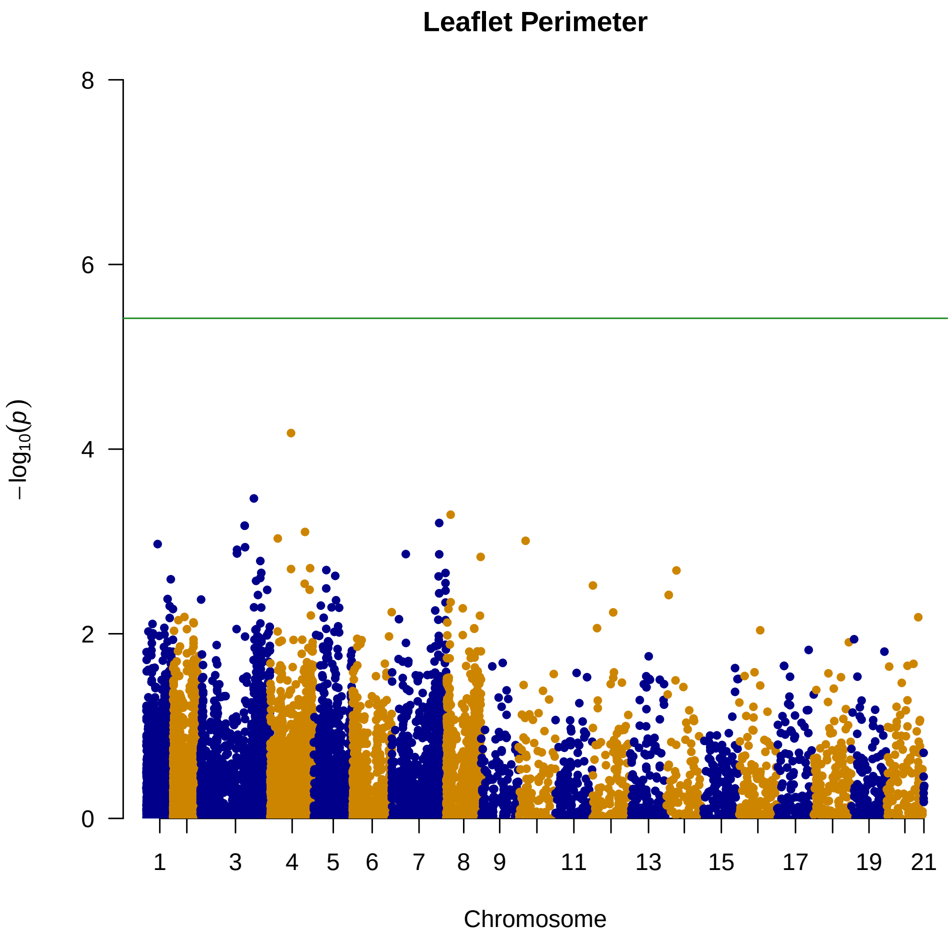
Genome-wide association studies were conducted for leaflet phenotypes. Manhattan plots were drawn for each pea pseudo-chromosome and the probability of each SNP being associated with the phenotype alongside Q-Q plots for the probability of each SNP association for the phenotypes for leaflet area, perimeter, length and width (Figure 4.31, 4.32, 4.33 and 4.34 respectively). For each of these phenotypes, no statistically significant associations were found. For all phenotypes, the highest peak was found in the pseudo-chromosome 4. For leaflet area, the next highest peak was found in pseudo-chromosome 3, for perimeter the next highest peak was found in pseudo-chromosomes 3 and 8, for leaflet length the next highest peak was found in pseudo-chromosomes 8 and for leaflet width the next highest peak was found in pseudo-chromosome 3. This suggests that although not statistically significant, potential loci of interest for leaflet lies in pseudo-chromosomes 3, 4 and 8. Q-Q plots show some distension away from the curve in the right-hand region. Most of the SNPs will fall along the reference line due to not being statistically significant. Deviations away from this line are expected in the right-hand region due to being associated SNPs. However, since they do not lie above the statistically significant threshold, this may indicate spurious associations.



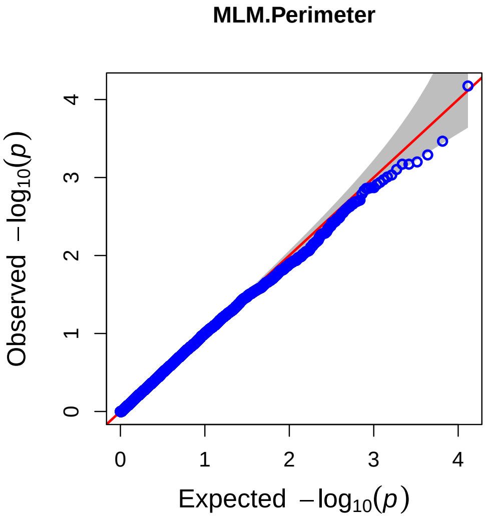
(A) Manhattan Plot

(B) Quantile-Quantile Plot

FIGURE 4.31: **Genome-wide Association Study for Leaflet Area.** Manhattan plot (A) and Quantile-Quantile plot (B) for leaflet area of the core collection grown in the glasshouse. Original in colour.

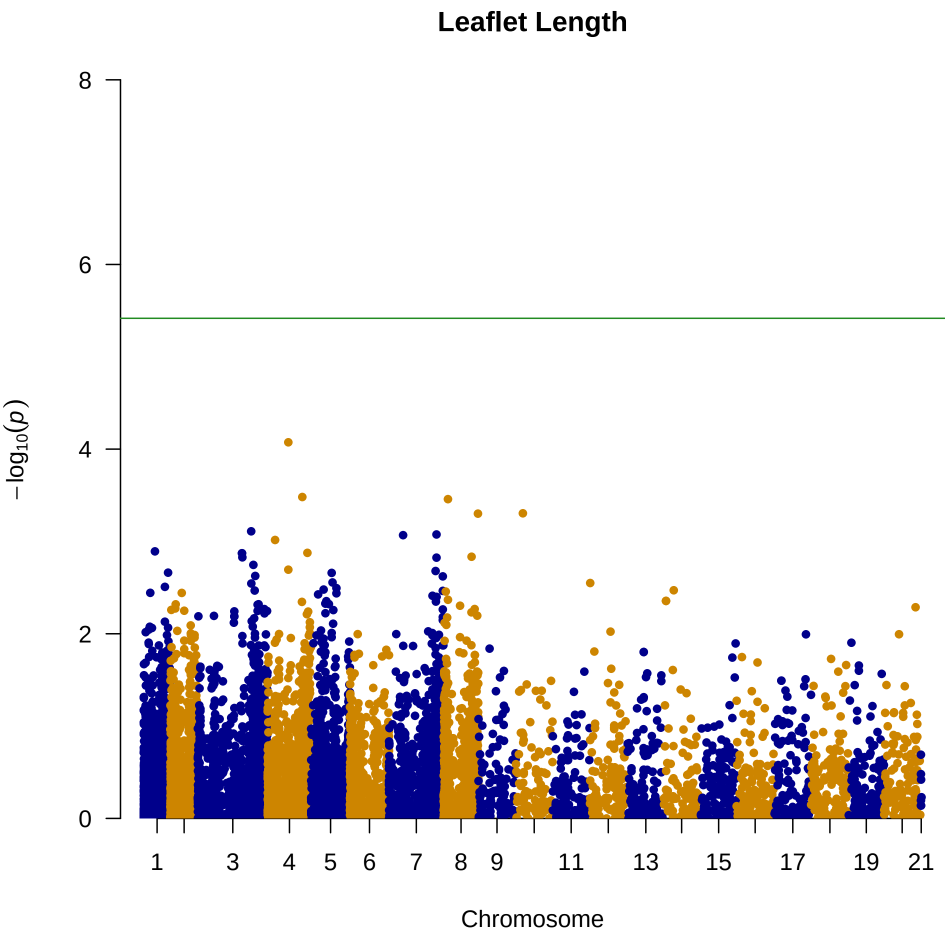


(A) Manhattan Plot

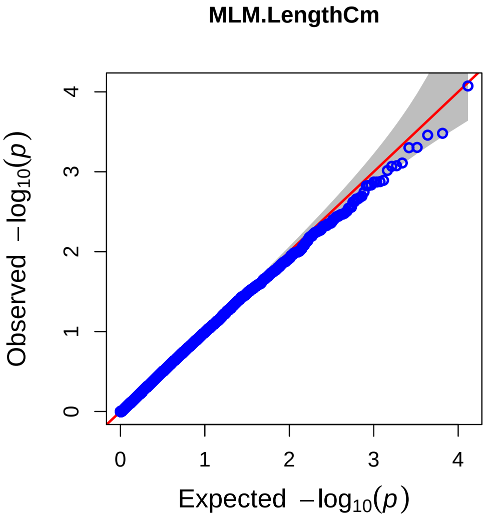


(B) Quantile-Quantile Plot

FIGURE 4.32: **Genome-wide Association Study for Leaflet Perimeter.** Manhattan plot (A) and Quantile-Quantile plot (B) for leaflet perimeter of the core collection grown in the glasshouse. Original in colour.

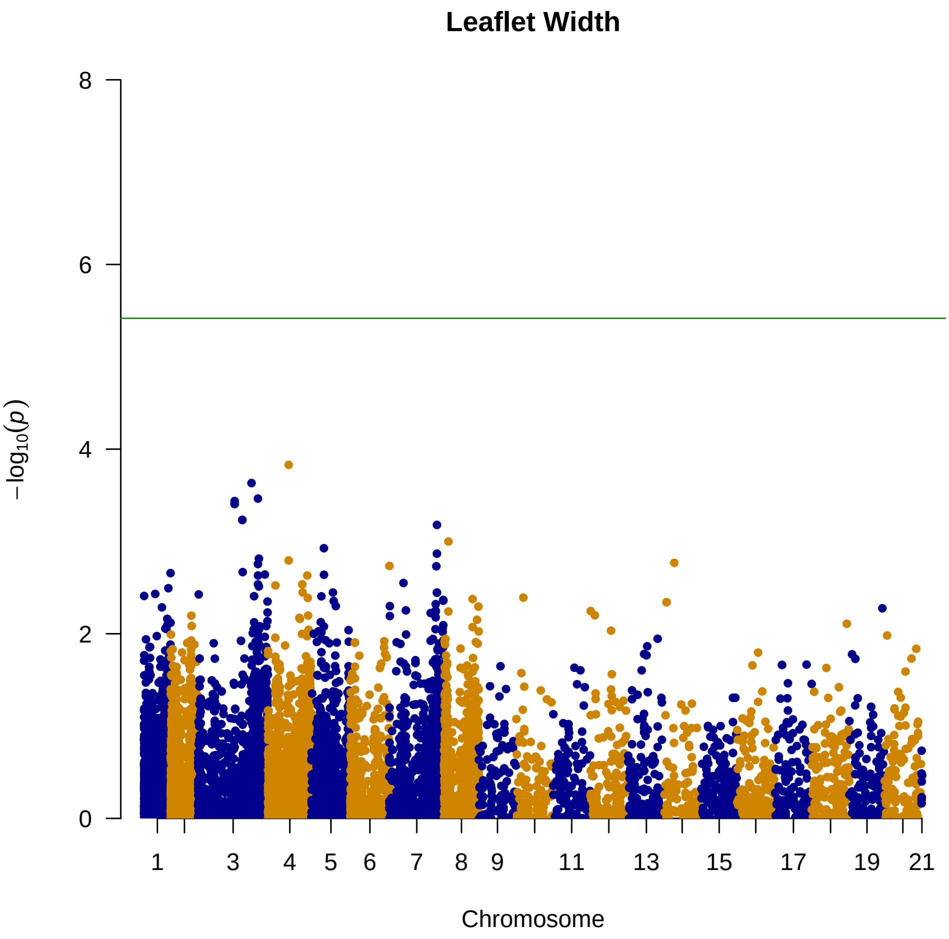


(A) Manhattan Plot

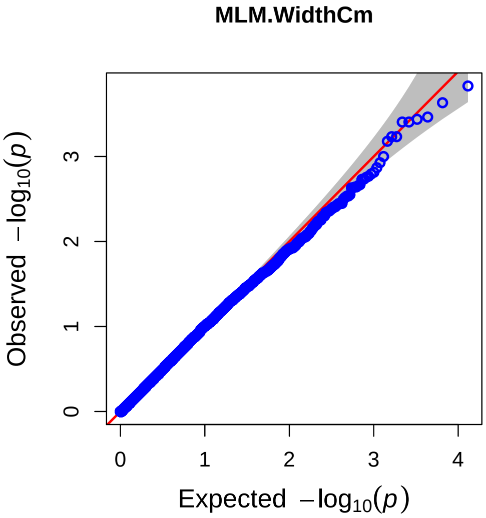


(B) Quantile-Quantile Plot

FIGURE 4.33: **Genome-wide Association Study for Leaflet Length.** Manhattan plot (A) and Quantile-Quantile plot (B) for leaflet length of the core collection grown in the glasshouse. Original in colour.



(A) Manhattan Plot



(B) Quantile-Quantile Plot

FIGURE 4.34: **Genome-wide Association Study for Leaflet Width.** Manhattan plot (A) and Quantile-Quantile plot (B) for leaflet width of the core collection grown in the glasshouse. Original in colour.

A genome-wide association study was performed on leaflet teeth. The statistically significant peak found was in pseudo-chromosome 1 (Figure 4.35 and 4.36). The Q-Q plot shows distension away from the reference line (Figure 4.37), which is expected, suggesting a number of SNPs are associated with this phenotype, however, since only one SNP was statistically significant, this could indicate the presence of spurious associations. Additional SNPs were found close to the cut-off threshold in pseudo-chromosomes 2 and 3 (Figure 4.38). This indicates that they may be loci of potential interest for leaflet teeth in pseudo-chromosomes 1, 2 and 3.

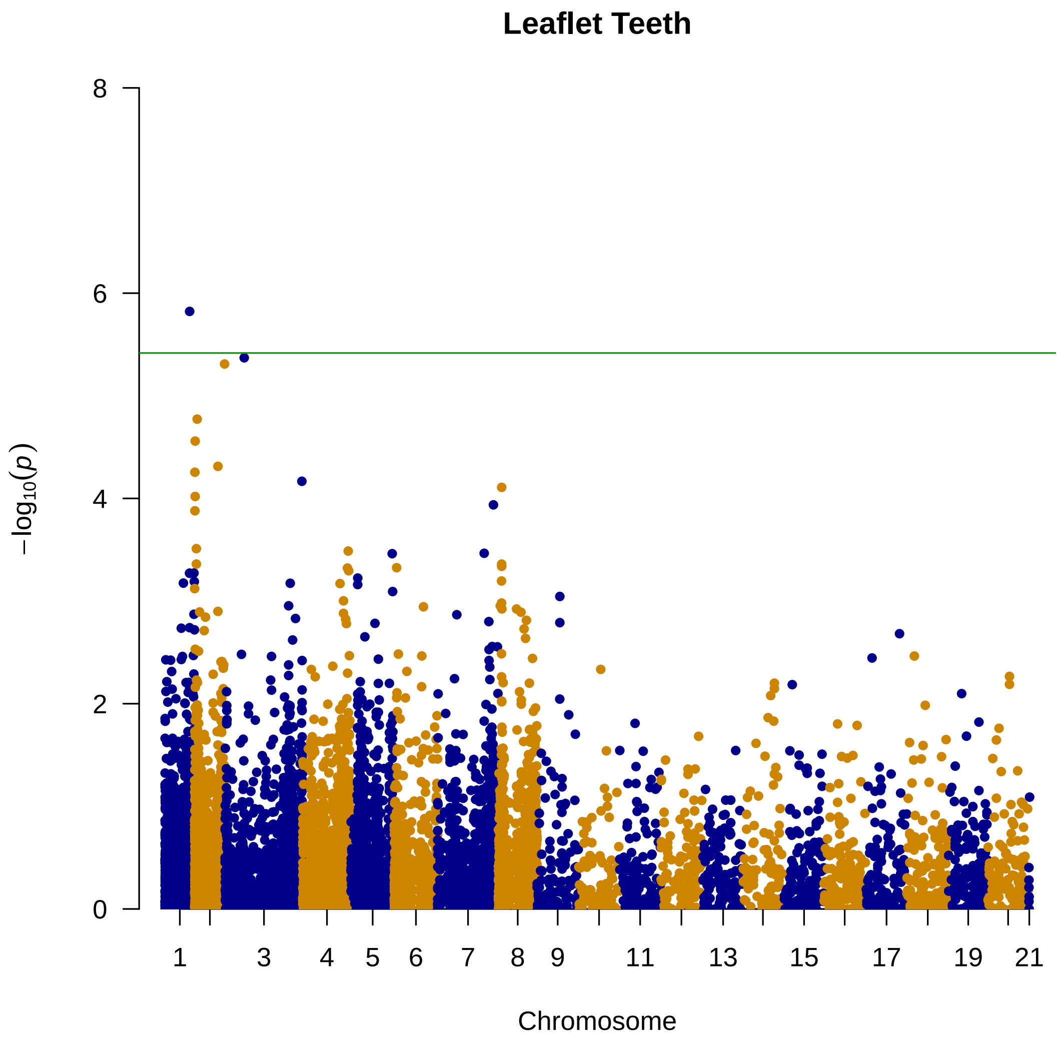


FIGURE 4.35: **Genome-wide Association Study for Leaflet Teeth.** Manhattan plot for leaflet teeth of the core collection grown in the glasshouse. Original in colour.

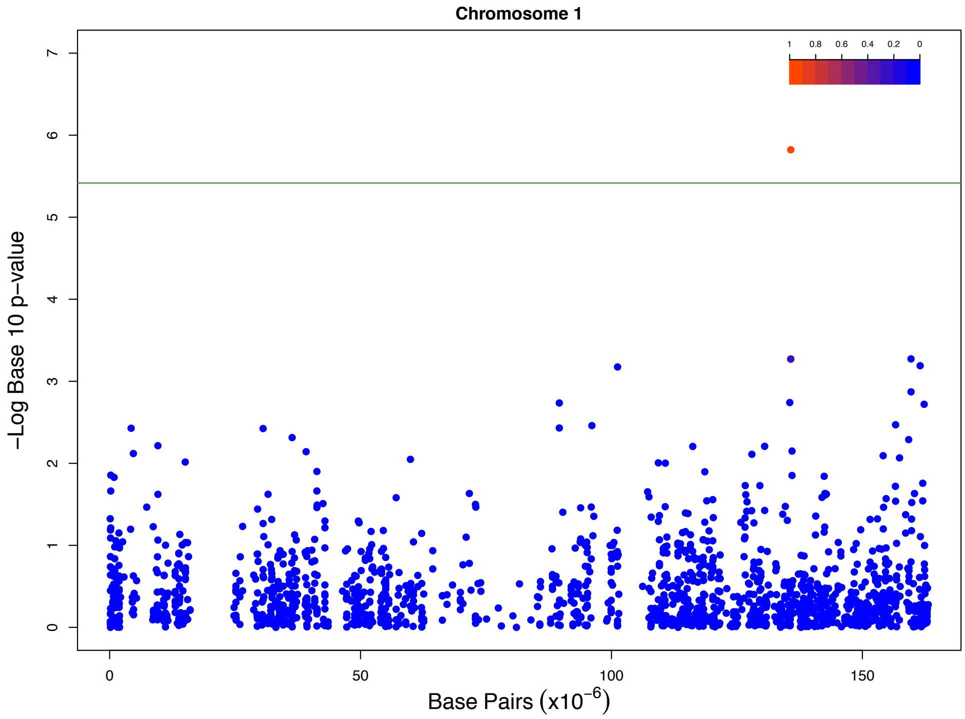


FIGURE 4.36: **Genome-wide Association Study for Leaflet Teeth.** Manhattan Plot of Chromosome 1 for Leaflet Teeth. Original in colour.

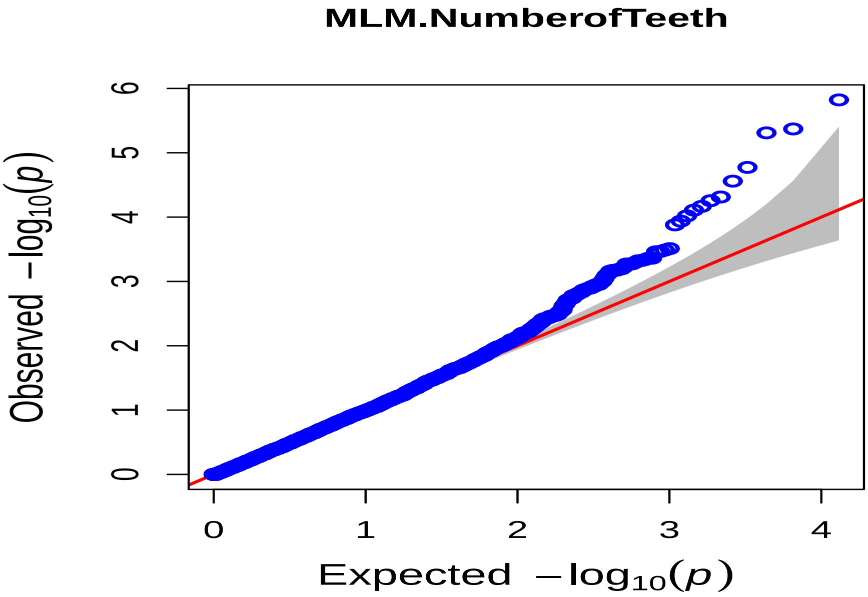
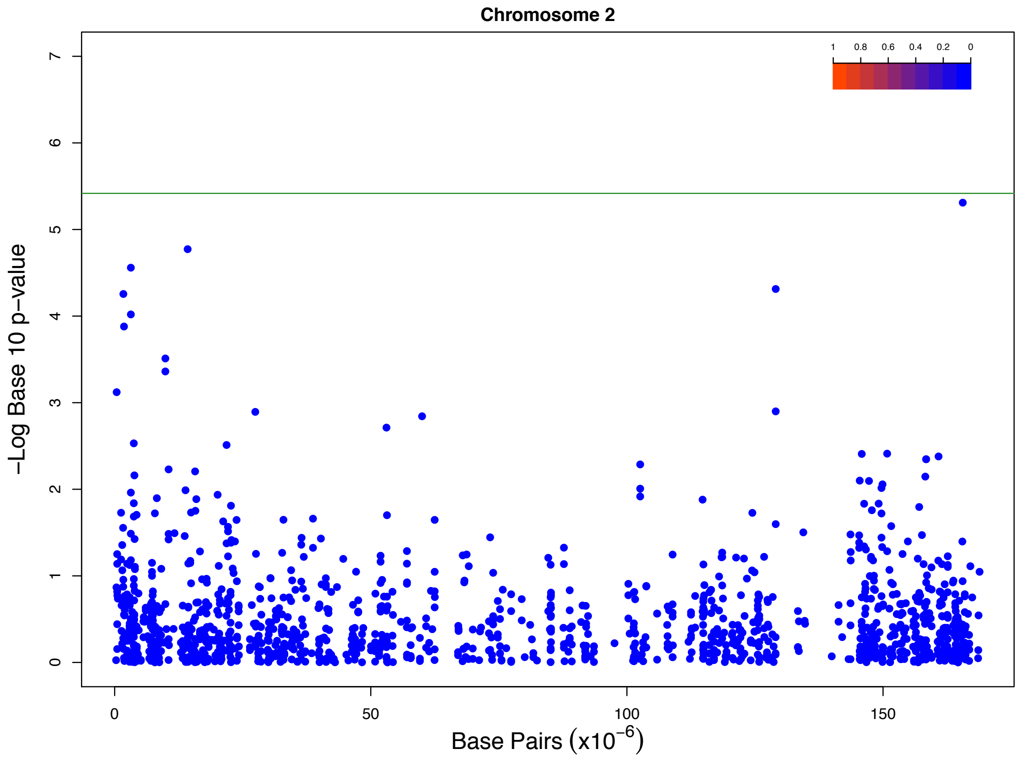


FIGURE 4.37: **Genome-wide Association Study for Leaflet Teeth.** Quantile-Quantile plot for leaflet teeth of the core collection grown in the glasshouse. Original in colour.



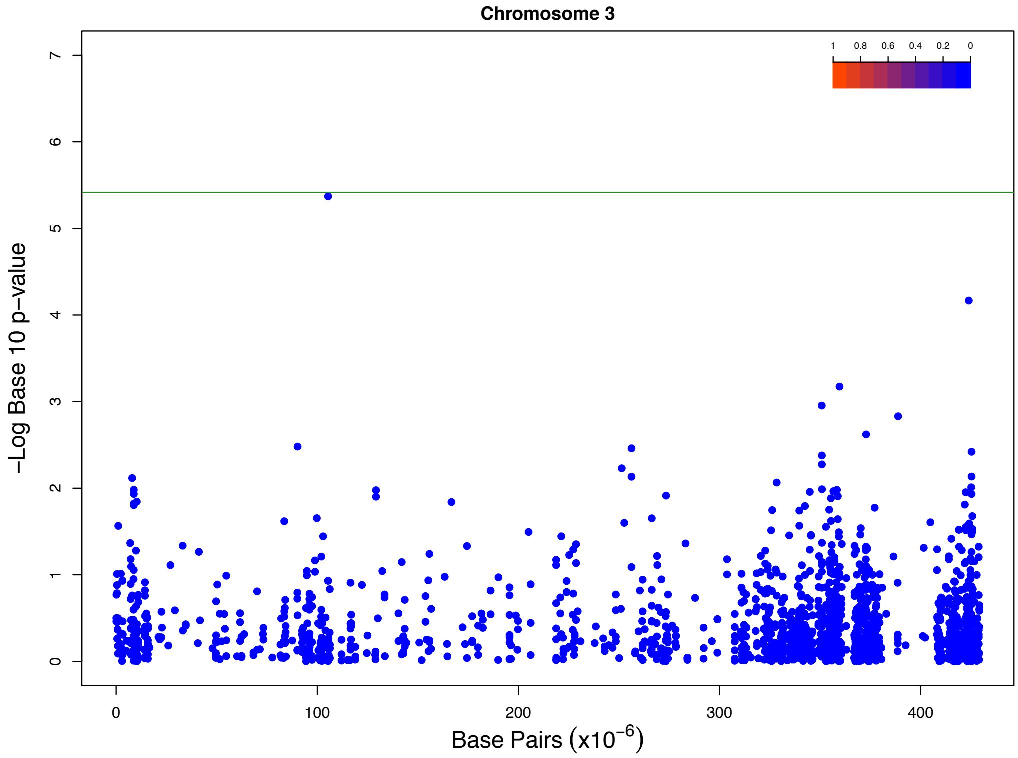
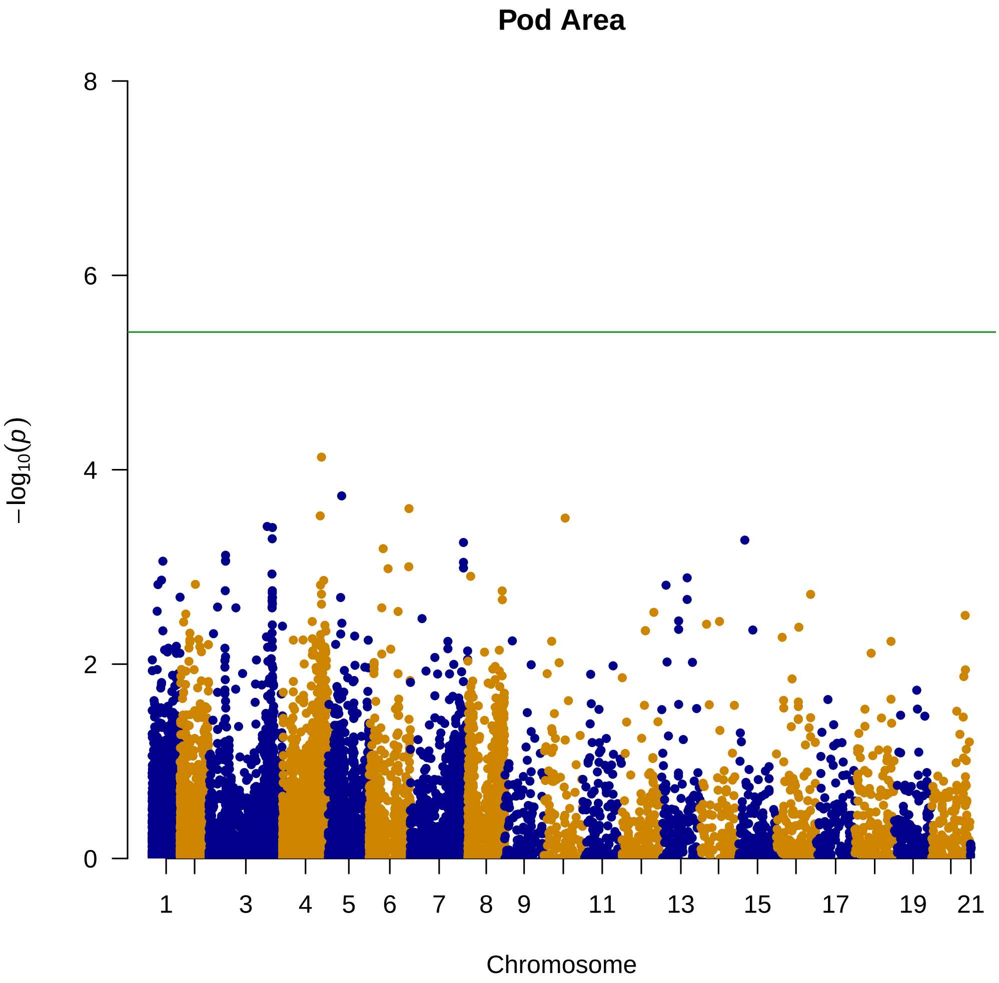


FIGURE 4.38: **Manhattan plot of Chromosome 2 and 3 from the Genome-wide Association Study for Leaflet Teeth.** The Manhattan plot displays the statistically significant threshold as a green line and there are no SNPs above this statistically significant threshold for these two chromosomes. Original in colour.

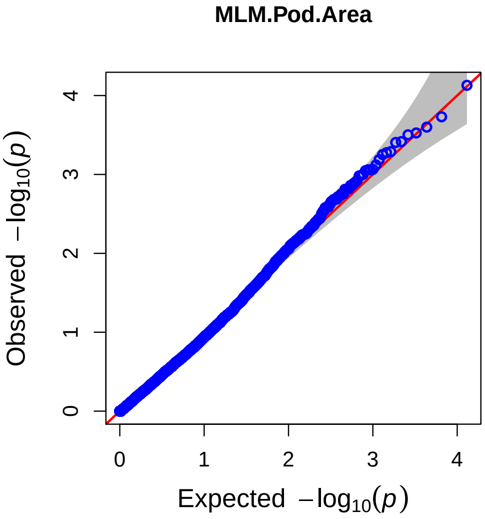
Pod

Genome-wide association studies were conducted for pod phenotypes. Manhattan plots were drawn for each pea pseudo-chromosome and the probability of each SNP being associated with the phenotype alongside Q-Q plots for the probability of each SNP association for the phenotypes for pod area, length, width and perimeter (Figure 4.39, 4.40, 4.41, 4.42 respectively). For each of these phenotypes, no statistically significant associations were found. For pod area, the highest peak is seen in in pseudochromosome 4. For pod length, the highest peak was seen in pseudochromosome 13. For pod width, the highest peak was seen in pseudochromosome 4. For pod perimeter, the highest peaks were observed in pseudochromosomes 1, 4 and 13.

Q-Q plots show some distension away from the curve in the right-hand region. Most of the SNPs will fall along the reference line due to not being statistically significant. Deviations away from this line are expected in the right-hand region due to being associated SNPs. However, since they do not lie above the statistically significant threshold, this may indicate spurious associations.

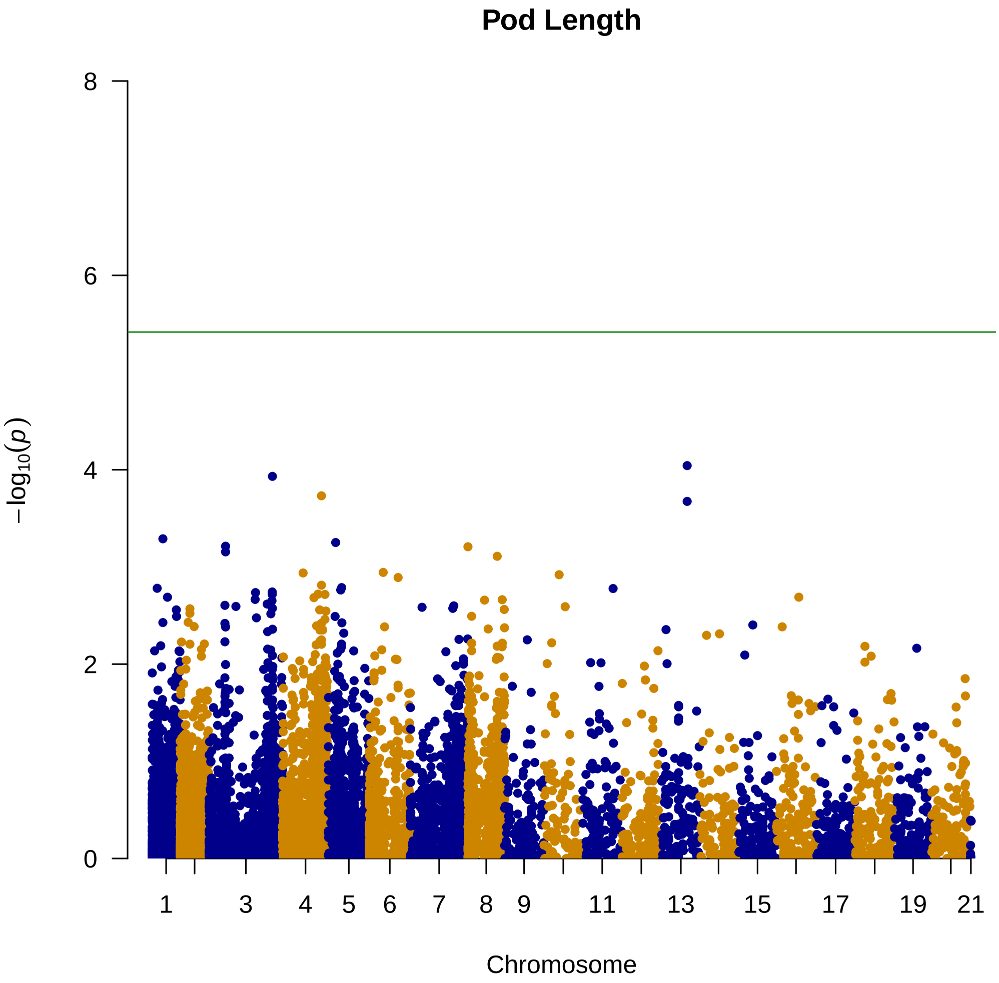


(A) Manhattan Plot

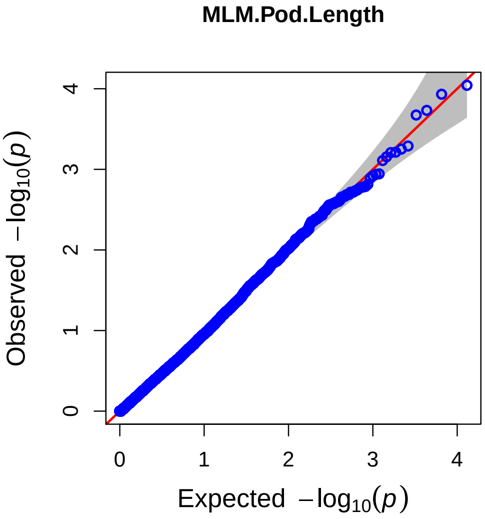


(B) Q-Q Plot

FIGURE 4.39: **Genome-wide Association Study for Pod Area.** The Manhattan plot (A) and Q-Q plots (B) for pod phenotypes of the core collection grown in the glasshouse. Original in colour.

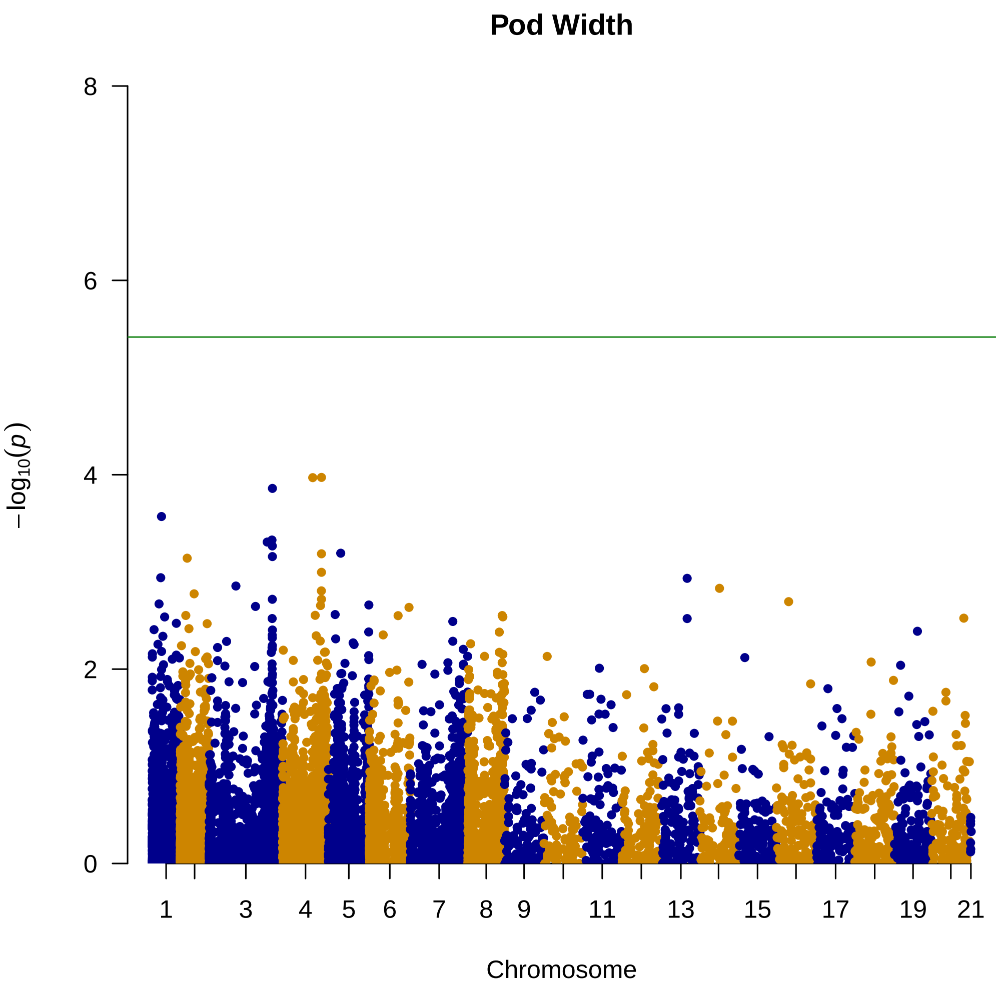


(A) Manhattan Plot

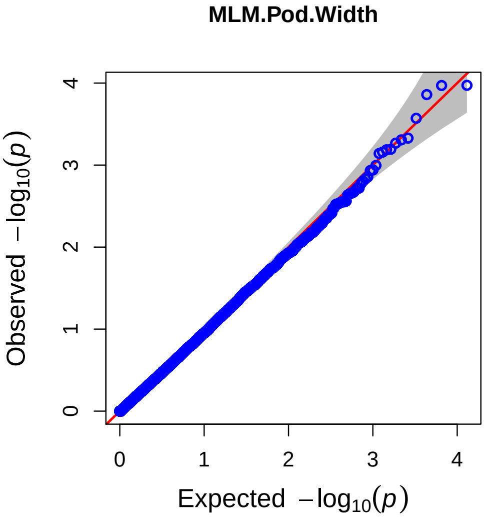


(B) Q-Q Plot

FIGURE 4.40: **Genome-wide Association Study for Pod Length.** The Manhattan plot (A) and Q-Q plots (B) for pod phenotypes of the core collection grown in the glasshouse. Original in colour.

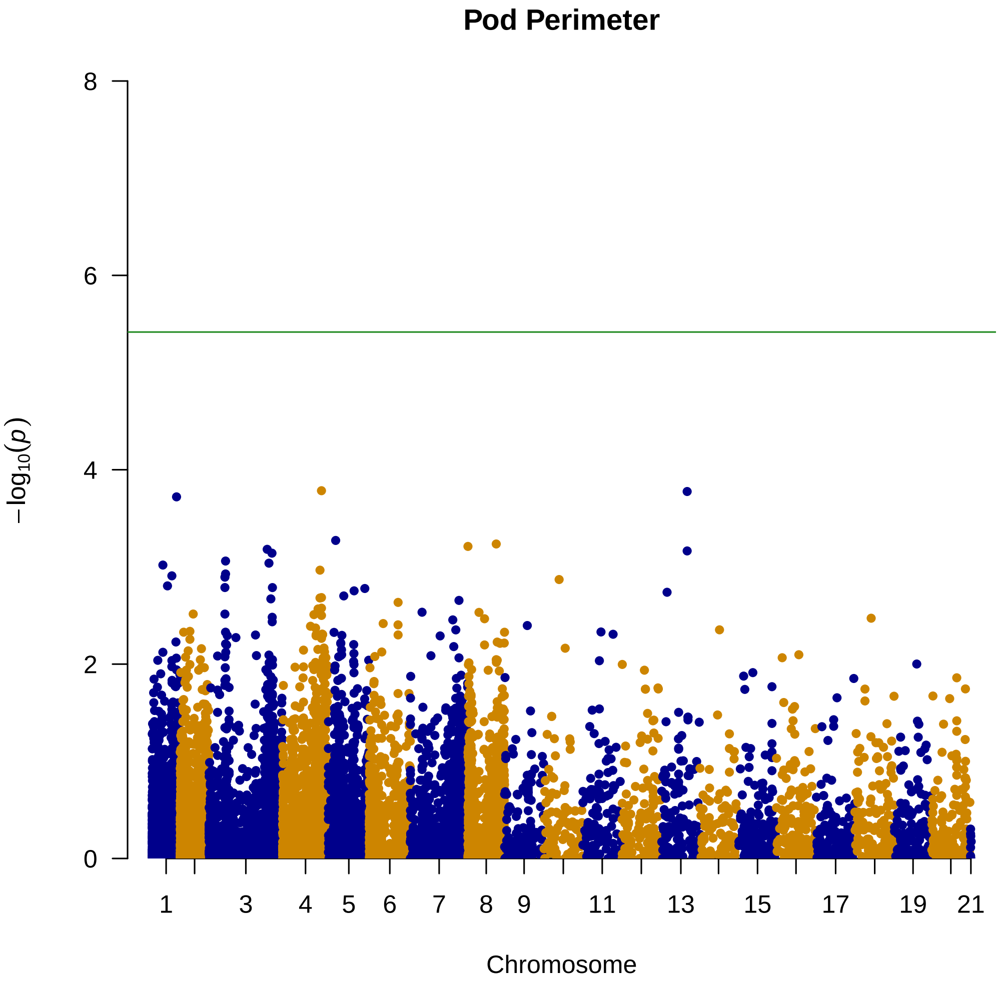


(A) Manhattan Plot

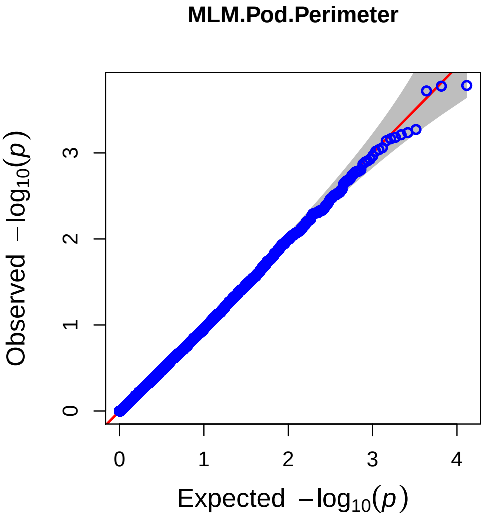


(B) Q-Q Plot

FIGURE 4.41: **Genome-wide Association Study for Pod Width.** The Manhattan plot (A) and Q-Q plots (B) for pod phenotypes of the core collection grown in the glasshouse. Original in colour.



(A) Manhattan Plot



(B) Q-Q Plot

FIGURE 4.42: **Genome-wide Association Study for Pod Perimeter.** The Manhattan plot (A) and Q-Q plots (B) for pod phenotypes of the core collection grown in the glasshouse. Original in colour.

4.3.6 Statistically associated loci

For the statistically associated loci found in leaflet teeth, seed eccentricity and seed perimeter the nearest flanking markers were taken and locally aligned against *Medicago truncatula*. Table 4.6 shows the flanking regions obtained for each phenotype.

TABLE 4.6: **Flanking regions of statistically associated loci.** Table shows the organ and phenotype, the chromosome the statistically associated loci is found on, the positions of the left flanking marker, the SNP, and the right flanking marker and length of the total flanking region (bp)

Organ	Phenotype	Chr	Left Flank	SNP	Right Flank	Total Flank
Leaflet	Teeth	1	135,727,814	135,729,915	135,981,694	253,880
Seed	Eccentricity	8	132,523,741	133,948,908	133,949,171	1,425,430
Seed	Eccentricity	3	64,321,570	68,285,374	68,285,586	3,964,016
Seed	Perimeter	7	329,112,507	329,113,583	329,136,382	23,875

Table 4.7 refers to the orthologs found in *Medicago truncatula* and *Arabidopsis thaliana* for the phenotype leaflet teeth. These results show that a common *A.thaliana* ortholog is found, AT5G60390.1, which is annotated as a GTP binding Elongation Factor Tu protein. This ortholog is described as being involved in translation and functions in calmodulin binding.

Other *A.thaliana* orthologs found in Table 4.7 are annotated as the gene POWERDRESS which has DNA binding transcription factor activity, a gene encoding a pseudouridine synthase family protein which has RNA binding properties and pseudouridine synthase activity and a S-adenosyl-L-methionine transporter-like (SAMTL) gene which functions in calcium binding and S-adenosyl-L-methionine transmembrane transporter activity (AT3G52250.1, AT3G52260.3 and AT2G35800.1 respectively).

Furthermore, there were two *M.truncatula* orthologs which were annotated as hypothetical proteins, one of which (Medtr1g095600) held an *A.thaliana* ortholog annotated as a gene that encodes for a defence response to bacterium. The other *M.truncatula* ortholog, Medtr1g048480, held no ortholog in *A.thaliana*.

These results indicate that the orthologs found in the flanking region for leaf teeth are involved in RNA binding, RNA splicing and translation elongation.

TABLE 4.7: **Orthologs aligning in leaflet teeth flanking region.** *Medicago truncatula* (Mt) Gene Name and Annotation, corresponding top hit for *Arabidopsis thaliana* (At) Gene Ortholog and Annotation, *Arabidopsis thaliana* (At) Gene Ontology terms for Biological Process (BP) and Molecular Function (MF).

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr1g013680, Medtr1g024175, Medtr1g085410, Medtr1g095650 *GTP-binding elongation factor Tu family protein*	**AT5G60390.1** *GTP binding Elongation factor Tu family protein*	**BP** : involved in translation, translational elongation **MF** : functions in calmodulin binding, has GTPase activity translation elongation factor activity
Medtr1g048480 *hypothetical protein*	N/A	**BP** : N/A **MF** : N/A
Medtr1g095570 *Myb DNA-binding domain protein*	AT3G52250.1 POWERDRESS PWR	**BP** :involved in RNA splicing, floral meristem determinacy, histone H3-K9 deacetylation, negative regulation of transcription by RNA polymerase II, positive regulation of histone H3-K9 acetylation, regulation of leaf senescence **MF** :functions in DNA binding protein binding has DNA-binding transcription factor activity

Continued on next page

Table 4.7 – *Continued from previous page*

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr1g095580 *pseudouridine synthase Rlu family protein*	AT3G52260.3 Pseudouridine synthase family protein	**BP** :involved in pseudouridine synthesis **MF** :has RNA binding pseudouridine synthase activity
Medtr1g095600 *hypothetical protein*	AT4G25030.1 ATNHR2B NON HOST RESISTANCE 2B	**BP** :defence response to bacterium, incompatible interaction, phosphorylation **MF** :functions in protein binding has kinase activity
Medtr1g095780 *substrate carrier family protein*	AT2G35800.1 S-ADENOSYL METHIONINE TRANSPORTER-LIKE. SAMTL	**BP** :N/A **MF** :functions in calcium ion binding has S-adenosyl-L-methionine transmembrane transporter activity

Continued on next page

Table 4.7 – *Continued from previous page*

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr1g101870	AT5G60390.1	
elongation factor 1-alpha, Medtr1g495705 GTP-binding elongation factor Tu family protein, Medtr1g095703 GTP-binding elongation factor Tu family protein	EF1ALPHA, GTP binding Elongation factor Tu family protein	**BP** :involved in translation, translational elongation **MF** :functions in calmodulin binding has GTPase activity, translation elongation factor activity

Table 4.8 refers to the orthologs found in *Medicago truncatula* and *Arabidopsis thaliana* for the seed eccentricity flanking region found in pseudo-chromosome 8. The results indicate that most of the orthologs are involved in metabolic processes. This includes, a hypothetical *Medicago truncatula* gene (Medtr8g047050) and its *Arabidopsis thaliana* ortholog (AT4G26270.1) is annotated as Phosphofructokinase 3 (PFK3) with GO terms that indicate it is involved in metabolic biology, as well as ATP and metal binding (Table 4.8).

Other genes include, Medtr8g055930, which has an *Arabidopsis thaliana* ortholog annotation as a transmembrane protein (DUF616) and Medtr8g056030 which is annotated as cullin 3B with its *Arabidopsis thaliana* ortholog annotated as involved in light response and catabolic processes.

Medtr8g055840 has an *Arabidopsis thaliana* ortholog which is annotated as an S-adenosyl-L-methionine-dependent methyltransferase and involved in methylation (AT1G26850.1). Another gene, Medtr8g055860, is putatively annotated as a phytochrome kinase substrate protein and the *Arabidopsis thaliana* ortholog annotation is also phytochrome kinase substrate

1 (PKS1) which is involved in phototropism, gravitropism and protein binding. Another gene, Medtr8g056100, and its ortholog AT2G46210.1, is involved in response to cold but also in oxidation-reduction processes.

Some *Medicago truncatula* orthologs have been annotated in other functions, there are two genes (Medtr8g467580 and Medtr8g467590) involved in responses to a plant hormone named cytokinin. Their *Arabidopsis thaliana* orthologs are annotated as HXXXD-type acyl-transferases (AT3G26040.1 and AT3G26040.1). The remaining ortholog, Medtr8g467560, is annotated as centromere C-like protein and involved in cell division.

TABLE 4.8: **Orthologs aligning in Seed Eccentricity flanking region on Chromosome 8.** *Medicago truncatula* (Mt) Gene Name and Annotation, corresponding top hit for *Arabidopsis thaliana* (At) Gene Ortholog and Annotation, *Arabidopsis thaliana* (At) Gene Ontology terms for Biological Process (BP) and Molecular Function (MF).

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr8g047050 *hypothetical protein*	**AT4G26270.1** PFK3 PHOSPHO-FRUCTOKINASE 3	**BP** : involved in fructose 6-phosphate metabolic process; glycolytic process; root epidermal cell differentiation **MF** :has 6-phosphofructokinase activity; ATP binding; metal ion binding
Medtr8g055840 methyltransferase PMT2-like protein	**AT1G26850.1** S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	**BP** :involved in methylation **MF** :has methyltransferase activity

Continued on next page

Table 4.8 – *Continued from previous page*

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr8g055860 *phytochrome kinase substrate protein; putative*	**AT2G02950.1** PHYTOCHROME KINASE SUBSTRATE 1; PKS1	**BP** :involved in phototropism; positive gravitropism; red or far-red light signalling pathway; red; far-red light phototransduction; response to far red light; response to red light **MF** :functions in protein binding
Medtr8g055930 *transmembrane protein*	**AT2G02910.1** transmembrane protein (DUF616)	**BP** :involved in ceramide metabolic process **MF** :has hydrolase activity; acting on carbon-nitrogen (but not peptide) bonds; in linear amides; transferase activity; transferring glycosyl groups
Medtr8g056030 *cullin 3B*	**AT1G26830.1** ATCUL3; ATCUL3A; CUL3; CUL3A; CULLIN 3; CULLIN 3A	**BP** :involved in SCF-dependent proteasomal ubiquitin-dependent protein catabolic process; embryo development ending in seed dormancy; endosperm development; positive regulation of flower development; proteasome-mediated ubiquitin-dependent protein catabolic process; response to red or far red light; ubiquitin-dependent protein catabolic process **MF** :functions in protein binding, has protein binding; ubiquitin protein ligase activity; ubiquitin protein ligase binding; ubiquitin-protein transferase activity

Continued on next page

Table 4.8 – *Continued from previous page*

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr8g056100 *fatty acid/sphingolipid desaturase*	**AT2G46210.1** ATSLD2; SLD2; SPHINGOID LCB DESATURASE 2; Fatty acid/sphingolipid desaturase	**BP** :involved in cellular response to cold; oxidation-reduction process; sphingolipid biosynthetic process **MF** :has metal ion binding; oxidoreductase activity; sphingolipid delta-8 desaturase activity
Medtr8g467490 *NAC domain class transcription factor*	**AT1G26870.1** ANAC009; ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 9; FEZ;NAC-domain protein.	**BP** :involved in multicellular organism development; positive regulation of asymmetric cell division; regulation of transcription; DNA-templated; response to auxin; root cap development; somatic stem cell division **MF** :has DNA binding; DNA-binding transcription factor activity
Medtr8g467560 *centromere C-like protein*	**AT1G15660.1** CENP-C; CENP-C HOMOLOGUE; CENTROMERE PROTEIN C;	**BP** :involved in attachment of mitotic spindle microtubules to kinetochore; attachment of spindle microtubules to kinetochore involved in homologous chromosome segregation; cell division; kinetochore assembly **MF** :has centromeric DNA binding

Continued on next page

Table 4.8 – *Continued from previous page*

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr8g467580 *HXXXD-type acyltransferase family protein*	**AT3G26040.1** HXXXD-type acyltransferase family protein	**BP** :involved in response to cytokinin **MF** :has transferase activity
Medtr8g467590 *alcohol acyltransferase*	**AT3G26040.1** HXXXD-type acyltransferase family protein	**BP** :involved in response to cytokinin **MF** :has transferase activity
Medtr8g468030 *galactose oxidase/kelch repeat protein*	**AT2G02870.1** Galactose oxidase/kelch repeat superfamily protein	**BP** :involved in protein ubiquitination

Table 4.9 refers to the orthologs found in *M.truncatula* and *A.thaliana* from the locus found in chromosome 3 that associates with seed eccentricity. The results indicate that these genes are involved in defence against infection, transcription, nucleic acid binding, DNA repair and oligopeptide transport. Some genes are annotated as hypothetical in *M.truncatula* but have orthologs with functional annotation in *A.thaliana* such as Medtr3g069337, Medtr3g033235, Medtr3g448950, Medtr3g036640, Medtr3g062298, Medtr3g089640, and Medtr3g047635. One gene, Medtr3g021030, is hypothetical in both *M.truncatula* and *A.thaliana*.

TABLE 4.9: **Orthologs aligning in Seed Eccentricity flanking region on Chromosome 3.** *Medicago truncatula* (Mt) Gene Name and Annotation, corresponding top hit for *Arabidopsis thaliana* (At) Gene Ortholog and Annotation, *Arabidopsis thaliana* (At) Gene Ontology terms for Biological Process (BP) and Molecular Function (MF).

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr3g069500 *peptide/nitrate transporter plant*	AT4G23160.1 CRK8; CYSTEINE-RICH RLK (RECEPTOR-LIKE PROTEIN KINASE) 8	**BP** : involved in defence response to bacterium; protein phosphorylation **MF** : has ATP binding; kinase activity; protein kinase activity; protein serine/threonine kinase activity
Medtr3g022830 *GRAS family transcription factor*	AT1G50420.1 SCARECROW-LIKE 3(SCL-3)	**BP** : involved in regulation of transcription; DNA-templated; response to gibberellin **MF** : has DNA-binding transcription factor activity; protein binding; sequence-specific DNA binding; transcription coregulator activity
Medtr3g069337 *hypothetical protein*	AT2G24830.1zinc finger (CCCH-type) family protein / D111/G-patch domain-containing protein;	**BP** : N/A **MF** : functions in nucleic acid binding
Medtr3g021030 *hypothetical protein*	- -	**BF** : N/A **MF** : NA

Continued on next page

Table 4.9 – *Continued from previous page*

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr3g022850 *crossover junction endonuclease MUS81-like protein*	**AT5G39770.1** pseudogene homologous to AtMSU81	**BP** : involved in DNA repair; double-strand break repair via break-induced replication; intra-S DNA damage checkpoint; resolution of meiotic recombination intermediates **MF** : functions in nucleic acid binding has 3′-flap endonuclease activity;crossover junction endodeoxyribonuclease activity;endonuclease activity
Medtr3g033235, Medtr3g448950 *hypothetical protein*	**AT4G26590.1** ARABIDOPSIS THALIANA OLIGOPEPTIDE TRANSPORTER 5	**BP** : involved in oligopeptide transport; protein transport **MF** : has oligopeptide transmembrane transporter activity
Medtr3g036640 *hypothetical protein*	**AT2G16810.1** F-box and associated interaction domains-containing protein	**BF** : involved in ubiquitin-dependent protein catabolic process **MF** : has ubiquitin-protein transferase activity
Medtr3g062298 *hypothetical protein*	**AT2G42760.1** DUF1685 family protein	**BF** : N/A **MF** : N/A

Continued on next page

Table 4.9 – *Continued from previous page*

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr3g089640 *hypothetical protein*	**AT4G17140.3** pleckstrin homology (PH) domain-containing protein	**BF** : N/A **MF** : N/A
Medtr3g047635 *hypothetical protein*	**AT3G14172.1** GPI-anchored adhesin-like protein	**BF** : N/A **MF** : N/A

Table 4.10 refers to the orthologs found in *Medicago truncatula* and *Arabidopsis thaliana* for the phenotype seed perimeter. There are 3 genes identified with the seed perimeter phenotype, involved in cell division, N-glycan fucosylation and the third has not been annotated with a GO function. One of these genes, Medtr7g115550 is annotated as a putative cell division control protein and its *Arabidopsis thaliana* ortholog holds evidence for this role. Another gene, Medtr7g115570, is annotated as a coatomer protein but holds an annotation in the *Arabidopsis thaliana* ortholog as an alpha-(1,6)-fucosyltransferase. The third gene, Medtr7g115580, is annotated as an ankyrin repeat 13B-like protein and holds a similar annotation in its *Arabidopsis thaliana* ortholog.

TABLE 4.10: **Orthologs aligning in Seed Perimeter flanking region.** *Medicago truncatula* (Mt) Gene Name and Annotation, corresponding top hit for *Arabidopsis thaliana* (At) Gene Ortholog and Annotation, *Arabidopsis thaliana* (At) Gene Ontology terms for Biological Process (BP) and Molecular Function (MF).

Mt Gene Name and Annotation	At Gene Ortholog (TOP HIT) and Annotation	At GO terms
Medtr7g115550 cell division control protein; putative	**AT1G09770.1** ARABIDOPSIS THALIANA CELL DIVISION CYCLE 5	**BP** :involved in DNA repair; RNA splicing; cell cycle; defense response signaling pathway; resistance gene-dependent; defense response signaling pathway; resistance gene-independent; defense response to bacterium; defense response to fungus; innate immune response; mRNA processing; regulation of transcription; DNA-templated **MF** :functions in DNA binding, has DNA binding; DNA-binding transcription factor activity; protein binding
Medtr7g115570 coatomer protein	**AT5G28910.2** alpha-(1,6)-fucosyltransferase	**BP** :involved in N-glycan fucosylation; protein N-linked glycosylation **MF** :has alpha-(1,6)-fucosyltransferase activity; transferase activity; transferring glycosyl groups
Medtr7g115580 ankyrin repeat 13B-like protein	**AT3G04470.1** Ankyrin repeat family protein	**BP** : N/A **MF** : N/A

4.4 Discussion

The study above performed a GWAS in *Pisum sativum* using the first publicly available pea genome reference, using Simple and Morphological Shape Descriptors (SMSDs) and manual phenotypes on a core collection of the John Innes Collection that is largely statistically representative based on seed weight across wild, landrace and cultivated varieties. This GWAS identified statistically associated genomic loci for leaflet teeth, seed perimeter and seed eccentricity phenotypes. This is potentially useful in the context of pea domestication. Wild leaves are known to be serrated, seeds are known to be larger in size in domesticated varieties and wrinkled seeds are known to be favoured in cultivated varieties due to their sweeter taste (Holdsworth et al., 2017; Rayner et al., 2017).

4.4.1 Major Findings

Leaflet teeth

In leaflet teeth, a statistically associated SNP was found and a flanking region was taken and aligned to *M.truncatula*. The flanking region aligned with several orthologs encoding for Elongation Factor 1-alpha, a ubiquitously expressed protein involved in translation (Durso and Cyr, 1994). This does not necessarily suggest that that Elongation Factor 1-alpha is the causative gene but may be an indirect association, in other words, in linkage disequilibrium with the causal variant involved in the production of leaflet teeth (Bush and Moore, 2012).

Another ortholog, pseudouridine synthase, is also involved in RNA binding. However, a study (Tran et al., 2004) has identified this gene as not invoking drought response, since it does not hold a NAC recognition site which interacts with the Arabidopsis gene EARLY RESPONSIVE TO DEHYDRATION STRESS 1 (ERD1) and therefore does not provide a water stress response. In combination with another study which identified a low presence of teeth in wet climates (Peppe et al., 2011), it may be possible to suggest that the gene, pseudouridine synthase, may be expressed in wetter climates. The two outcomes of a GWAS are either direct or indirect association (Bush and Moore, 2012) and it is important to note that pseudouridine synthase may not be a causal gene but instead, could be in LD with the causative variant.

Another gene, POWERDRESS (PWR), was also identified in the flanking region. This gene is known to regulate leaf size by increasing cell proliferation and mutations in the POWERDRESS gene causes small leaves to be observed (Suzuki et al., 2018). The SNP identified may be a direct or an indirect association (Bush and Moore, 2012) and further work needs to be performed to determine which type of association it is, nevertheless, if the SNP was identified as a causal variant, POWERDRESS could be a possible candidate gene of interest.

In section 1.3, we previously explored how the Serrated *Ser* gene is known to cause a serrated leaf due to the presence of leaf teeth (Weeden and Ambrose, 2004). The *Ser* gene was not identified in this GWAS peak and its flanking region, however, this does not rule out the possibility that this SNP is some distance away from the *Ser* gene.

Seed eccentricity

In seed eccentricity, two statistically significant loci were found. The locus with the most significant p-value was in Chromosome 8 and the less significant p-value was in Chromosome 3.

For the most significant locus, a flanking region was taken and aligned to *M.truncatula* and this corresponded to the orthologous gene Medtr8g047050, annotated as encoding for a hypothetical protein. This gene was then aligned against *A.thaliana* and the resultant gene, AT4G26270.1, is annotated as **PHOSPHOFRUCTOKINASE 3 (PFK3)** and its Gene Ontology for Biological Process was listed as involved in fructose 6-phosphate metabolic process.

PFK3 breaks down fructose 6-phospate (a molecule found in the pea carbohydrate metabolism pathway) into a smaller substrate (Mustroph, Sonnewald, and Biemelt, 2007). Fructose 6-phospate is a key precursor to starch production (Casey et al., 1998) and PFK3 may affect its bioavailability. This change in the pea carbohydrate metabolism pathway has potential to change starch structure with downstream effects on pea wrinkledness (Rayner et al., 2017). In this thesis, seed eccentricity is used as a metric of seed roundness. The SNP has not been identified as a direct or indirect association. If it were to be identified as an indirect association this SNP would be in linkage disequilibrium with the casual variant but if this SNP were to be identified as a causal variant, PFK3 could be a gene of potential interest to seed wrinkledness.

For the second most significant loci for seed eccentricity, a flanking region was taken and aligned to *M.truncatula* and *A.thaliana*. The *M.truncatula* and *A.thaliana* orthologs are annotated as involved in root nodulation (Table 4.9). For example, Medtr3g069500, is a peptide/nitrate transporter and its *A.thaliana* ortholog's gene ontology suggests it is involved in plant defence (Wrzaczek et al., 2010) and the gene Medtr3g022830, is a GRAS transcription factor and some forms of GRAS transcription factors are involved in the nodulation of legumes (Kaló et al., 2005; Smit et al., 2005). One gene in *M.truncatula*, Medtr3g047635, is annotated as encoding for a hypothetical protein, but its *A.thaliana* ortholog is annotated as a GPI-anchored adhesin-like protein and these maybe involved in symbiosis (Brewin, 2004). Another gene in *M.truncatula*, Medtr3g069337, annotated as hypothetical but its *A.thaliana* ortholog is involved nucleic acid binding and Medtr3g022850 is involved in DNA repair, according to Table 4.9. Infection of the pea root by the bacteria *Rhizobium leguminosarum biovar viceae* is explored in Chapter 1. The orthologous genes for root nodulation found in the flanking region associated with seed eccentricity are may be due to the SNP being some distance away from the causative gene itself as a result of indirect association (Bush and Moore, 2012).

Section 1.3 explored key genes that affect seed wrinkledness. Section 1.3 explored the key genes of pea carbohydrate metabolism, *r* (Bhattacharyya et al., 1990; Rayner et al., 2017), *rb* (Hylton and Smith, 1992), *rug-3* (Harrison et al., 2000), *rug-4* (Casey et al., 1998), *rug-5* (Craig et al., 1998) and *lam* (Bogracheva et al., 1999). These genes change the structure of starch. Pea seeds with lower starch (higher sugar concentration) will become wrinkled and those with higher starch (lower sugar concentration) will become round. All of the above genes cause wrinkledness in pea and are also sweeter in taste due to a higher sugar concentration. The *r* gene is also thought to not only cause wrinkledness but to also reduce seed size (Weeden, 2007). In this thesis, we have used seed eccentricity as one way to measure how round or wrinkled a seed would be. The finding of the orthologous gene, PFK3, in the flanking region of the GWAS peak is interesting since it may breakdown a key precursor in the pea carbohydrate metabolism pathway, nevertheless, it is important to note that the SNP found to be statistically associated with this trait may be in linkage disequilibrium and some distance away from the causative gene. However, the other finding of orthologous genes for root nodulation was unexpected.

Seed perimeter

For the loci found for seed perimeter, a flanking region was taken and aligned to *M.truncatula* and *A.thaliana* and orthologs were found. One gene, Medtr7g115570, is annotated to be a coatomer protein, these proteins are present in the peribacteroid membrane of bacteria infected nodules (Larrainzar and Wienkoop, 2017). Another gene, Medtr7g115550 is thought to be a putative cell division protein, and its *A.thaliana* orthologs is annotated as encoding for a protein CELL DIVISION CYCLE 5, involved in the cell cycle. Another gene, Medtr7g115580, is annotated, as an ankyrin repeat protein, proteins which contain this protein are known to aid symbiosis in other legumes (Kumagai et al., 2007). Orthologous genes responsible for root symbiosis, coatomer protein, and cell division are found in this flanking region. It is unclear why these genes are associated with seed perimeter but these genes may be unlikely to be causative since the SNP could be some distance away from the causative variant itself due to linkage disequilibrium (Bush and Moore, 2012).

In this thesis, we also used seed perimeter to assist with helping determine larger or more wrinkled seeds. In Section 1.3 we explored genes that affect seed size. Key QTLs such as I, IV, and VII (Weeden, 2007) were not found in the flanking region of the GWAS peak and pleiotropic genes that also have a secondary effect on seed size such as *Np* and *r* (Weeden, 2007) were not found in this region either. It is important to note that whilst these genes were not found in the flanking region, it does not mean these genes are likely to be the causative gene since the statistically associated SNP may be in high linkage disequilibrium with the functional SNP and a large genetic distance away from the causative gene (Bush and Moore, 2012).

Other phenotypes

The study did not find significantly associated loci for plant height or seed weight which are known to be domesticated. This may not be surprising as these traits are likely to be polygenic (Boyle, Li, and Pritchard, 2017; Nawab et al., 2008), meaning that there are several genes which could contribute to this phenotype, hence, this might be a cause of the absence of a statistically significantly associated SNP. Furthermore, no statistically significant associated SNPs were found for seed area, seed equivalent diameter, leaflet area, leaflet perimeter, leaflet length, leaflet width, pod length, pod width, pod area or pod perimeter. The reasons for this lack of statistically significant SNPs for these traits could include penetrance (the size of the effect of the SNP), number of samples obtained and number of SNPs obtained (Bush and Moore, 2012). An additional reason for the lack statistically significant SNPs could include limitations in the experimental design mentioned in Section 3.4.4.

If significantly associated loci were to be found one would have expected the flanking region for plant height to perhaps correspond with the *Le* gene (Sherriff et al., 1994), a QTL near *GA2bOH* gene on linkage group IV or the QTLs *ht1,2* and *3* (Weeden, 2007) or the following markers: *cttg7*, *caag4* and *cagg5* (Tar'an et al., 2003). For the leaflet size SMSDs, one would have expected the flanking region of any associated loci to have corresponded with the *afila* gene (Lafond, Evans, and Ali-Khan, 1981) and for pod SMSDs we would have expected to see correspondence with the *nn* (Wehner and Gritton, 1981). However, again, it is important to reiterate that the statistically associated SNP could be an indirectly associated SNP through LD and therefore a large distance away from the causative gene (Bush and Moore, 2012).

4.4.2 Assessment of sequencing data

The GBS reads found pea had a GC content of 47-48%, whereas literature states the GC content of pea as 37.7% (Smýkal et al., 2012). This difference may occur from the nature of the GBS protocol enriching for a particular motif as well as the GC biases in Illumina data. In contrast, the whole genome shotgun sequencing provided a GC content 36-37%.

The *de novo* reference assembly created in this chapter proffered a genome with a N50 metric of 12,597 and is over 6x larger in contiguity than INRA's reference which holds a N50 metric of 1,986. Hence, this genome is more contiguous than that of the International Consortium. This is unexpected, since the Consortium hold PacBio reads (known for their long reads), and Mate-Pair reads for this genome, however, perhaps this difference relates to the use of w2rap as an assembly tool for the reference created in this thesis. The genome produced in this chapter could be useful since it is more contiguous and could also be a vital resource, since it has been made publicly accessible. Furthermore, the community will find the choice of cultivar particularly beneficial due to its key phenotypes being of great interest to researchers, but also, that the cultivar is present in both the John Innes *Pisum* collection and the USDA collection, which improves accessibility to all. The N50 of the assembly generated may hold some effect on the ordering and orientation of genomic markers. The larger the N50 metric of an assembly is the less fragmented the assembly is. The benefit of a more contiguous assembly is the more accurate ordering of markers.

This project also found that higher MAF leads to a reduction in site number however, in some cases can sometimes increase depth. The results show high missing data for both MAFs perhaps, also in part, due to the nature of the GBS sequencing.

One aforementioned study (Holdsworth et al., 2017), performed GWAS on pea flower colour using all SNPs filtered with MAF 0.01 and missing data 20%, instead of those which fit these criteria and hold 5x coverage per SNP. This study observed 25 associated loci. Here, a potential argument is proposed: what confidence can be held in the data if coverage does not reach a minimum of 5x coverage per SNP? In this chapter, SNPs were called with a minimum of 5x per SNP. This argument brings up a well-known trade off: Stringency versus Leniency. More stringent approaches are conservative methods that do not always yield a high number of results, however the results that are found are likely to be more accurate. Conversely, lenient approaches are more liberal, allowing greater parameter flexibility and therefore more results, however confidence in the fidelity of these results can diminish. Therefore, it is best to select the most appropriate approach for the correct application in question. In GWAS, the tendency towards inflation of false positives (Lipka et al., 2012), may mean that a more stringent approach is more pertinent. Both approaches filtered SNPs using MAF 0.01 and missing data 20% however, the lack of coverage for SNPs in the Holdsworth paper does not account for sequencing error. Here, we argue that a 5x coverage in addition to SNP filters maybe a more suitable approach for analysis in pea Genome-wide Association Study.

Additionally, the difference on numbers of associated loci may differ between this study and the Holdsworth et al. (2017) study due to the type of phenotype. Flower colour can be a discrete phenotype, either white or pink, for example. Conversely, other phenotypes such as seed weight or plant height may be more continuous. Furthermore, the phenotype being polygenic or the penetrance of the gene effect may affect the number of statistically associated loci found in a GWAS.

Figure 4.15 displays unexpected clustering. This figure shows many landraces clustered alongside cultivars but also shows wild relatives clustered together with landrace material. This could potentially be due to backcrossing of cultivated material with its landrace material or landraces with their wild progenitors.

An alternative explanation for the unexpected clustering would be that these are potentially paraphyletic groups, a phenomenon that occurs when a common ancestor is present in a phylogenetic tree but only some the descendants are present (McLennan, 2010), ultimately resulting in incorrect assumptions that the accessions are unrelated, when actually they are related. This could be due to sampling a subset of the JIC Pisum collection or human error mis-characterising these peas when first characterised. Jing et al. (2010) even states that Group 1 whilst predominantly landrace, is not made up of entirely landrace material, but also includes some cultivars, which may explain this observation.

Furthermore, the method may in part play a role in the presence of unexpected clustering. Firstly, in section 1.5.4, we explore how the nature of GBS sequencing can contribute to missing data through mutations in the cut site may causing allele drop out. Therefore, not all loci were used in the generation of the phylogenetic tree. The missing data could be improved upon thought imputation of data using the reference this thesis has generated. Whilst this could reduce missing data, it also has potential to incorporate errors or spurious SNPs.

Alternatively, it could be more prudent to use SNPs from a single locus across all accessions to generate the phylogenetic tree which could produce tighter clustering and that perhaps the statistically associated loci found in this thesis would be good candidates to use. Finally, this tree was drawing using a Neighbour Joining algorithm and perhaps alternative methods such as UPGMA, Maximum Parsimony or Maximum Likelihood methods would yield more resolved clustering.

Discriminant Analysis of Principle Components was performed to explore genetic variability between groups. This differs to PCA which only considers variance as a whole, whereas DAPC increases variance between groups and decreases variance within groups (Jombart, 2008). DAPC does this based on the alleles provided. It could be suggested that the different loadings (allele co-efficients) provide some insights into domestication and genetic causes for common bottlenecks such as the domestication bottleneck and the modernisation bottleneck. Group 4 appears further apart from the domesticated material across the horizontal eigenvector suggesting a domestication bottleneck (Wouw et al., 2010) whereas the vertical eigenvector indicates group 4 sits slightly closer to group 1 (landraces) than it does to group 2 (cultivated) indicating a potential smaller modernisation bottleneck (Wouw et al., 2010).

4.4.3 Limitations

The limitations with the GWAS method are four-fold: sample size, incomplete genotyping, genetic heterogeneity and genetic background (Bush and Moore, 2012).

Whilst this GWAS held 350 samples, increasing the sample size would improve statistical power. Some samples did not have genotypic data due to sample dropout in the lab or allele drop out from removal of reads based on coverage. Additionally, not all SNPs would pass the filter of MAF 0.01 and missing data 20% meaning incomplete genotyping will have occurred based on allele dropout. Whilst imputation, the use of a reference to impute genotypes that are not present in the analysis, would have reduced the incompletion of the genotype, it also carries a risk of imputing genotypes that may not necessarily be present, further adding to the false positives found in GWAS. Here, we suggest that the use of a cultivated variety as a reference may not be suitable for imputation in wild varieties (Figure 4.15). Furthermore, not all phenotypes were measured for reasons such as the loss of a sample. GWAS can only be performed where all samples hold a corresponding phenotypic value and genotypic data.

One must consider the likelihood of the associated loci being the causative SNPs for these associated loci. GWAS does not guarantee that statistically associated SNP is the causative effect of the phenotype, but may be a "tag" SNP in linkage disequilibrium with the SNP (Bush and Moore, 2012). They may also not lie in a gene area but in an intergenic region, however, the method of using a methylation-sensitive enzyme, ApeKI, in GBS may lead to more SNPs found in genic regions since ApeKI will not cut methylated Cytosines (Elshire et al., 2011). Genes which are being silenced, such as repetitive regions, will not be cut and therefore not sequenced (Elshire et al., 2011). This is particularly important in a repetitive genome such as pea.

Genetic background is a common limitation of GWAS but was adjusted for by the use of Mixed Linear Models to account for kinship and population substructure for the core collection. DAPC was used to identify if any samples needed to be removed and there were no samples identified as requiring removal.

GWAS is also additionally limited to the type of variants it detects. It is best suited to common variants with small penetrance (Bush and Moore, 2012) meaning that rare alleles are not detected.

The main limitation of the *de novo* assembly is that long mate-pair data was not used. This is an expensive dataset but this could have further improved the chances of a producing a more contiguous assembly by helping to scaffold the contigs produced and perhaps allow for gap closing to reduce the fragmentation of an assembly.

Furthermore, the peaks on the GWAS suggest statistically significant association at genetic loci, but the gene within this loci cannot be determined without annotating the genome. The lack of an annotation is a major limitation for other downstream analysis but the presence of transcriptome data can help improve the annotation from an *ab initio* annotation which requires no prior knowledge to evidence-based annotation with the use of a transcriptome and more importantly, a functional annotation would provide the most use in comparison to a structural one.

4.4.4 Future work

Therefore, astute future work would be to include Long Mate-Pair data for scaffolding the current contigs. This will improve on contiguity. Additionally, additional sequencing to improve coverage of the reference and GBS datasets would improve confidence. Combining these approaches with existing transcriptomic data, alongside the performance of epigenetic datasets would help to fine-map any loci we might find, particularly in conjunction with a functional annotation.

Fine-mapping by integrating multi-omic data sets may assist in identifying SNPs that are more likely than others to be better candidate genes (Schaid, Chen, and Larson, 2018). There are tools available for fine-mapping such as PAINTOR (Kichaev et al., 2014), gwasMP (Wu et al., 2017), RiVIERA (Li and Kellis, 2016) and FINEMAP (Benner et al., 2016) however these tools are predominantly designed for human datasets, given the wealth of sequencing data in this organism.

An alternative way to validate candidate genes would be to use reverse genetics approaches to either edit the gene (using techniques like CRISPR/Cas (Cong et al., 2013)) or to knock-out the gene (using techniques like homologous recombination (Puchta, 2002)) to identify and validate the gene which causes an observable change in phenotype.

4.4.5 Conclusion

In conclusion, this thesis chapter has created a whole genome assembly of *Pisum sativum* and combined Genotyping by Sequencing data to perform a genome-wide association study. The results have identified potential regions of interest for leaflet teeth, seed perimeter and seed eccentricity. The new knowledge obtained from this study has potential to be used to improve current breeding practices and ultimately, food security.

Chapter 5

Discussion

Peas are important crops in the context of food security since they are legumes and capable of fixing nitrogen (Graham and Vance, 2003). Legumes can act as way to reduce fertiliser (Venkateshwaran and Ané, 2011), improve soil quality (Tilman et al., 2002), reduce the need for making fertiliser through the Haber-Bosch process - which often require energy from fossil fuels (Menge, Wolf, and Funk, 2015) - but also as a way of pest and disease management (Abawi and Widmer, 2000) and reduce weed growth through intercropping (Liebman and Dyck, 1993).

Peas are a nutritious food source. In the current political climate, where countries such as Britain are considering leaving the European Union, we must consider the effect of food imports on national human health. Europe has been known to supply foods rich in vitamins A and C (Macdiarmid et al., 2018). Co-incidentally, peas happen to be rich food sources in these particular key nutrients in comparison to other legumes and other grain cereals (Table 1.1), raising a possible candidate crop of interest agriculturally.

Understanding current elite cultivars is important. Over the course of time, humans breeding crops with agriculturally and commercially beneficial traits has been a common theme throughout evolutionary history. Human survival depends on it, but there can be serious consequences for the plant caused by genetic erosion, making plants more susceptible and less able to fight abiotic and biotic stresses (Wouw et al., 2010). Therefore, understanding key genes that may have been artificially selected is vital to improving food security. This thesis captures the genetic diversity of pea through the study of wild, landrace and cultivated material with a view to exploring domestication.

Darwin's theory of evolution by natural selection explains how selection pressures mean some alleles are selected for and passed on to the next generation. Mendel explained how these alleles are inherited together using pea as example through phenotypic observation and statistical rigour (Mendel, 1996). Miescher was the first to document the extraction of genomic material - the material that encodes the concepts of Darwin and Mendel's work (Dahm, 2008). The work outlined in this thesis builds on the early foundations of these scientists work in order to understand the domestication of the common pea.

This thesis presents its major novel contributions as the development of MktStall a user-friendly, fully-automated, multi-organ image analysis tool and the production of a publicly available pea genome reference. Based on these major contributions produced, statistically significant associated loci for leaflet teeth (Figure 4.35), seed eccentricity (Figure 4.28) and seed perimeter (Figure 4.25) were obtained and potential candidate genes identified. Furthermore, these contributions rest on a new core collection constructed from the John Innes Centre *Pisum* germplasm collection based on seed weight. The output of these contributions augments the understanding of pea domestication.

5.1 Domestication

This thesis has been based on a previous study using RBIP markers (Jing et al., 2010) on all accessions within John Innes Centre *Pisum* germplasm which categorised them into Group 1 (predominantly landrace), Group 2 (cultivar) and Group 4 (wild, but named in the paper as group 3 as it ignores the unclassified group). This entire thesis was built on this population structure. The core collection was based on representing all 3 groups using seed weight, which is a known domesticated trait. The findings show divergence between wild and more domesticated material and this presence can be observed clearly on a phylogenetic tree (Figure 4.15). The analysis of the alleles obtained portrays possible bottlenecks in the genome's domestication history observed using Discriminant Analysis of Principle Components, one indicating a domestication bottleneck (between wild and landrace) and another indicating a modernisation bottleneck (between landrace and cultivars). This fits with part of the proposed domestication hypothesis outlined by Wouw et al. (2010) but the dispersal bottleneck is harder to elucidate.

Some traits are known to be observed in domesticated material. This thesis outlined a phenotypic analysis of domesticated traits. Domesticated phenotypes are important because often there is a beneficial by-product associated with this. For example, large seed weight linked to better plant vigour (Peksen et al., 2004), small plant height with reduced risk of lodging (Tar'an et al., 2003), increased number of leaf teeth for plants growing in cold environments (Xu et al., 2009; Royer et al., 2005), seed wrinkledness and perceived sweetness (Carpenter et al., 2017; Rayner et al., 2017) or blunt pods and lower rates of seed abortion (Lee, 1988).

Additionally, this thesis identified loci for leaflet teeth, seed perimeter and seed eccentricity, in Chapter 4. For leaflet teeth the candidate gene identified, POWERDRESS, is known to play a role in leaflet size (Suzuki et al., 2018). For seed eccentricity, the candidate gene identified, PHOSPHOFRUCTOKINASE-3, is known to play a role in a glycolysis (Mustroph, Sonnewald, and Biemelt, 2007), and it is the changes in starch structure that affect seed wrinkledness and sweetness.

5.2 Impact of this research on agriculture

The contributions made by this thesis will be beneficial to agriculture. The publicly available genome will unquestionably be of use to the international community because it is the only available pea reference. The knock-on effects within Marker Assisted Selection and Genomics Assisted Breeding to develop improved elite cultivars will only become evident in time. The core collection produced through this thesis can be re-used by others for additional studies of the germplasm collection. The phenotypic work has been carried out, so this information can be passed on and used to inform other studies. The GWAS can be used to identify potential loci for other researchers investigating similar traits. Most notably, MktStall will have an impact on other researchers who currently use general image analysis tools or a combination of different specific image analysis tools. It has the potential to replace current tools since it is a fully-automated, multi-organ image analysis tool.

5.3 Key considerations

5.3.1 Phenotyping

Mendel had a team of 8 helpers and phenotyping took him 8 years, admittedly he was performing crosses too but he looked at 2000 plants (Mendel, 1996). This thesis explored a similar number of plants by investigating 350 accessions in replicates of three. In addition to the development of a core collection, sowing and growing peas, organ harvesting and image acquisition took a year. Additional workers in the phenotyping process would have ensured swift collection of samples avoiding issues such as senescence but could also help with additional sampling such as growing more accessions, growing additional replicates and even phenotyping additional developmental stages. Far more could have been achieved and additional investigations could have been carried out. The lack of manpower for the work outlined has been a considerable disadvantage.

Although Mendel had plenty of manpower in the form of his 8 helpers phenotyping is a physically challenging and burdensome task, but additionally manual phenotyping can also be a subjective process. The hard work and time put into obtaining phenotyping has potential to be wasted if the phenotyper has no expertise especially when phenotypes are subjective. MktStall is an fully-automated way to phenotype SMSDs. MktStall holds potential to phenotype more accessions with greater accuracy and consistency than Mendel's own work.

Imaging the plants through manual means such as scanners or cameras can be low-throughput, however high-throughput phenotyping platforms such as a PhenoSpex (*PhenoSpex*) can scan entire fields multiple times a day, however, not all scientists have access to such expensive frameworks. Oftentimes, within a scientific institute, many people will have to share these resources. Yet these high-throughput platforms can perform tasks quicker than an entire team of workers. Until such specialist kit becomes cheaper and more accessible, both manual phenotyping and image acquisition remain a bottleneck.

5.3.2 Genotyping

GBS is a reduced representative approach to sequencing and was used in this thesis as a means for sequencing the core collection. Alternative approaches would be a low coverage whole genome approaches called Pool-seq (Anand et al., 2016), however, this may not produce sufficient coverage to call SNPs confidently and is also an expensive approach. Other alternatives such as exome capture (Warr et al., 2015) would provide SNPs within gene regions, however, this too is an expensive approach and would require baits to be designed to capture the exome and therefore require a reference. This thesis produced a reference genome. This reference could be improved with the availability of Long Mate-Pair data and long PacBio reads to help reduce gaps and scaffold contigs.

The GWAS outlined in this thesis, could have benefited from imputation which would require the use of this reference. However, since this reference is not a “finished” genome, there is scope to incorporate errors into a method known for false positives causing spurious associations.

5.4 Future work

5.4.1 Annotation

There are two types of annotation: structural and functional. Structural annotations identify features such as genes, introns and exons whereas functional annotation also provides the function of the feature. There are also different ways of doing this, such as *ab initio* and evidence based approaches (Yandell and Ence, 2012). *Ab initio* approaches do not require evidence but often have to be trained and evidence based approaches require existing datasets to be used. In pea, *ab initio* methods may not be appropriate unless trained on pea data itself or closely related legume. Furthermore, Chapter 1 outlines potential datasets such as transcriptomes which could be used to help create an annotation. An appropriate example of a pipeline that will perform this would be MAKER (Campbell et al., 2014).

5.4.2 Fine mapping

If the annotation was performed, the GWAS loci identified could be explored to see if they lie in regulatory regions and epigenetics and functional annotation can be used to do this.

As chromatin becomes unpacked from the histone, it exposes a region - accessible DNA - for the transcription factor machinery to bind to and ATAC-seq (Buenrostro et al., 2015), ChIP-seq (Schmidt et al., 2009) and DNase-seq (Cumbie, Filichkin, and Megraw, 2015) can be used to explore this. Using ATAC-seq as an example, the accessible DNA is fragmented and then sequenced providing 50bp reads. When aligned to the references these reads pile up and peaks can begin to called using MACS (Zhang et al., 2008) or HOMER (Heinz et al., 2010). Further study of these peaks can sometimes show a trough in the peak, which is likely to be a transcription factor binding site and therefore indicates that this is a cis-regulatory element called an enhancer. These regulatory elements can be visualised in a genome browser such as JBrowse (Skinner et al., 2009). Addition of multi-omics datasets would be another informative option. For example, if existing transcriptomic studies were to be added we can correlate gene expression with accessibility to the regulatory element - essentially, linking the enhancer with the gene. The existing phenotypic data from this thesis could be used to find the SNPs involved with the changes in enhancer, ultimately, the regulatory element that can help drive transcription and also helping to look at functional annotation of the enhancer that enhances the transcription of the gene.

Many fine-mapping tools outlined in Chapter 4 are available, but have been designed for human, given the wealth of data available in publicly accessible databases. This provides scope for a novel fine-mapping post-GWAS analysis tool specifically for plants designed around existing plant resources.

5.4.3 Detection of antioxidant concentration in the common pea and its wild relatives

As discussed in Chapter 1, peas were mentioned as decreasing the risk of cancer, in particular, isoflavonoids were mentioned in Chapter 1 as an antioxidant known to fight cancer. It does this by preventing angiogenesis, one of the hallmarks of cancer (Hanahan and Weinberg, 2000). Angiogenesis is the formation of new blood vessels, and in the case of cancer, providing tumour cells with the provision of nutrients but also an exit for waste products and therefore, a route to spread throughout the body (Carmeliet and Jain, 2000). Since, isoflavonoids are anti-angiogenic (Fotsis et al., 1993; Perabo et al., 2008), they can help prevent cancer. Legumes are known to be high in this compound (Dixon and Sumner, 2003) and studies have shown legumes are shown to decrease incidence of many types of cancer (Zhu et al., 2015).

Future work could be targeted towards the detection of the concentration of isoflavonoids in pea. One interesting question is to explore if different varieties of pea are higher in concentration of this compound. Therefore, I propose to grow the same varieties of pea, across wild, landrace and cultivated varieties to explore this question. The first section of this future work would be isolate isoflavanoid from all 350 peas (Oldoni et al., 2011). Using my existing dataset of GBS and the genome assembly would be able to provide the ability to form a GWAS investigating which loci associated with high levels of isoflavanoids.

Furthermore, in Chapter 1, it was mentioned that isoflavonoids play a key role in root nodule organogenesis (Venkateshwaran and Ané, 2011). Isoflavanoid is a chemo-attractant that attracts bacteria to enter the root, thus causing the root nodule to be formed and ultimately for nitrogen to be fixed. Thus, there are two sinks, one in the root, and one in the pea seed that is consumed. Measuring isoflavonoids concentrations in seed and root would be wise, but also in the leaf (where it might be made - the source). Once, isoflavanoid is measured (Oldoni et al., 2011), we can obtain the varieties with the highest and lowest isoflavonoid concentrations to perform RNA-Seq on, with the middle concentrations as a control. A differential expression analysis on the three tissues in leaf, seed, and root can be performed and provide information on which genes are being expressed in different tissues and the effect on concentration.

Overall, this study has potential to identify which varieties have highest isoflavonoid concentration in seed - providing potentially important information for cancer and human health and which varieties have highest isoflavonoid concentration in root - providing potentially important knowledge for food security, agriculture and the environment.

It may be possible to determine which loci are statistically associated with concentration in each organ. If the reference was structurally annotated, we would see precisely where these are located. If the reference was functionally annotated, we would have an indication of which genes are involved (if the SNP is in a coding region), or if the SNP is in regulatory regions (if the SNP is in a non-coding regions).

5.4.4 MktStall and Machine Learning

MktStall can be expanded into its analogy, where on every market there is a market stall selling different produce. MktStall can subsequently be improved to incorporate analysis of other organs such as root and flower or to improve on existing organs by adding other phenotypes such as colour of seed or leaf venation. Moreover, there is potential for MktStall to incorporate the conversion of phenotypic data to other forms such as MIAPPE or ISA-Tab (Ćwiek-Kupczyńska et al., 2016; Rocca-Serra et al., 2010), to allow phenotypic data to be publicly shared and open-source. However, these phenotypic ways of documenting an experiment hold existing tools to convert the CSV file containing phenotypic information (such as the output from MktStall) to their respective formats.

MktStall also has the capacity to be used alongside machine learning methods in order to recognise and taxonomically classification pea varieties. The results from the work outlined in Chapter 3, can be used as a training dataset for supervised machine learning methods. The aim of this is to begin to recognise organs as part of a particular group, wild, cultivar, landrace and to help classify them appropriately. A number of tools do exist that recognise different species (Shafiekhani et al., 2017), but this work would be a more fine-grained approach in recognising and classifying varieties of pea across multiple organs in a fully-automated manner. This would greatly help taxonomy classification of pea accessions, particularly with new varieties, and begin to help us better understand key traits involved in domestication.

Additionally, MktStall could mark the start of great citizen science projects, encouraging non-scientists into understand the organism and to begin using ontologies. The "Great British Pea Census" project is an idea to encourage non-scientists to grow peas, image organs and obtain phenotypes from MktStall before uploading this data onto a web server which would track: variety of pea, location grown, environmental conditions and help build up a horticultural library which can then be used to begin to explore which type of peas grow best in certain areas and under different conditions. In combination with the sequencing data (Chapter 4) this could be a new avenue of research to explore changes in pea phenotypes across Britain and explore which varieties grow best in different locations whilst also incorporating the public into horticulture and generating vast amounts of useful data.

5.5 Conclusions

In conclusion, the work outlined in this thesis has produced novel contributions in the form of a publicly accessible pea genome reference, a novel multi-organ image analysis tool ("MktStall") and identifying potential candidate genes from the genome-wide association study performed through the combination of the above reference, GBS datatsets and output of MktStall.

Glossary

core collection a sample of a gene bank designed to be as diverse as possible. 84

domestication breeding for more desirable traits. 41

Genome Wide Association Studies statistically associating genotypes with phenotypes. 84

MktStall a multi-organ image analysis tool. 108

Ontology a set of definitions and/or relationships. 64

orthologs genes that are homologous in different organisms. 184

Reduced representation sequencing a sequencing method using restriction enzymes. 66

retrotransposon based insertion polymorphism a type of marker. 42

Appendix A

Image Analysis

TABLE A.1: **List of software tools from the Plant Image Analysis database for the different organs described in this thesis.** These are categorised into the following categories: Commerical, Specialised, Other Plant Specific, Other Phenotypes Specific, Not Cross Platform and Usable.

Category	Leaf	Seed	Fruit
Commercial	Assess (*Assess 2.0: Image Analysis Software for Plant Disease Quantification*)	LemnaLauncher (LemnaTec, 2018)	
	Skye (Skye Instruments Limited, 2018)	SeedCount (Next Instruments, 2016)	
	Sigmagis (Smart Imaging Technologies, 2013)	WinSeedle (WinSeedle, 2005)	
	WinFOLIA (WinFolia, 2001)		

Continued on next page

Table A.1 – *Continued from previous page*

Category	Leaf	Seed	Fruit
Specialised	Limani	SeedSize (Miller et al., 2018; Moore et al., 2013)	
	PDQuant (Fitzgibbon et al., 2013)		
	CellArchitect (Faulkner et al., 2017)		
	Phenotic (Rousseau et al., 2013)		
Other Plant Specific	ImAGE (Hill et al., 2011)		TomatoAnalyser (Brewer et al., 2006)
	Leaf Growth (Remmler and Rolland-Lagan, 2012)		Maize Kernel Ear-Cob Analysis (Miller et al., 2017)
	Leaf Recognition (Novotnỳ and Suk, 2013)		P-Trap (Faroq et al., 2013)
	LeafSnap (Kumar et al., 2012)		PANorama (Crowell et al., 2014)

Continued on next page

Table A.1 – *Continued from previous page*

Category	Leaf	Seed	Fruit
	Leaf Analyser (Weight, Parnham, and Waites, 2008) Phytotyping4d (Apelt et al., 2015) Lamina2Shape (Dornbusch and Andrieu, 2010)		PotatoSize
Other Phenotypes	Phenovein (Bühler et al., 2015) NEFI (Max Planck Institute, 2018) Leaf GUI (Price et al., 2010) Leaf Doctor (Pethybridge and Nelson, 2015) Bioleaf (Machado et al., 2016) CompuEye (Bakr, 2005)		

Continued on next page

Table A.1 – *Continued from previous page*

Category	Leaf	Seed	Fruit
	GROW Map-Leaf (Walter and Schurr, 2005) Identify (Joly et al., 2014)		
Not Cross Platform	BlackSpot (Varma and Osuri, 2013) Easy Leaf Area (Easlon and Bloom, 2014) Petiole (Limited, 2018)	Germinator (Joosen et al., 2010) SmartGrain (Tanabata et al., 2012) GrainScan (Whan et al., 2014)	
Usable	Lamina (Bylesjö et al., 2008) LeafProcessor (Backhaus et al., 2010) Leaver (Borianne and Brunel, 2012) Morpholeaf (Biot et al., 2016) Phenophyte (Green et al., 2012)	PlantCV (Gehan et al., 2017)	

Appendix B

MktStall Parameters

Stage	Technique	Value
Masking	Upper green threshold	(35,255,255)
	Lower green threshold	(10,100,20)
Denoising - Image	Bilateral filter	(15,75,75)
Denoising - Label	Fast Non-Local Means denoising algorithm	regulating filter strength = 50, templateWindowSize = 21, searchWindowSize =7
	Global Thresholding algorithm	(threshold value = 127, maximum value = 255, thresholding type = THRESH_BINARY)
	Otsu's thresholding	(threshold value = 0, maximum value = 255, thresholding type = cv2.THRESH_BINARY + cv2.THRESH_OTSU)
	Gaussian and Otsu's Thresholding Algorithm	(threshold value = 0, maximum value = 255, thresholding type = cv2.THRESH_BINARY + cv2.THRESH_OTSU) + 5x5 Gaussian Kernel

Appendix C

Kolmogorov-Smirnov Test

Phenotype	Kolmogorov-Smirnov Test Statistic	p-value
Plant Height	D = 1	p-value < 2.2e-16
Seed Weight (GH)	D = 0.51243	p-value < 2.2e-16
Seed Weight (Field)	D = 0.51084	p-value < 2.2e-16
Seed Area	D = 0.5	p-value < 2.2e-16
Seed Perimeter	D = 0.85977	p-value < 2.2e-16
Seed Equivalent Diameter	D = 0.59309	p-value < 2.2e-16
Seed Eccentricity	D = 0.58496	p-value < 2.2e-16
Leaflet Length	D = 0.94875	p-value < 2.2e-16
Leaflet Width	D = 0.85633	p-value < 2.2e-16
Leaflet Perimeter	D = 0.99961	p-value < 2.2e-16
Leaflet Area	D = 0.93577	p-value < 2.2e-16
Leaflet Teeth	D = 0.5	p-value < 2.2e-16
Leaflet Aspect Ratio	D = 0.87334	p-value < 2.2e-16
Leaflet Roundness	D = 0.72944	p-value < 2.2e-16
Leaflet Compactness	D = 1	p-value < 2.2e-16
Leaflet Rectangularity	D = 0.74438	p-value < 2.2e-16

Continued on next page

Table C.1 – *Continued from previous page*

Phenotype	Kolmogorov-Smirnov Test Statistic	p-value
Leaflet Perimeter Ratio of Length	D = 0.98782	p-value < 2.2e-16
Leaflet Perimeter Ratio of Length and Width	D = 0.94673	p-value < 2.2e-16
Leaflet Equivalent Diameter	D = 0.90713	p-value < 2.2e-16
Pod Length	D = 0.99255	p-value < 2.2e-16
Pod Width	D = 0.85328	p-value < 2.2e-16
Pod Area	D = 0.98513	p-value < 2.2e-16
Pod Perimeter	D = 1	p-value < 2.2e-16
Pod Aspect Ratio	D = 0.97429	p-value < 2.2e-16
Pod Rectangularity	D = 0.65731	p-value < 2.2e-16
Pod Equivalent Diameter	D = 0.94994	p-value < 2.2e-16

Appendix D

Quartile-Quartile Plots

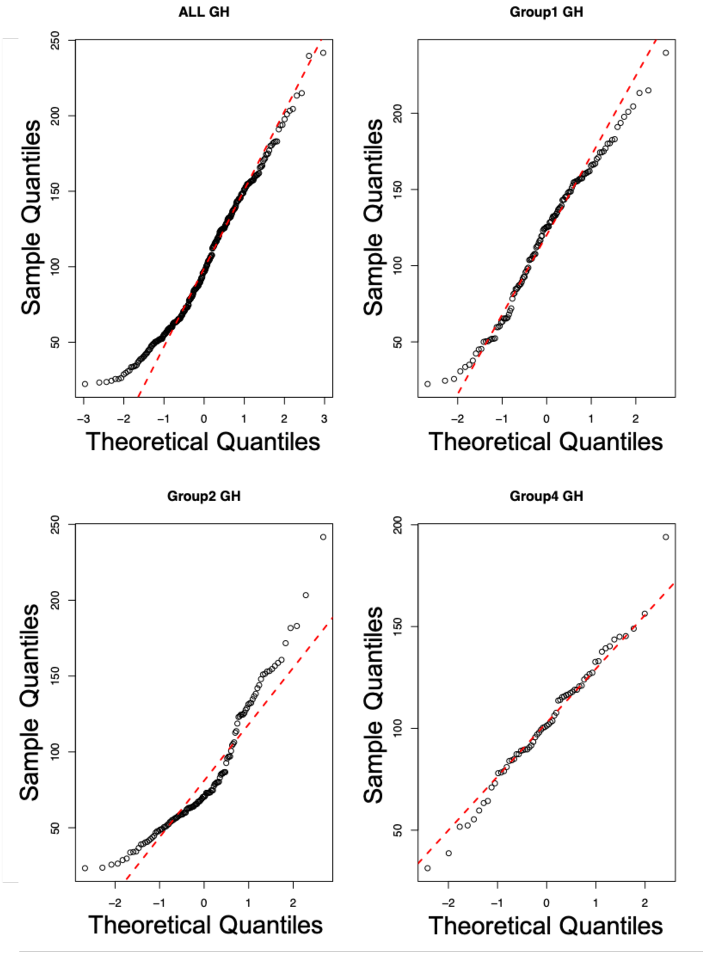


FIGURE D.1: **Q-Q plots for plant height in glasshouse environments.** All GH represents the QQ plot for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild), the corresponding Q-Q plots are also shown.

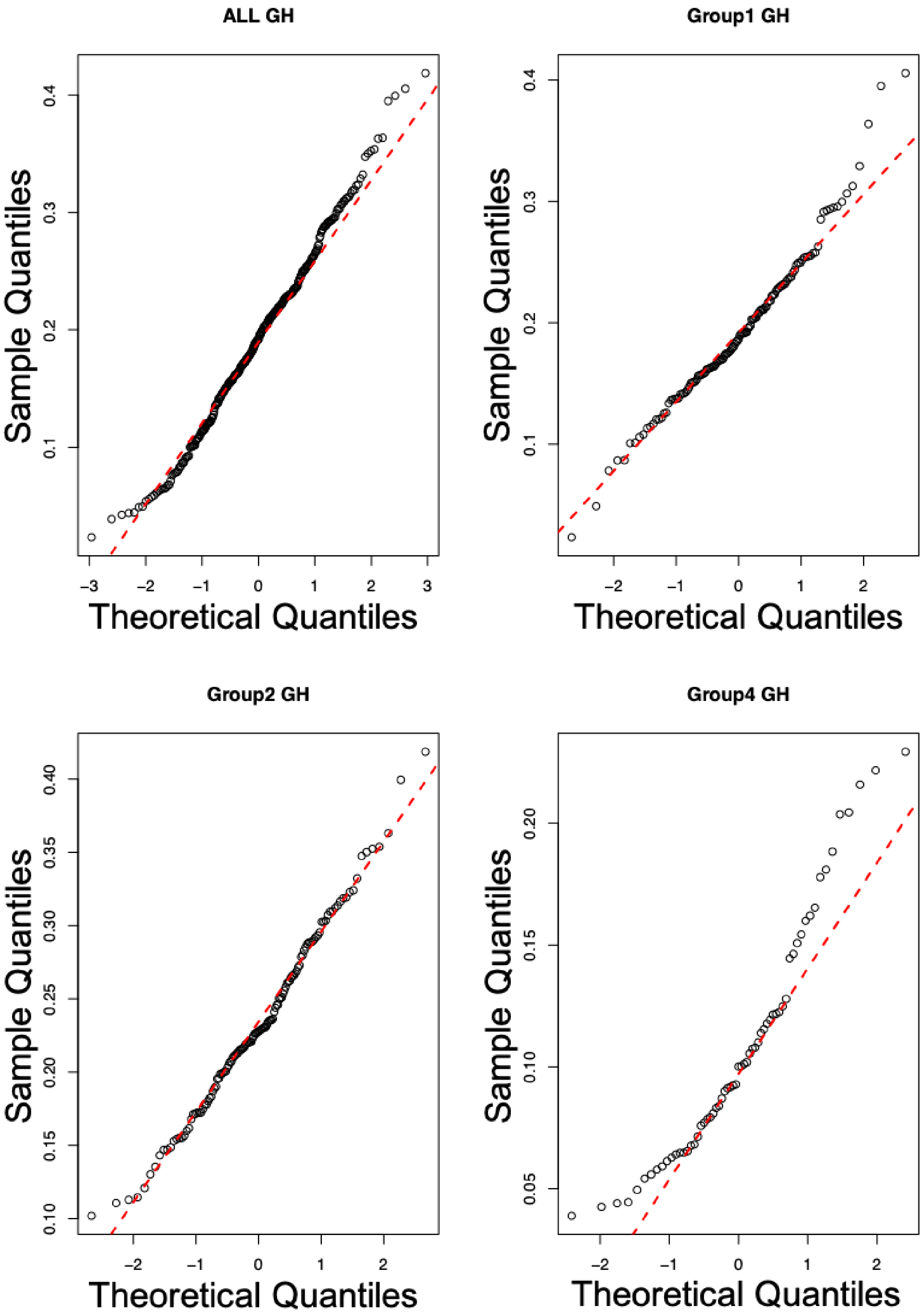


FIGURE D.2: **Q-Q plots for seed weight in glasshouse environments.** All GH represents the QQ plot for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

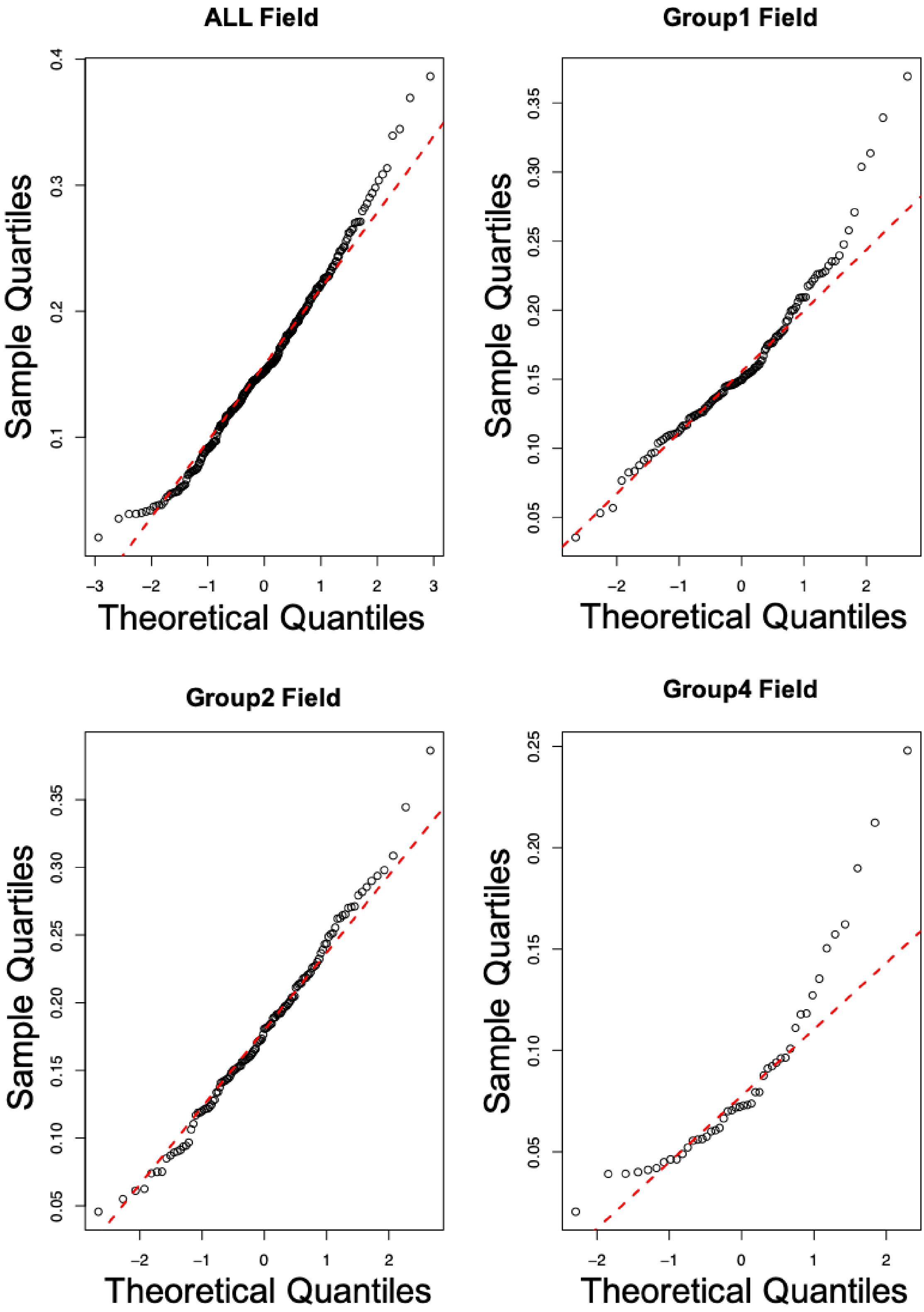


FIGURE D.3: **Q-Q plots for seed weight in field environments.** All GH represents the QQ plot for core collection grown in the field. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

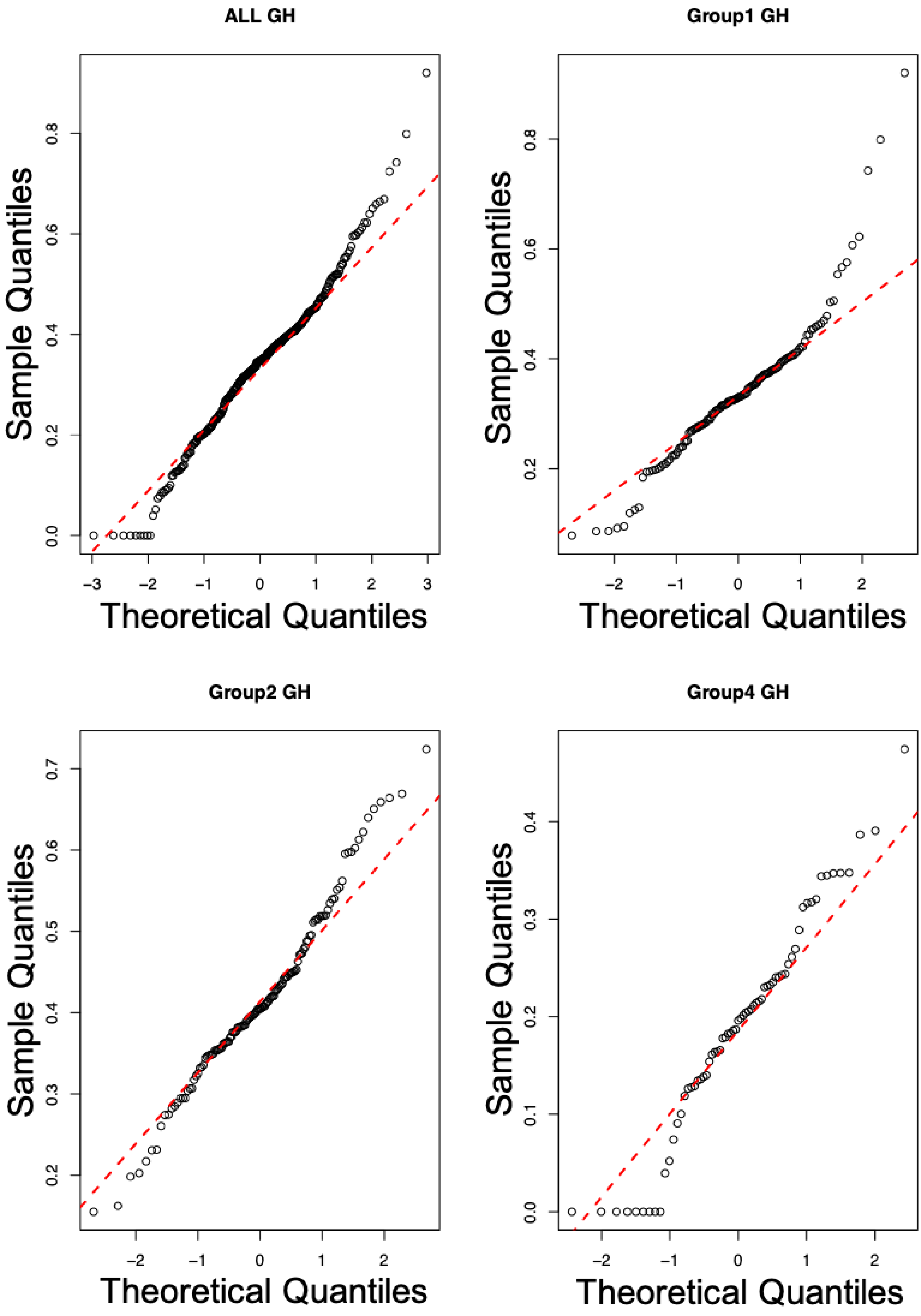


FIGURE D.4: **Q-Q plots for seed area in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

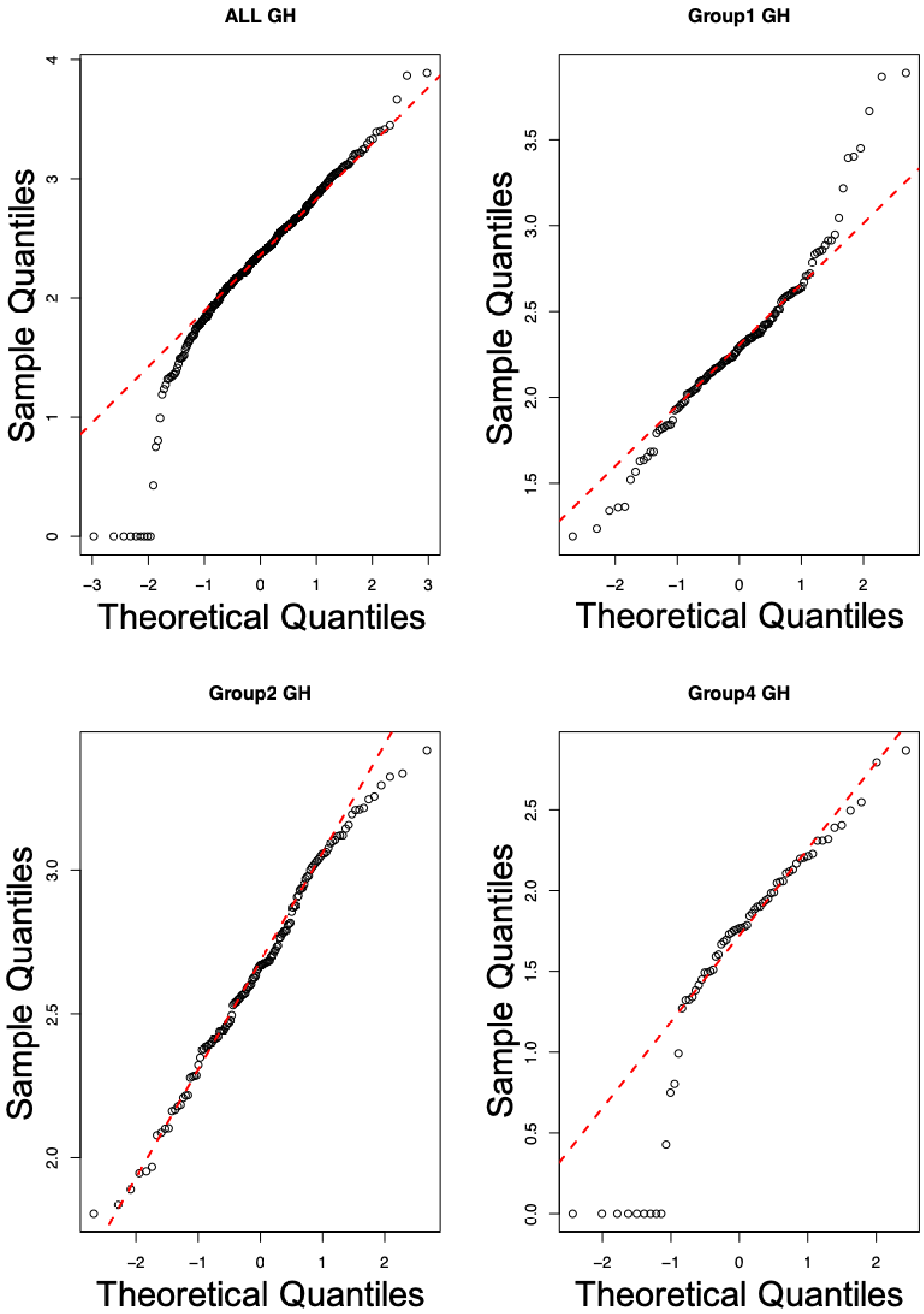


FIGURE D.5: **Q-Q plots for seed perimeter in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

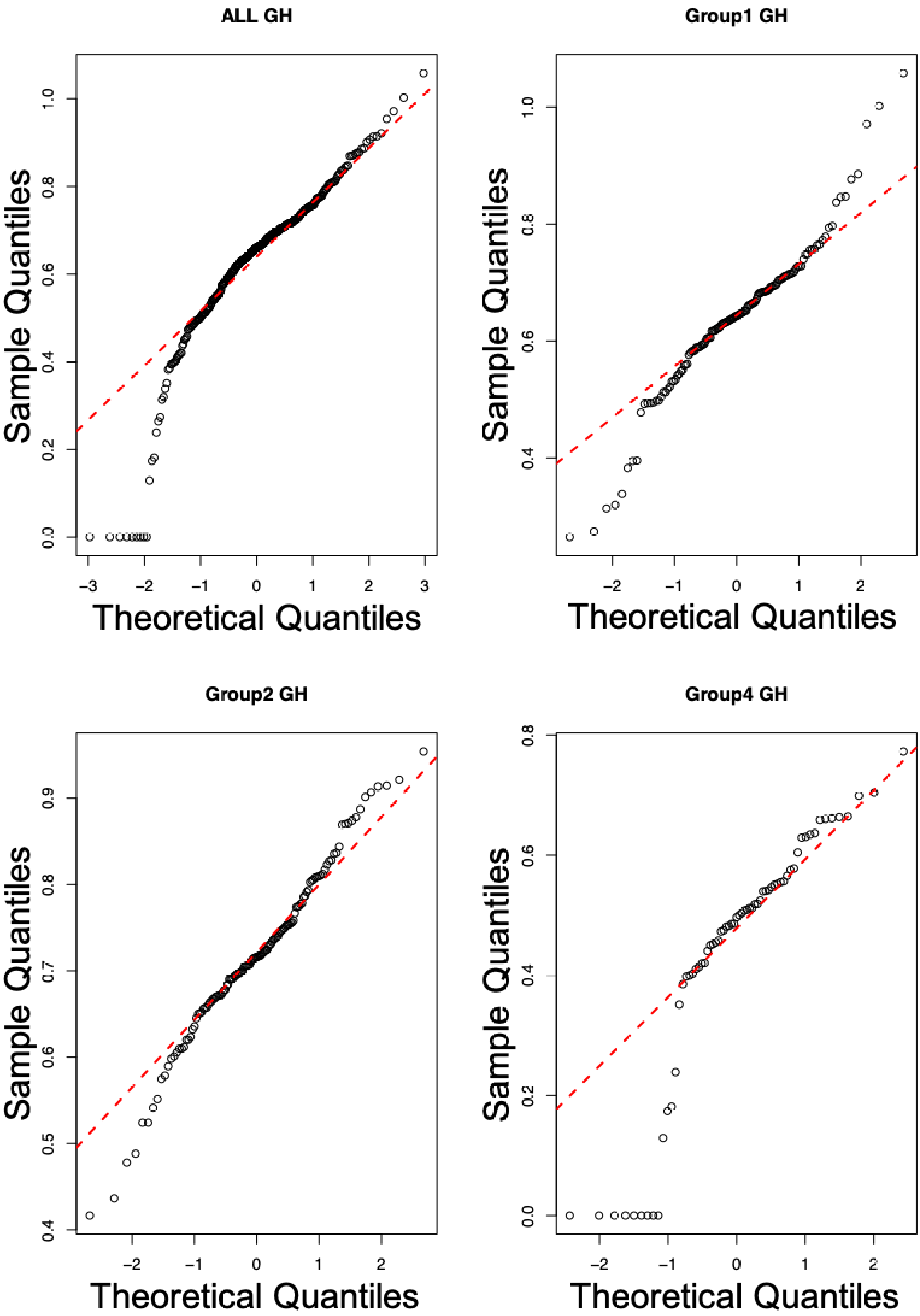


FIGURE D.6: **Q-Q plots for seed equivalent diameter in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

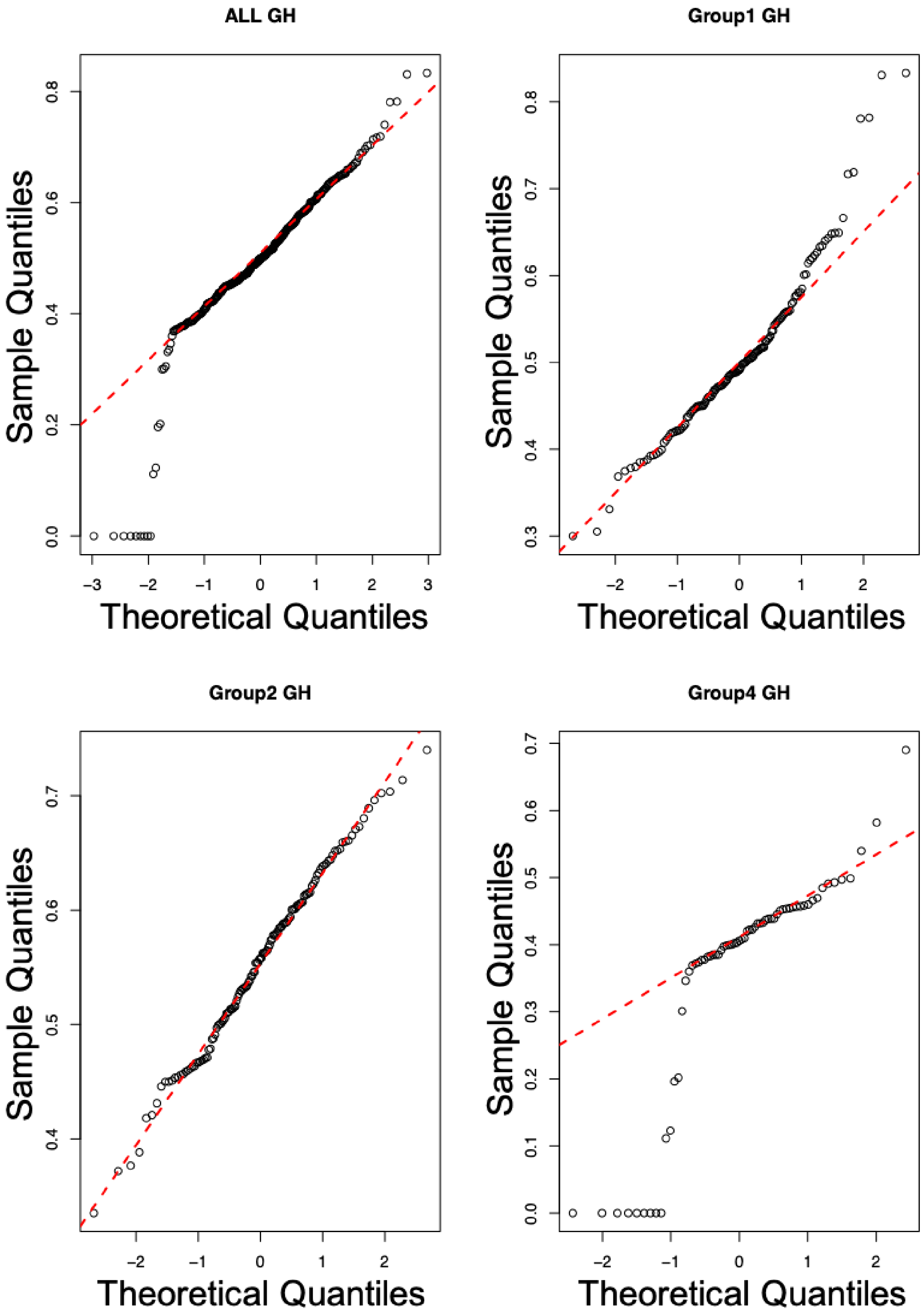


FIGURE D.7: **Q-Q plots for seed eccentricity in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

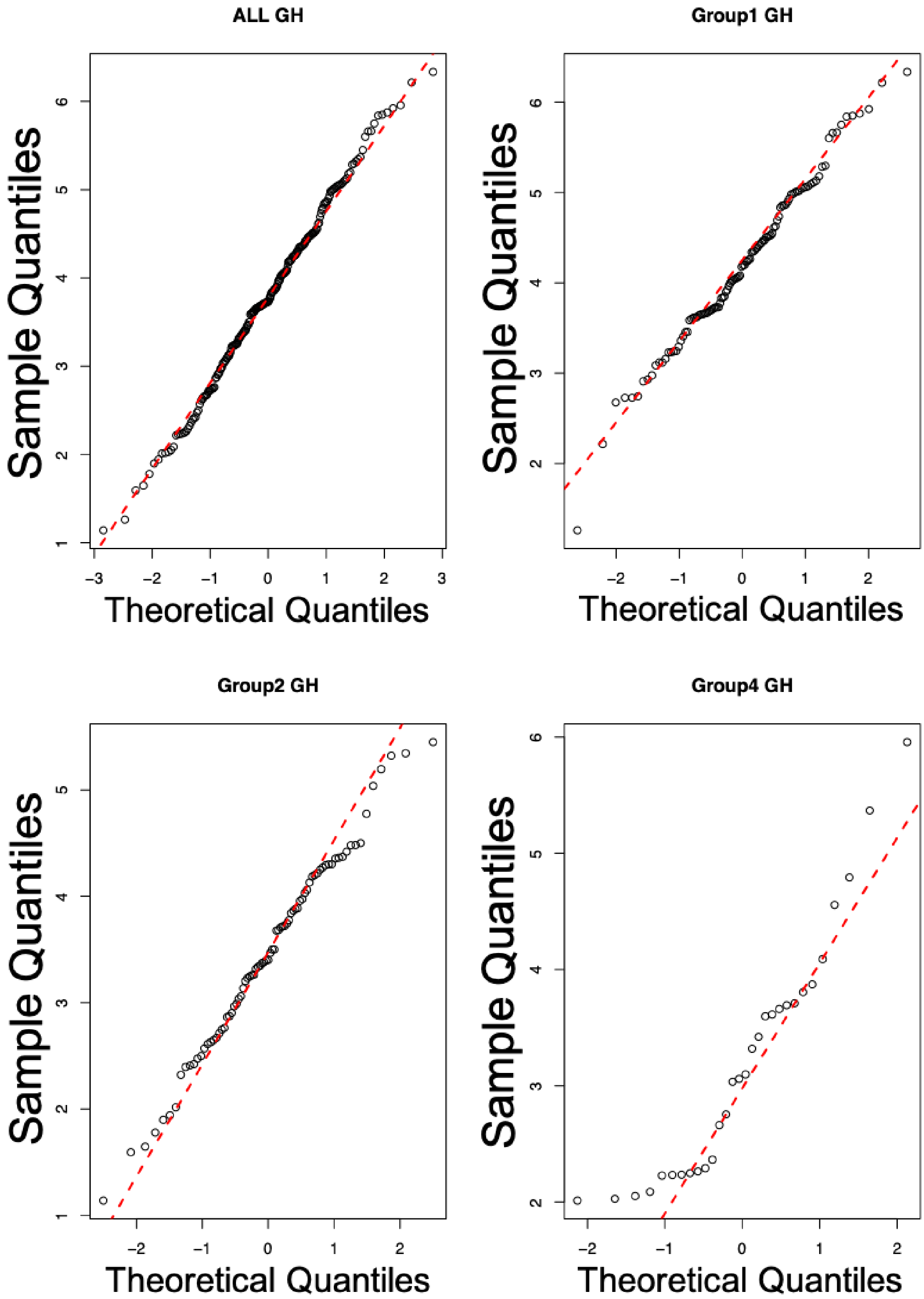


FIGURE D.8: **Q-Q plots for leaflet length in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

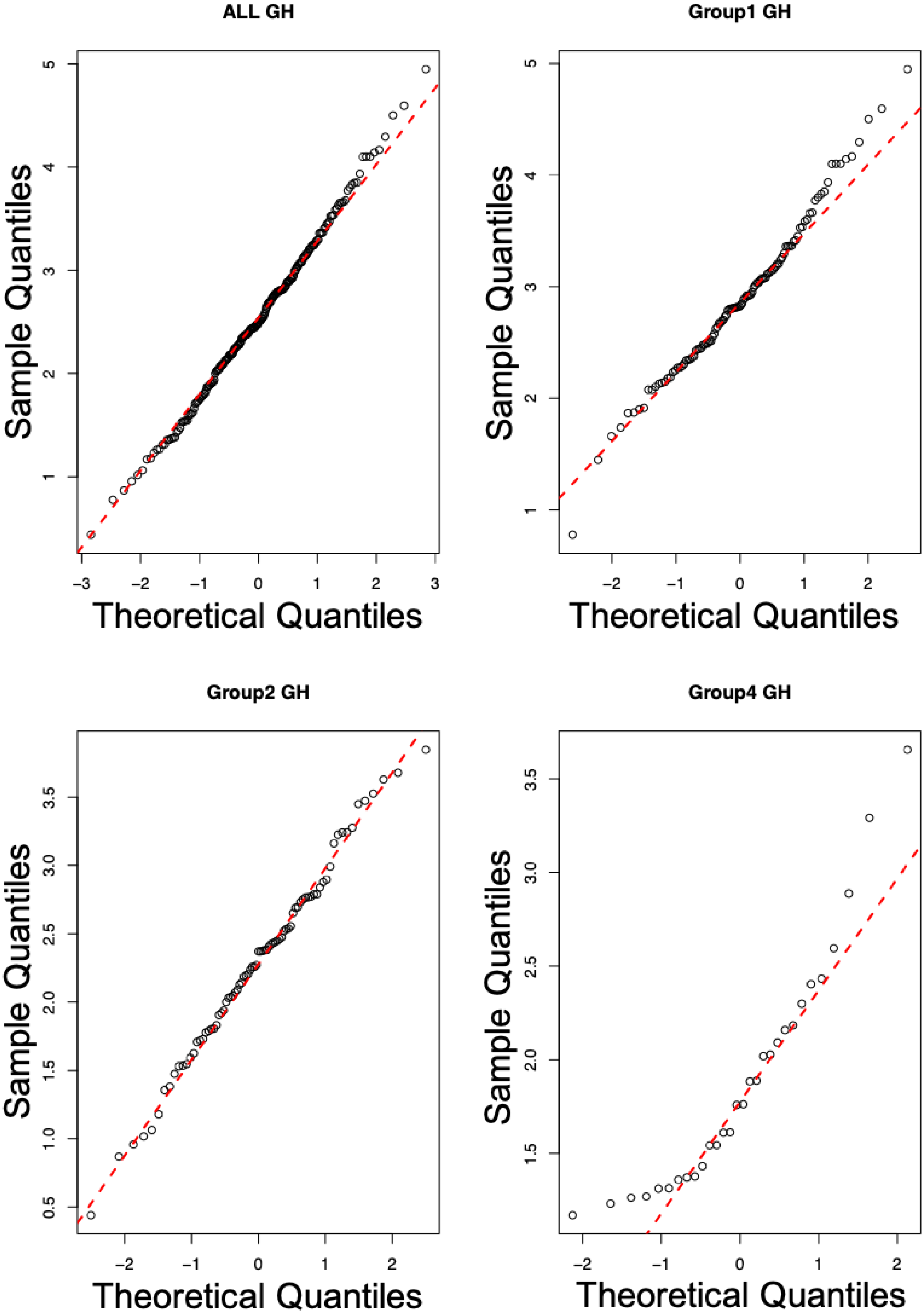


FIGURE D.9: **Q-Q plots for leaflet width in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

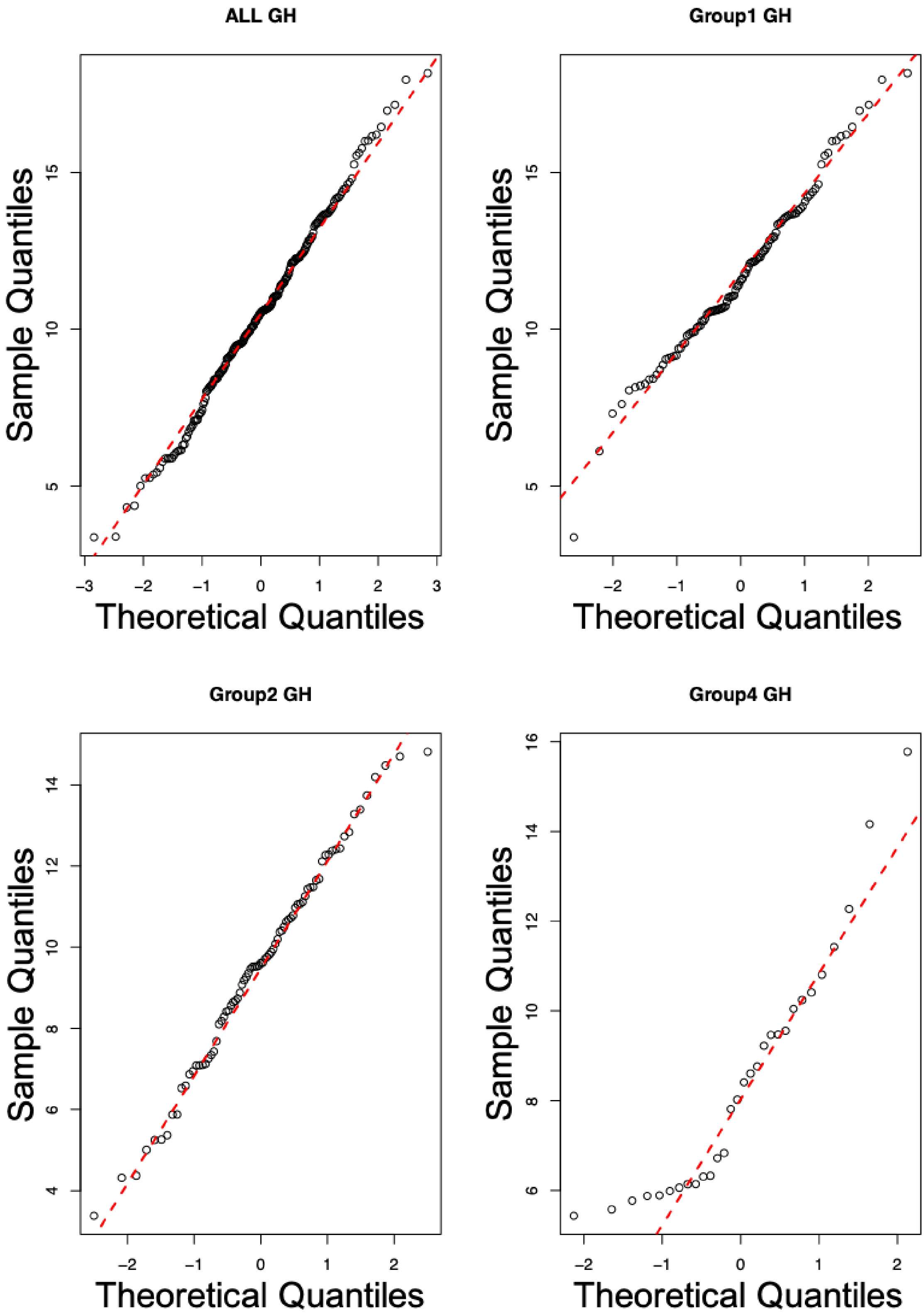


FIGURE D.10: **Q-Q plots for leaflet perimeter in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

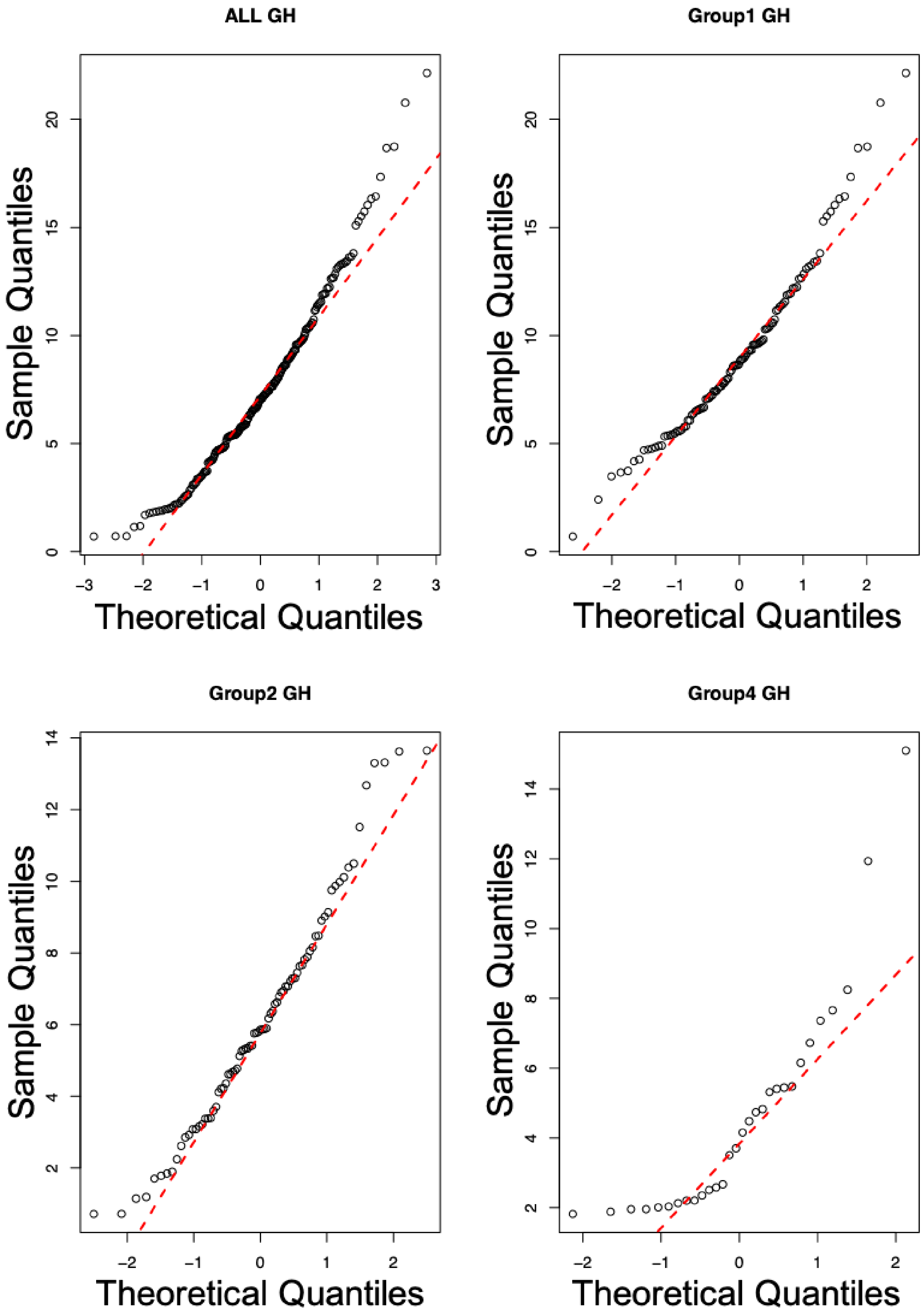


FIGURE D.11: **Q-Q plots for leaflet area in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

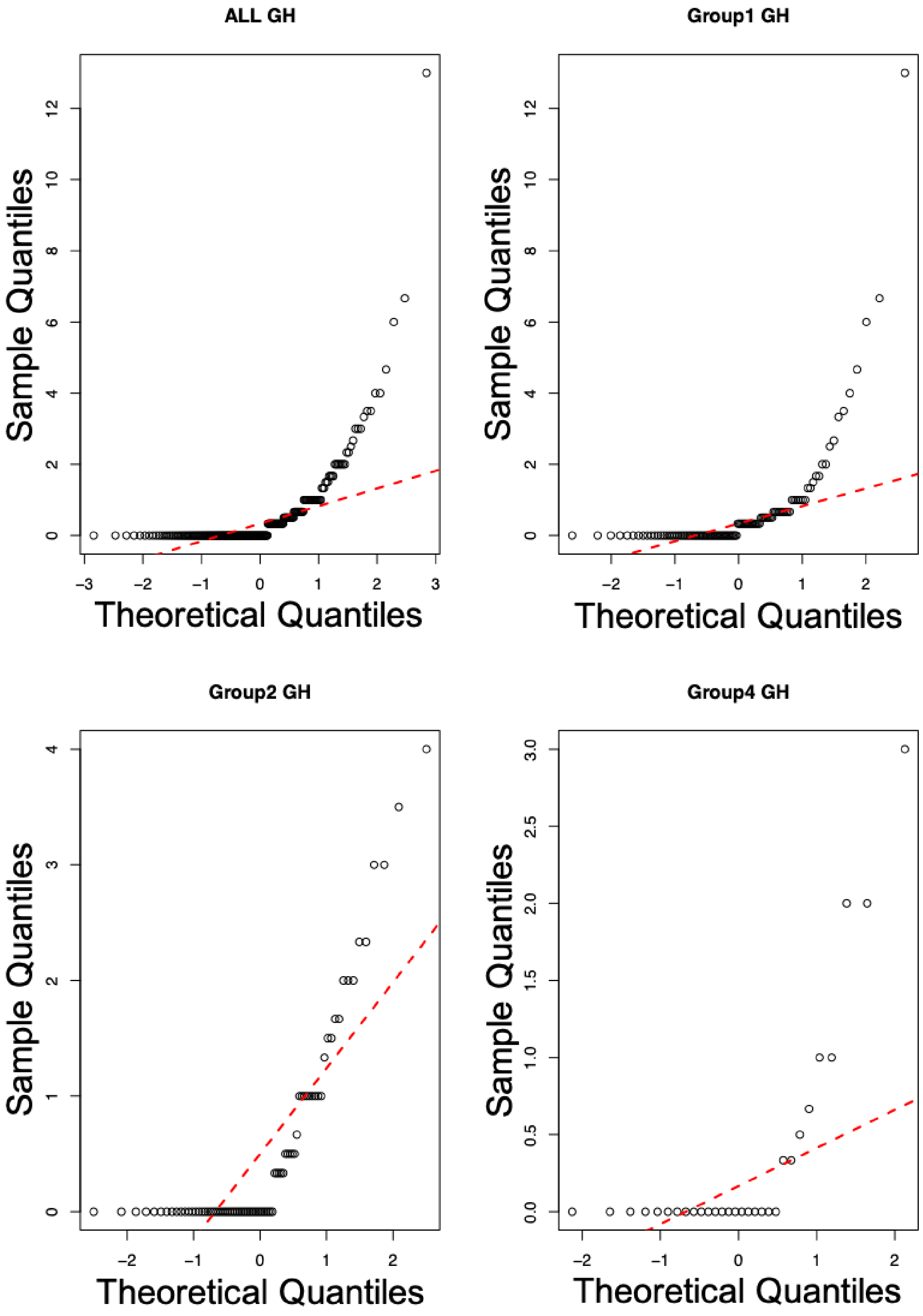


FIGURE D.12: **Q-Q plots for leaflet teeth in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

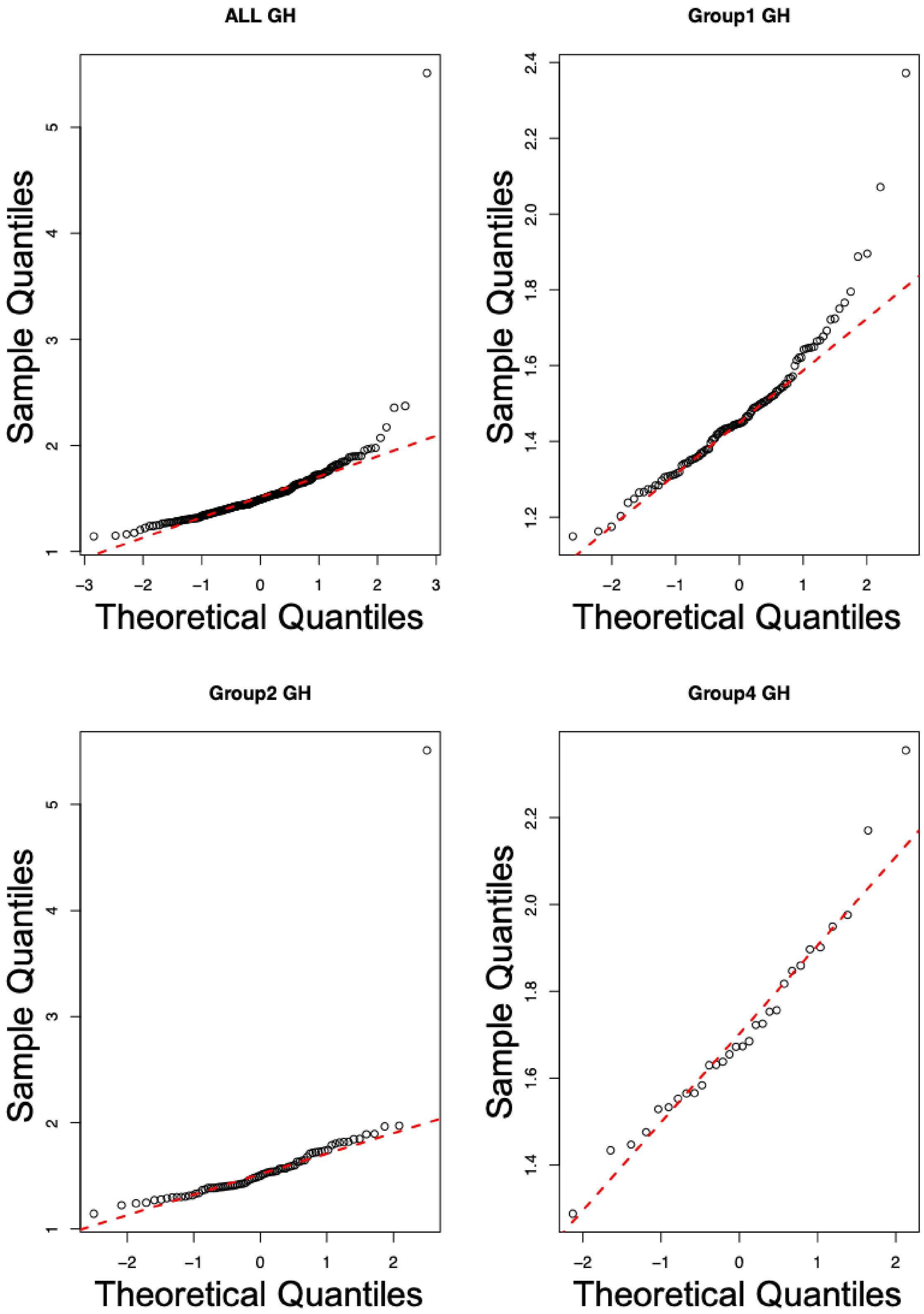


FIGURE D.13: **Q-Q plots for leaflet aspect ratio in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

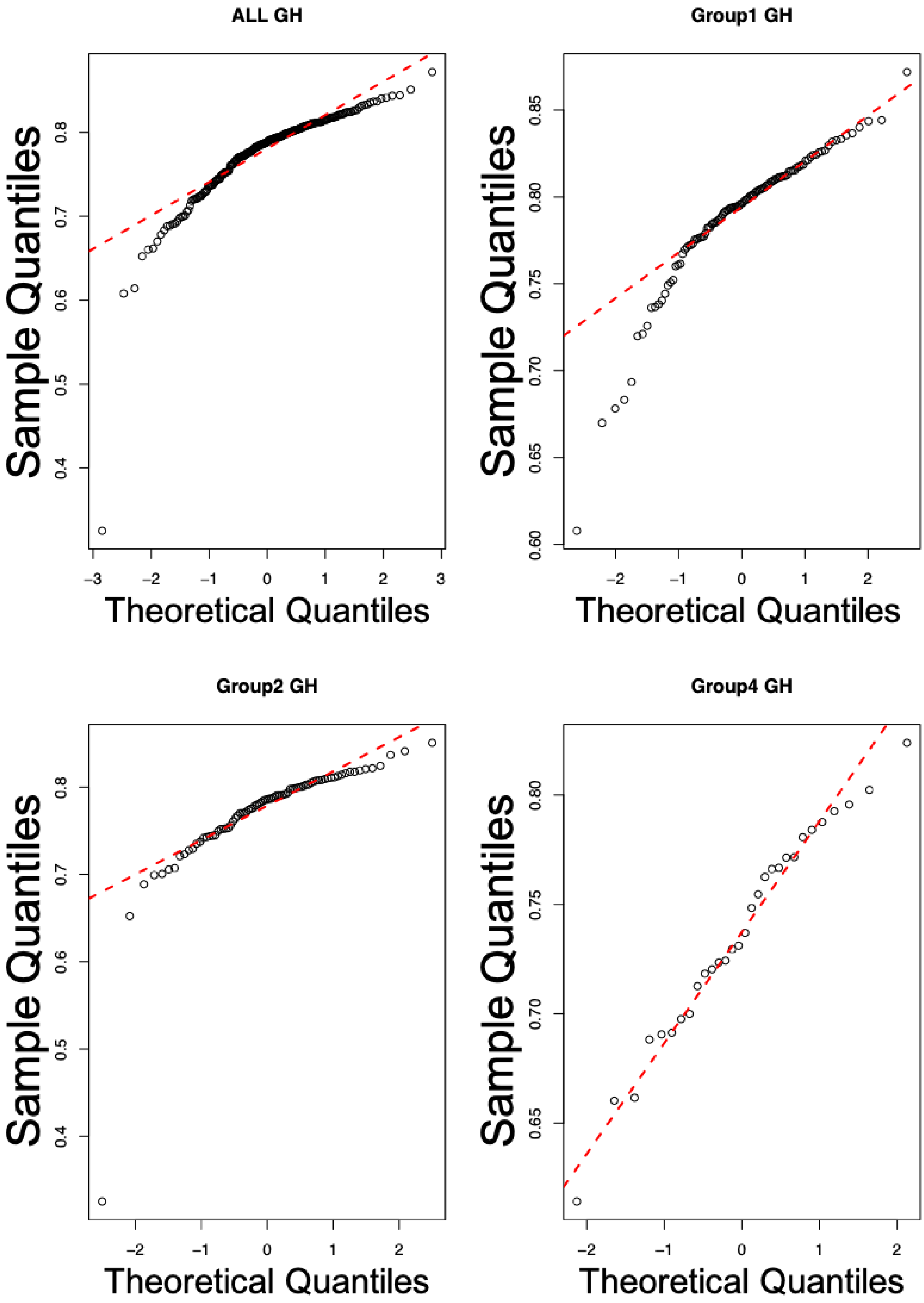


FIGURE D.14: **Q-Q plots for leaflet roundness in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

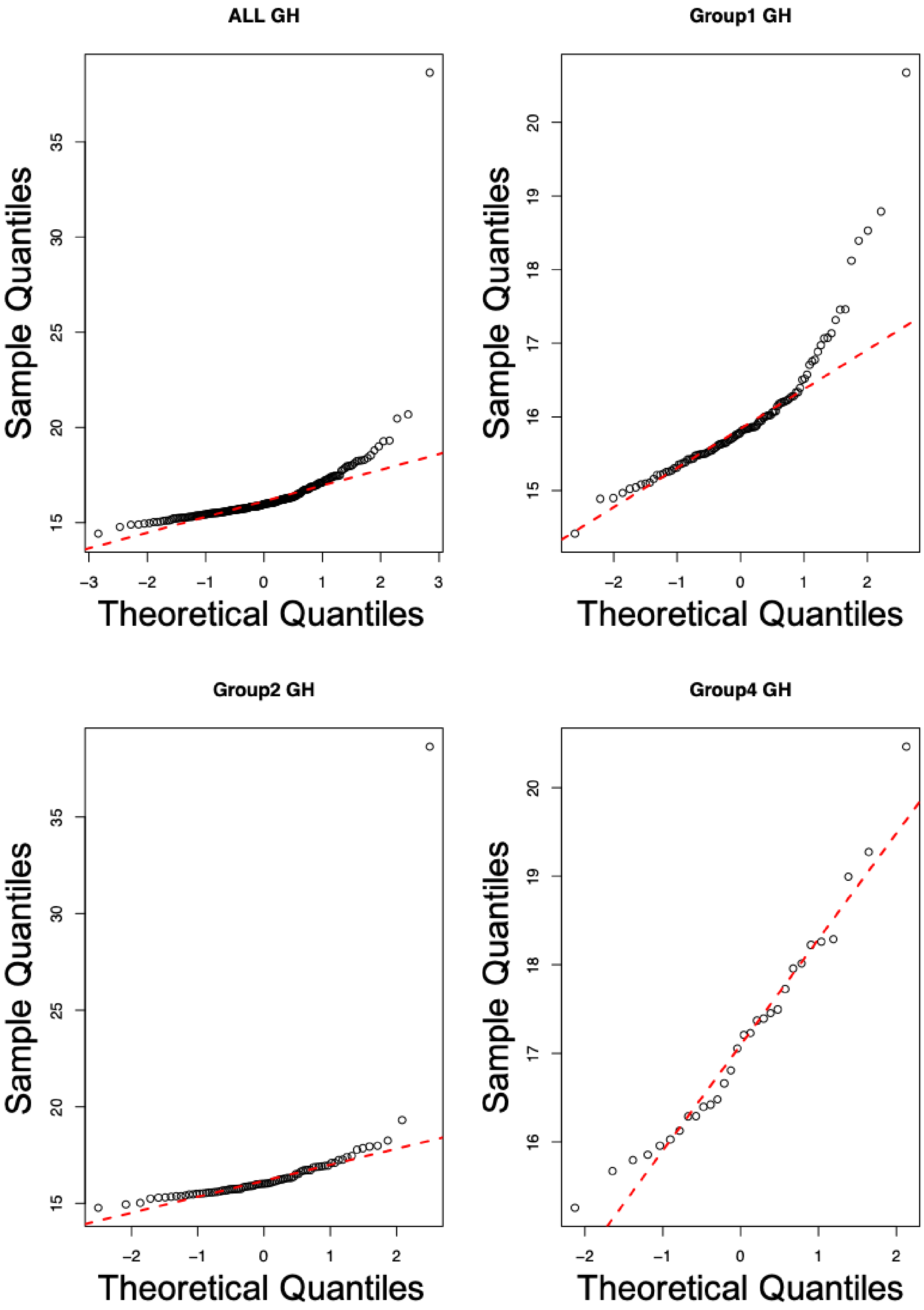


FIGURE D.15: **Q-Q plots for leaflet compactness in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

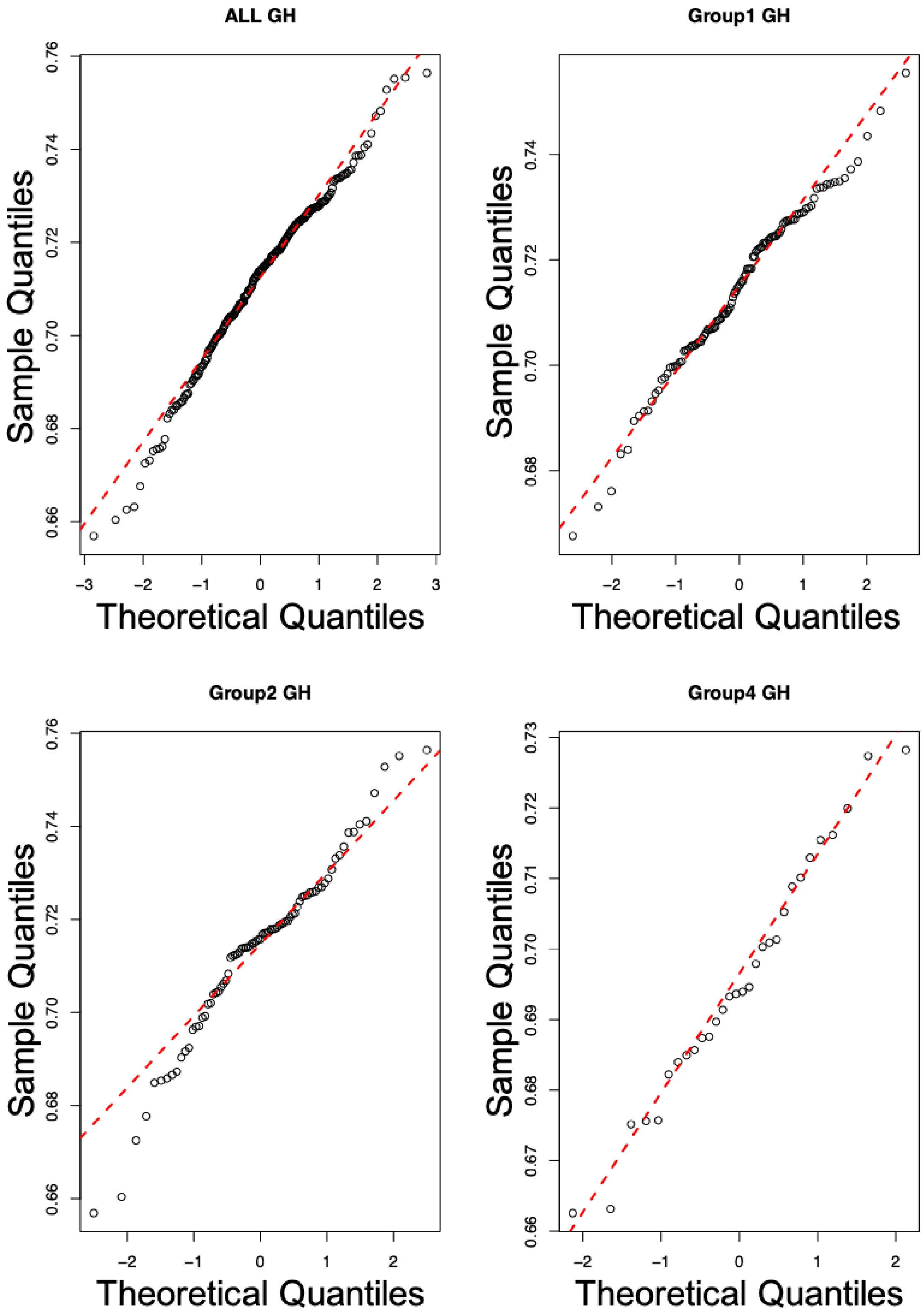


FIGURE D.16: **Q-Q plots for leaflet rectangularity in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

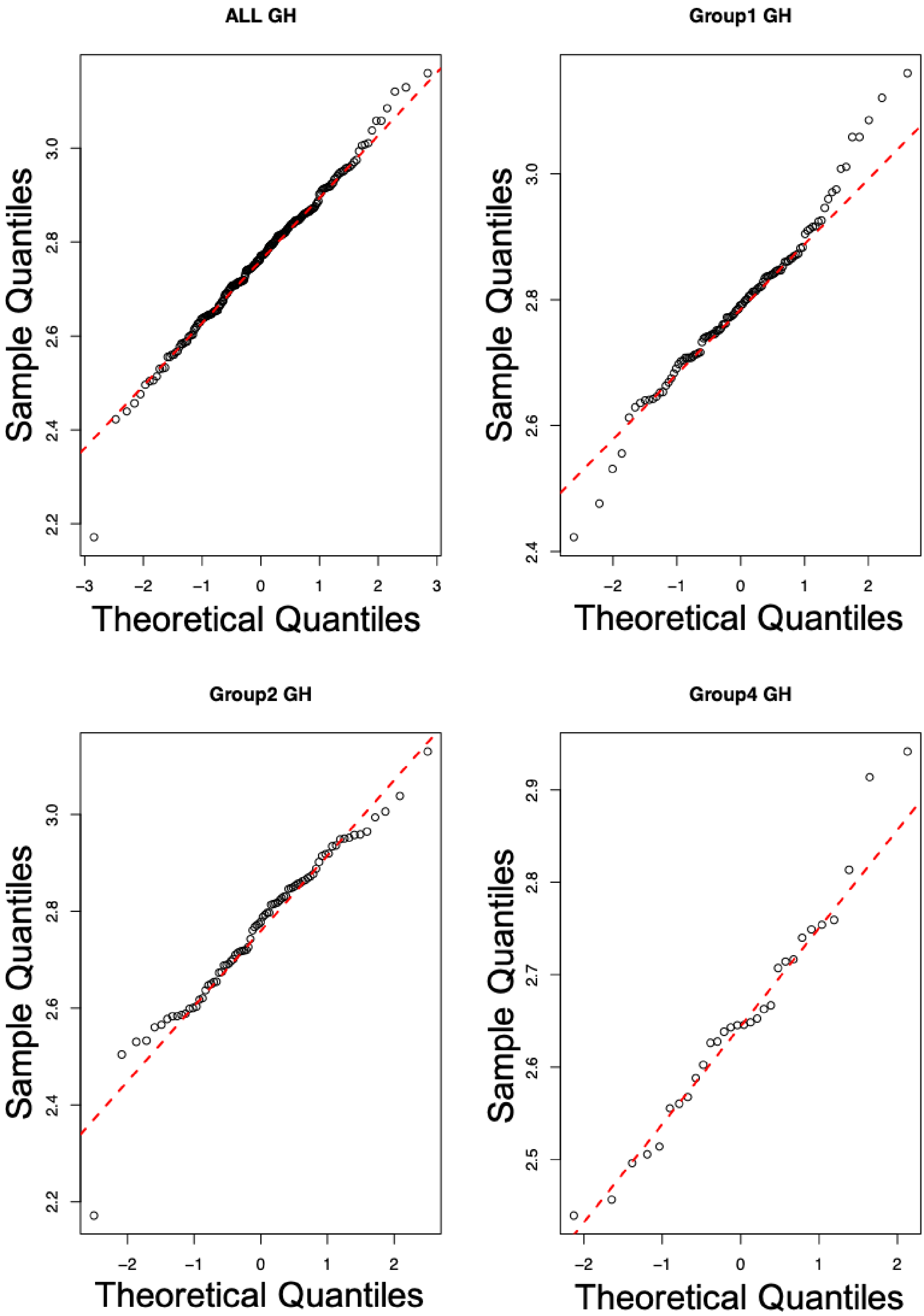


FIGURE D.17: **Q-Q plots for leaflet perimeter ratio of length in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

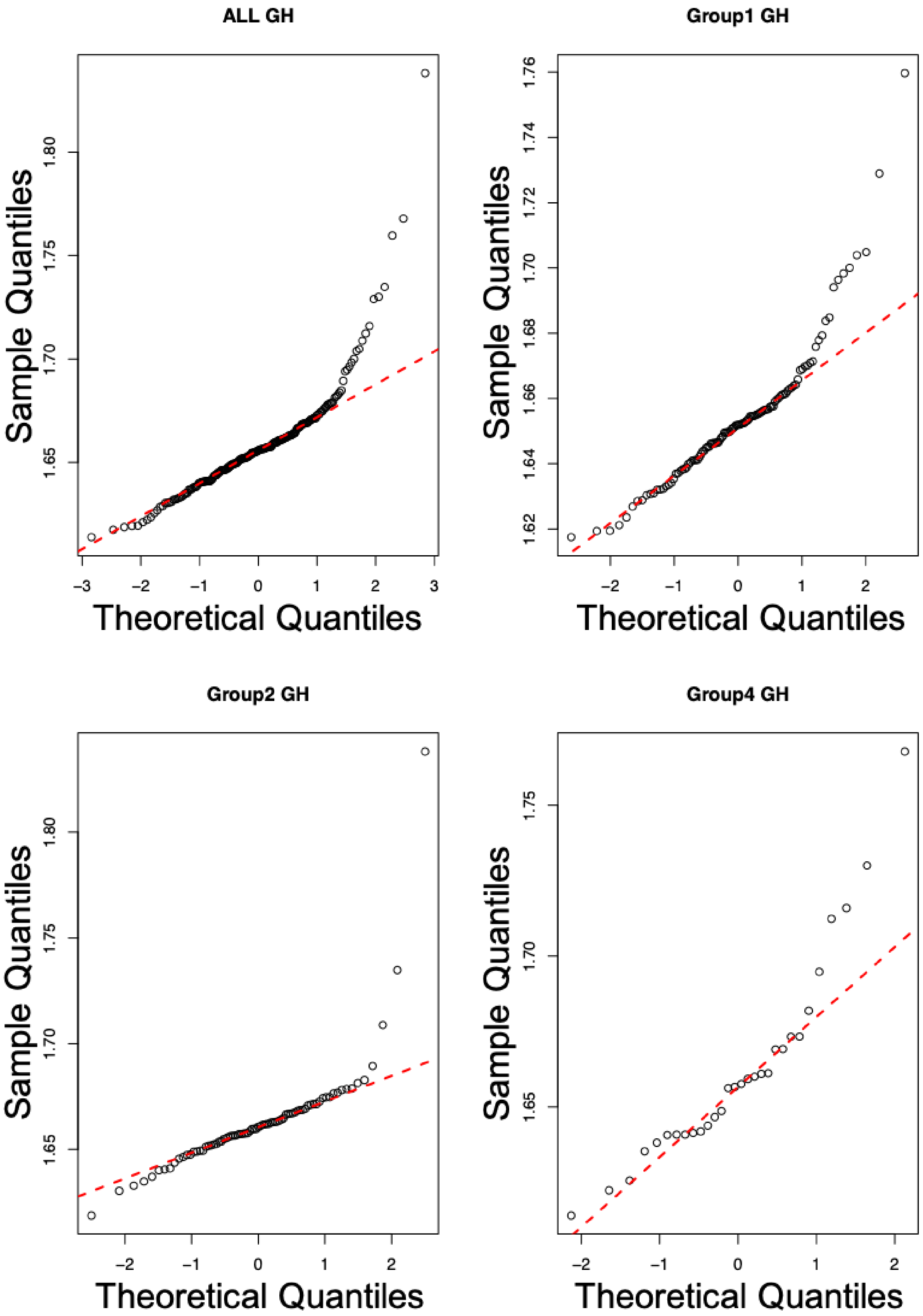


FIGURE D.18: **Q-Q plots for leaflet perimeter ratio of length and width in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

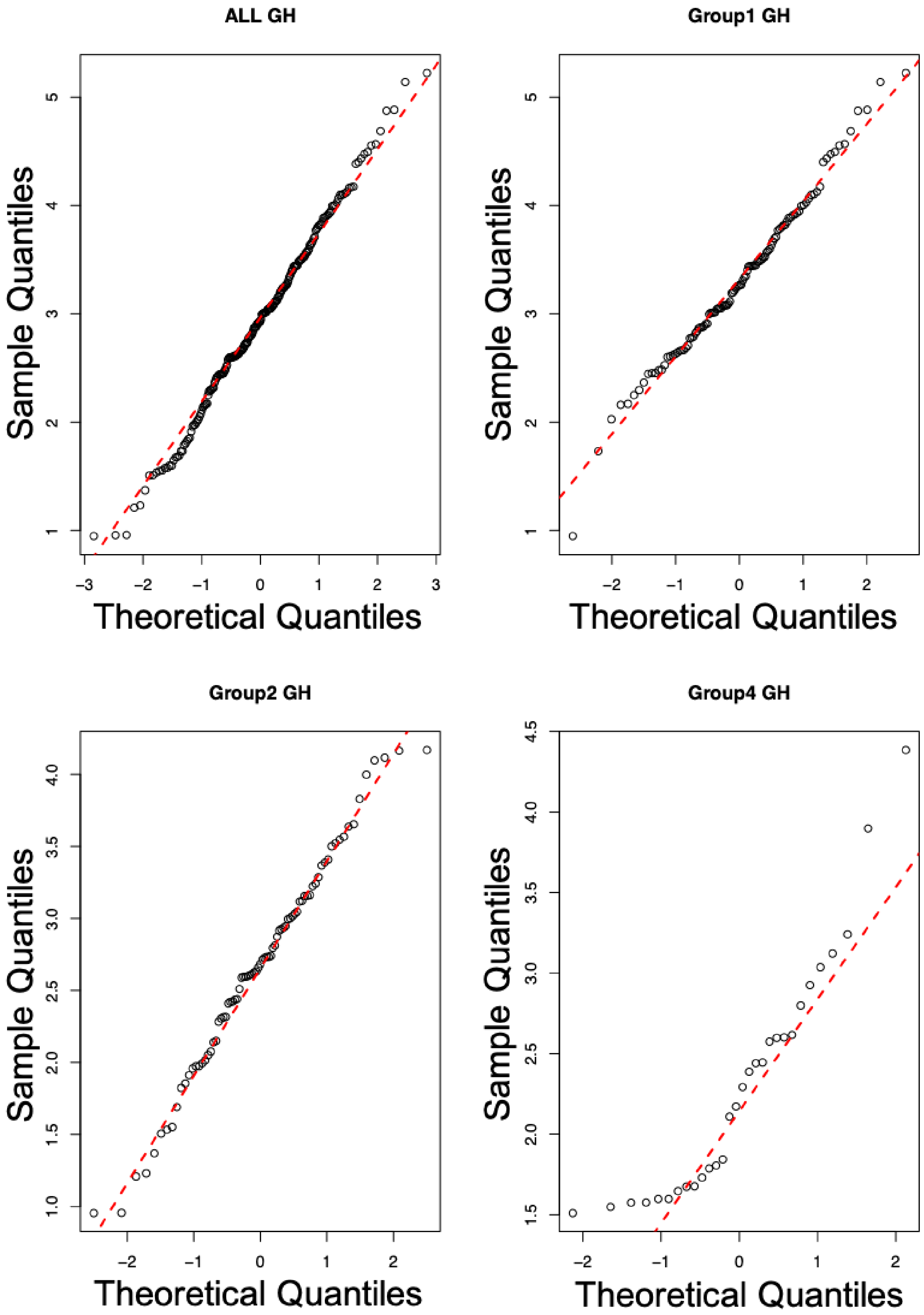


FIGURE D.19: **Q-Q plots for leaflet equivalent diameter in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

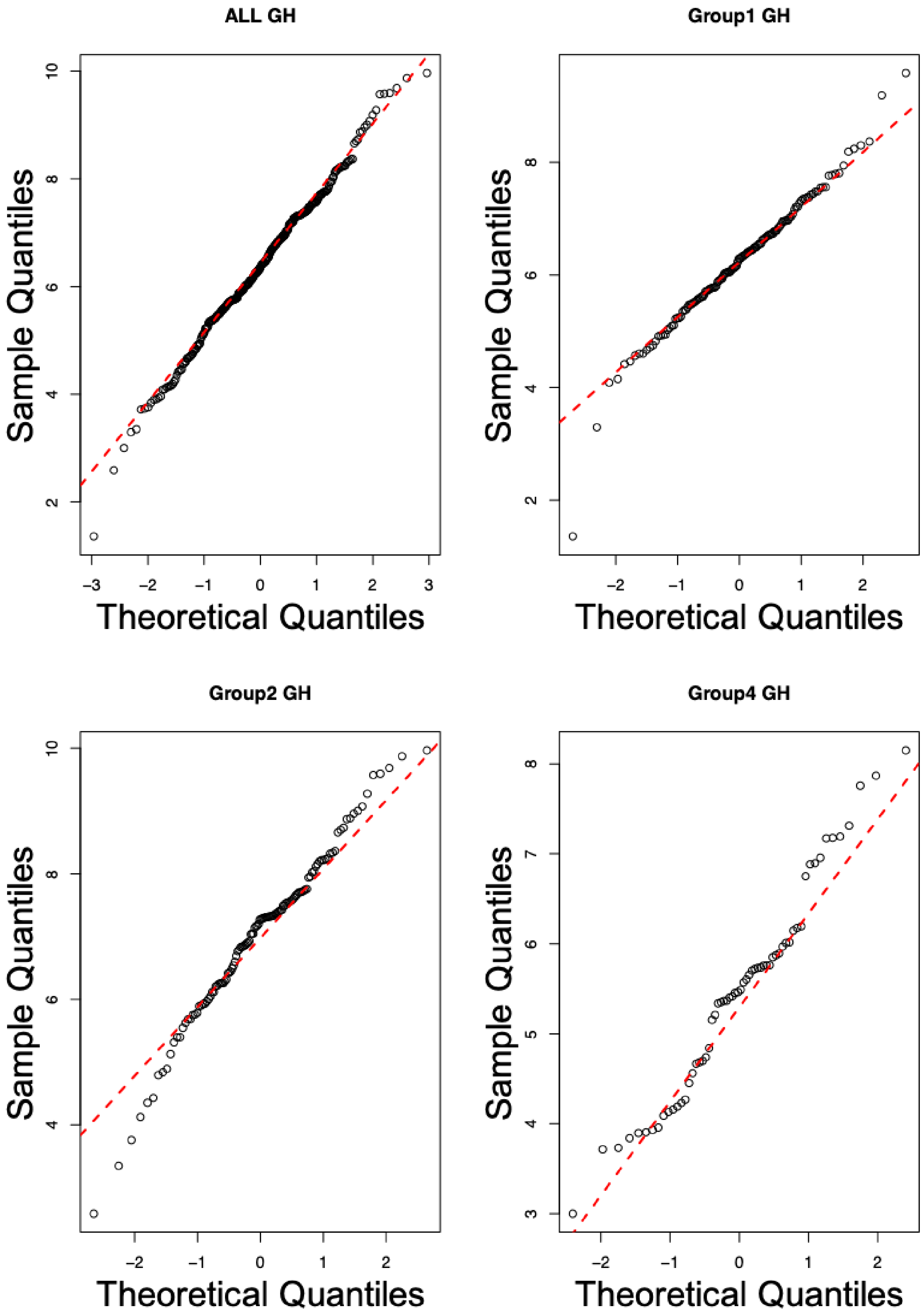


FIGURE D.20: **Q-Q plots for pod length in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

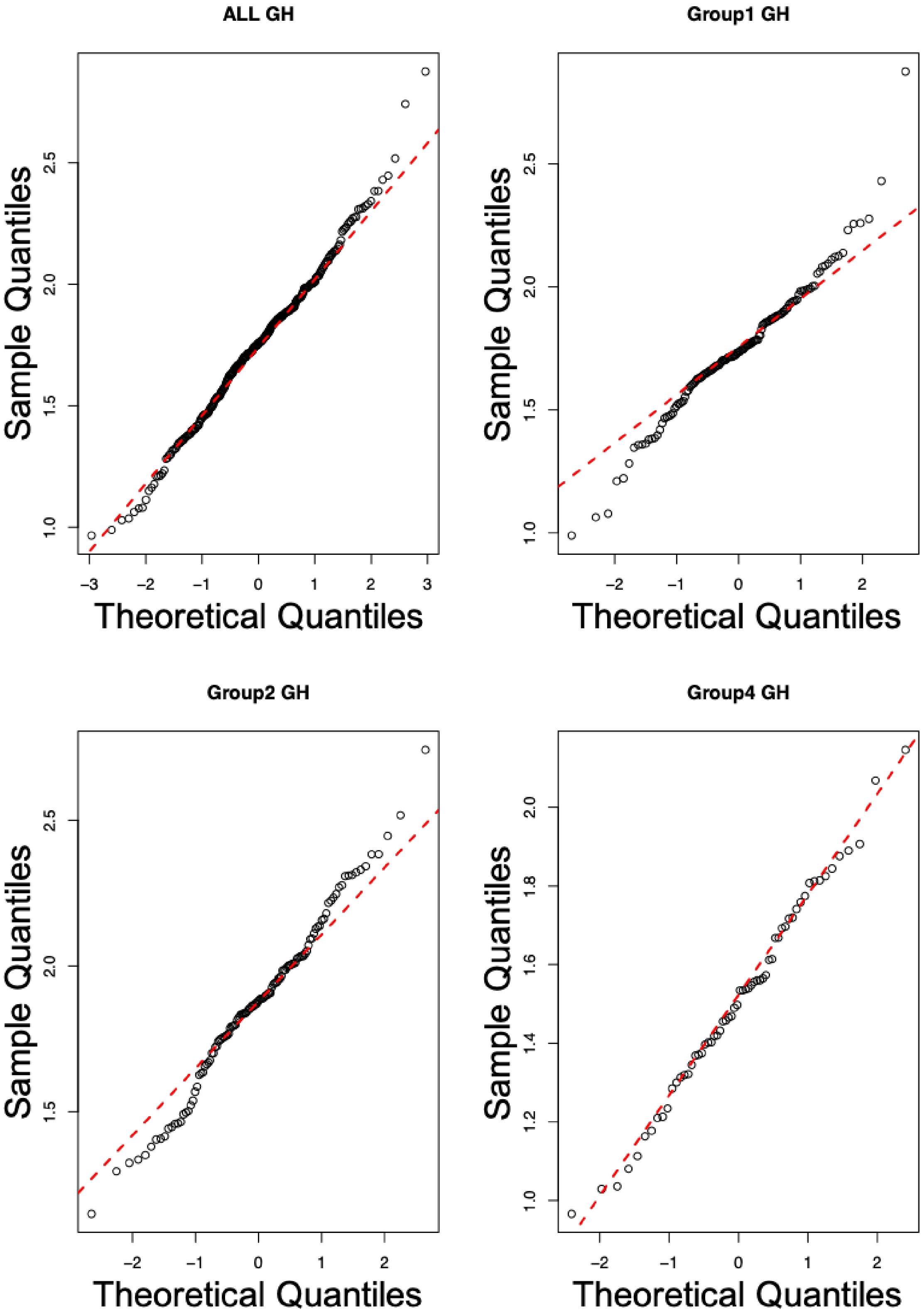


FIGURE D.21: **Q-Q plots for pod width in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

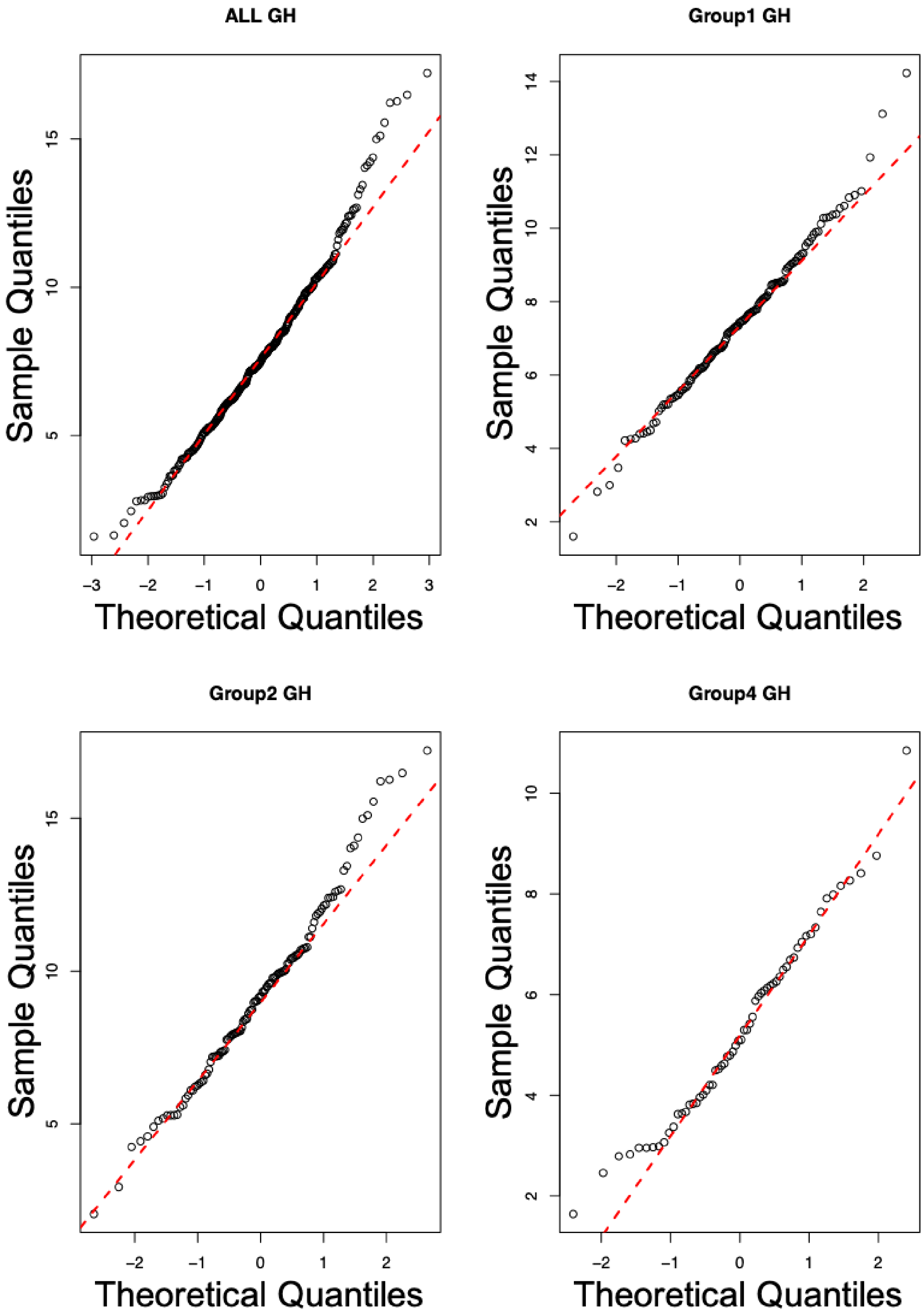


FIGURE D.22: **Q-Q plots for pod area in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

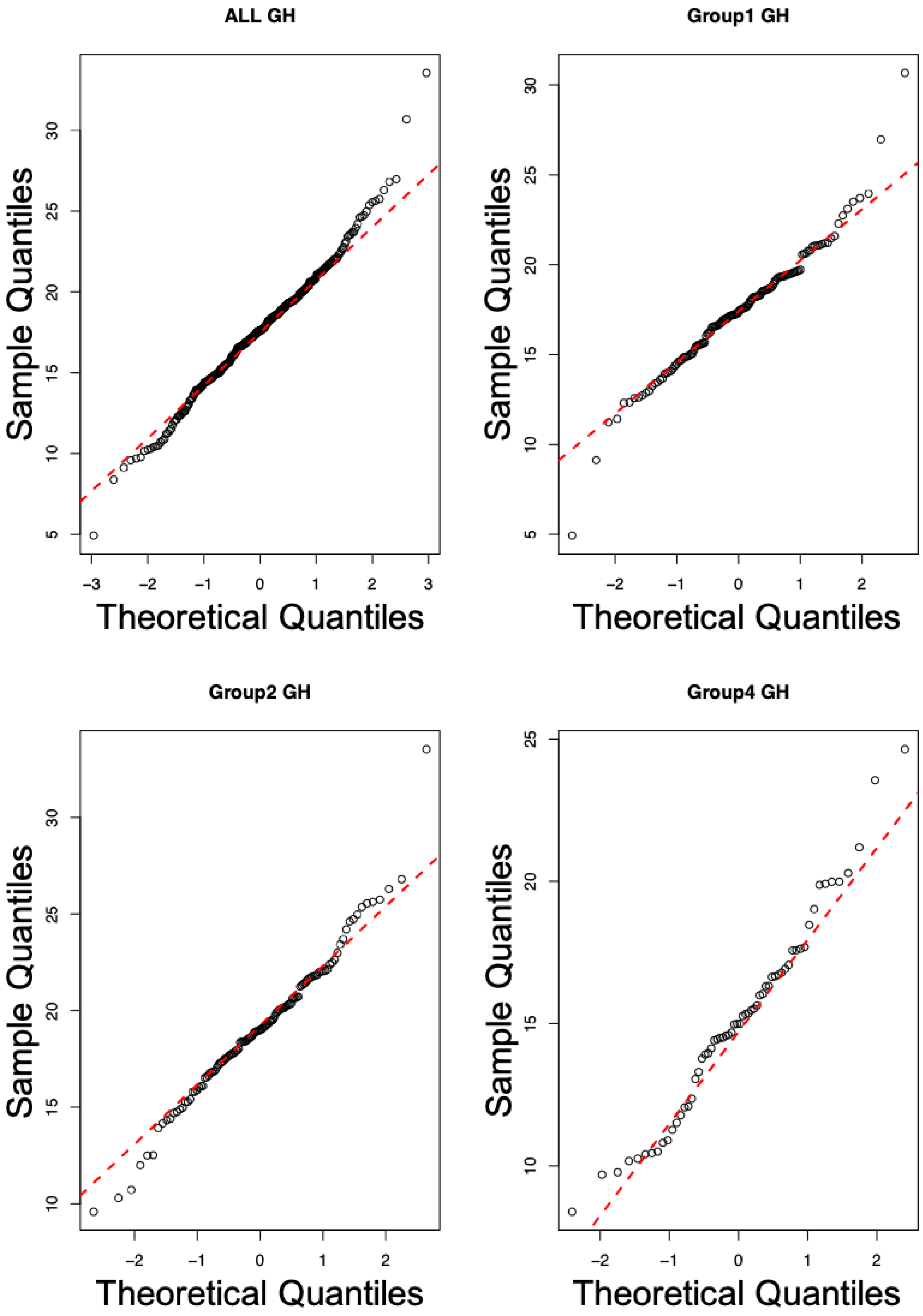


FIGURE D.23: **Q-Q plots for pod perimeter in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

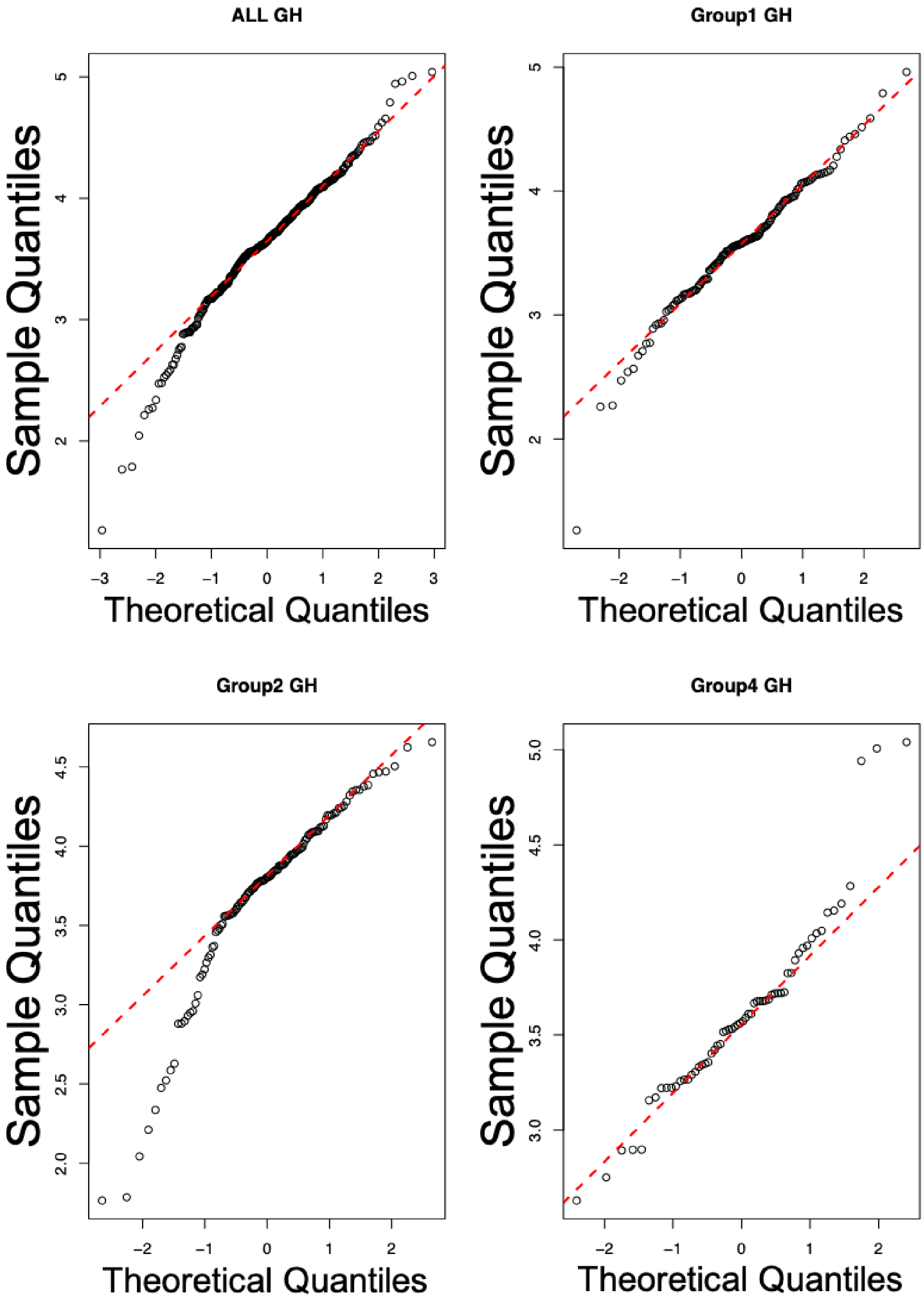


FIGURE D.24: **Q-Q plots for pod aspect ratio in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

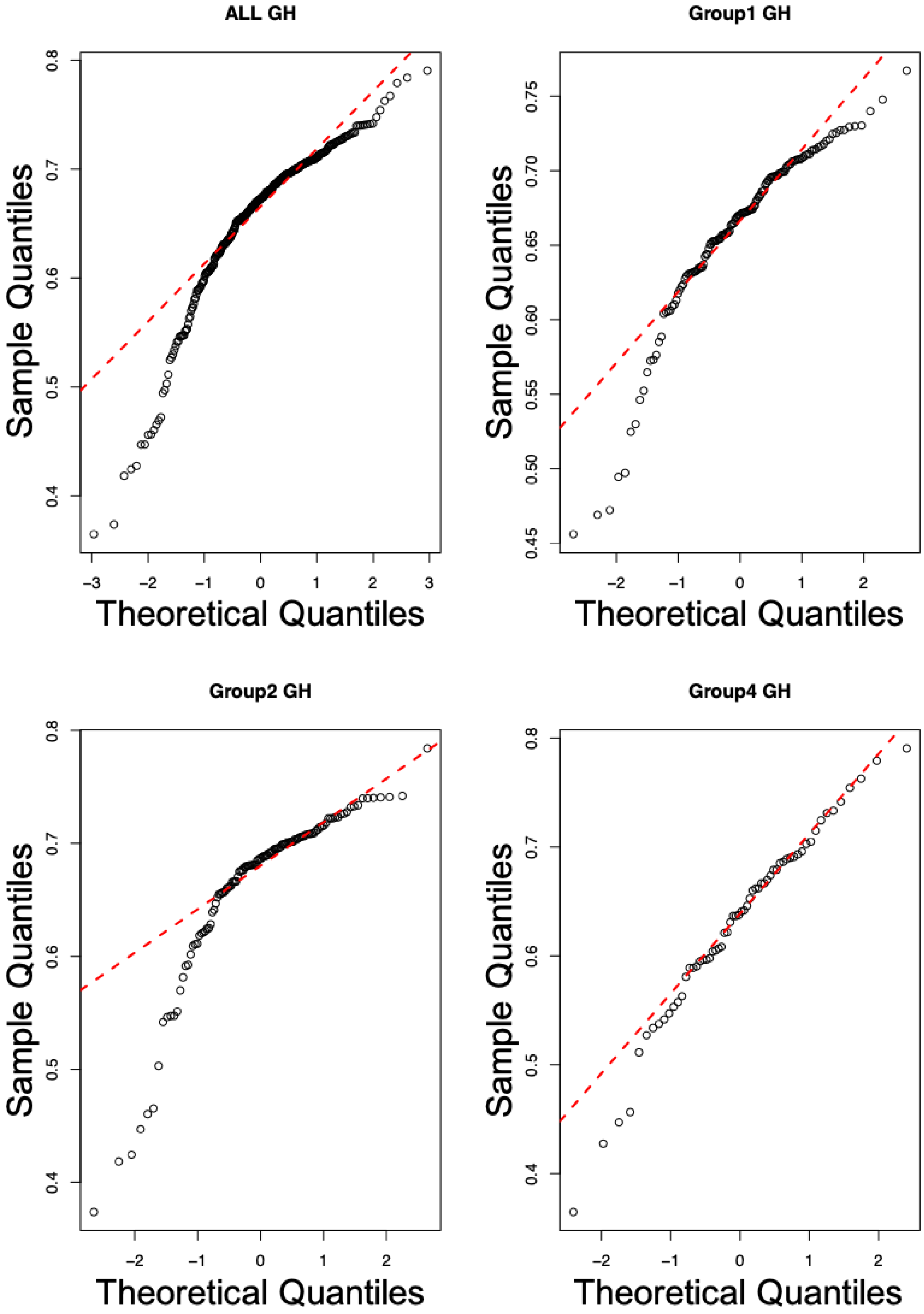


FIGURE D.25: **Q-Q plots for pod rectangularity in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

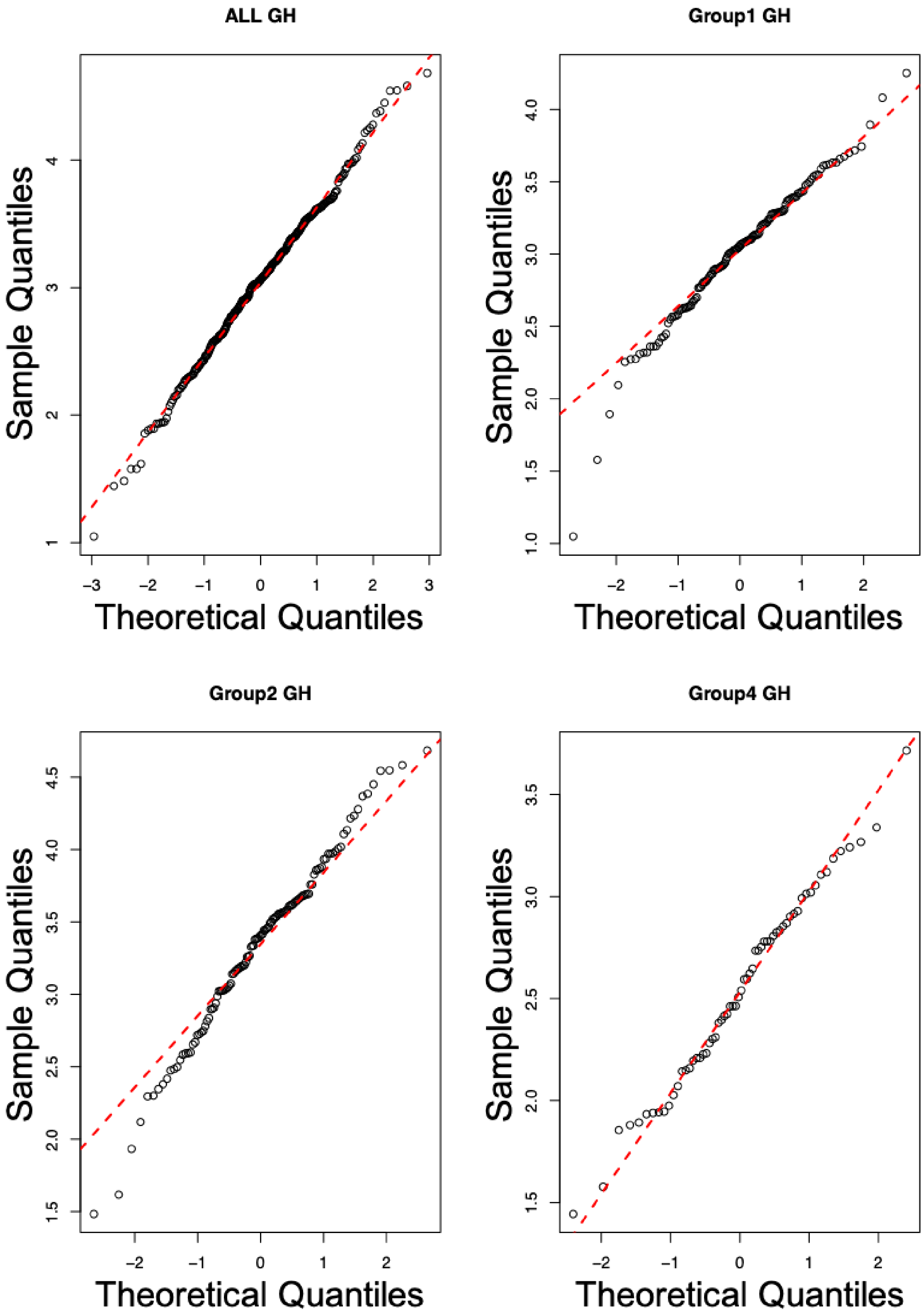


FIGURE D.26: **Q-Q plots for pod equivalent diameter in glasshouse environments.** All GH represents the values for core collection grown in the glasshouse. The core collection was also separated into groups: group 1 (landrace), group 2 (cultivar), group 4 (wild) for which the corresponding Q-Q plots are also shown.

Appendix E

Krukal-Wallis Test

Phenotype	Kruskal-Wallis chi-squared Statistical Test	Degrees of freedom	p-value
Plant height	45.188	df = 2	p-value = 1.54e-10
Seed Weight (GH)	119.71	df = 2	p-value < 2.2e-16
Seed Weight (Field)	72.249	df = 2	p-value < 2.2e-16
Seed Area	126.71	df = 2	p-value < 2.2e-16
Seed Perimeter	133.61	df = 2	p-value < 2.2e-16
Seed Equivalent Diameter	124.24	df = 2	p-value < 2.2e-16
Seed Eccentricity	108.68	df = 2	p-value < 2.2e-16
Leaflet Length	37.141	df = 2	p-value = 8.607e-09
Leaflet Width	55.15	df = 2	p-value = 1.058e-12
Leaflet Perimeter	44.876	df = 2	p-value = 1.8e-10
Leaflet Area	52.863	df = 2	p-value = 3.319e-12
Leaflet Teeth	2.9946	df = 2	p-value = 0.2237
Leaflet Aspect Ratio	29.396	df = 2	p-value = 4.137e-07
Leaflet Roundness	32.987	df = 2	p-value = 6.87e-08
Leaflet Compactness	32.509	df = 2	p-value = 8.726e-08
Leaflet Rectangularity	25.069	df = 2	p-value = 3.599e-06

Continued on next page

Table E.1 – *Continued from previous page*

Phenotype	Kruskal-Wallis Statistical Test	Degrees of freedom	p-value
Leaflet Perimeter Ratio of Length	25.914	df = 2	p-value = 2.359e-06
Leaflet Perimeter Ratio of Length and Width	15.146	df = 2	p-value = 0.0005141
Leaflet Equivalent Diameter	50.969	df = 2	p-value = 8.555e-12
Pod Length	66.675	df = 2	p-value = 3.325e-15
Pod Width	63.879	df = 2	p-value = 1.346e-14
Pod Area	88.623	df = 2	p-value < 2.2e-16
Pod Perimeter	54.124	df = 2	p-value = 1.767e-12
Pod Aspect Ratio	13.741	df = 2	p-value = 0.001038
Pod Rectangularity	14.325	df = 2	p-value = 0.0007753
Pod Equiavalent Diameter	77.493	df = 2	p-value < 2.2e-16

Appendix F

Nemenyi Test

Phenotype	Group 1 vs Group 2	Group 1 vs Group 4	Group 2 vs Group 4
Plant Height	2.10E-10	0.138	0.003
Seed Weight (GH)	2.50E-05	3.50E-12	2.00E-16
Seed Weight (Field)	0.031	4.40E-10	2.20E-16
Seed Area	7.20E-08	4.40E-10	< 2e-16
Seed Perimeter	3.80E-10	1.10E-08	< 2e-16
Seed Equivalent Diameter	8.10E-08	8.40E-10	< 2e-16
Seed Eccentricity	1.80E-05	2.40E-10	< 2e-16
Leaflet Length	5.0E-06	6.8E-06	0.42
Leaflet Width	6.7e-07	5.8e-10	0.031
Leaflet Perimeter	2.30E-06	1.10E-07	0.14
Leaflet Area	3.30E-07	4.30E-09	0.081
Leaflet Teeth	0.83	0.23	0.46
Leaflet Aspect Ratio	0.18211	4.30E-07	0.00041
Leaflet Roundness	0.0364	1.10E-07	0.0011
Leaflet Compactness	0.0384	1.30E-07	0.0012
Leaflet Rectangularity	0.98	8.60E-06	3.60E-05

Continued on next page

Table F.1 – *Continued from previous page*

Phenotype	Group 1 vs Group 2	Group 1 vs Group 4	Group 2 vs Group 4
Leaflet Perimeter Ratio of Length	0.5458	2.60E-06	0.0002
Leaflet Perimeter Ratio of Length and Width	0.00052	0.40414	0.39784
Leaflet Equivalent Diameter	6.40E-07	7.60E-09	0.082
Pod Length	1.70E-06	0.00044	2.50E-14
Pod Width	6.00E-04	3.10E-06	1.60E-14
Pod Area	5.80E-06	1.60E-07	< 2e-16
Pod Perimeter	0.00025	0.00019	3.30E-12
Pod Aspect Ratio	0.0027	0.9999	0.0259
Pod Rectangularity	0.17693	0.06282	0.00081
Pod Equivalent Diameter	6.60E-05	4.20E-07	< 2e-16

Appendix G

Descriptive Statistics

TABLE G.1: **Descriptive Statistics of Simple and Morphological Shape Descriptors (SMSDs).** The counts, mean, standard deviation, median and interquartile range (IQR) for all SMSDs found in Group 1 (landrace), Group 2 (cultivar), Group 4 (wild) or ALL (the total)

Phenotype	Group	Count	Mean	SD	Median	IQR
Height	ALL	349	101.6617	45.03017	97	70
	1	140	119.306	47.47152	125.3333	70.41667
	2	140	83.48259	41.58029	70.41667	50.25
	4	69	102.7641	29.96182	101.3333	35.66667
Seed Weight GH	ALL	350	0.192716	0.07538575	0.1923148	0.09301737
	1	140	0.1930383	0.06307631	0.1865608	0.07679631
	2	140	0.2334436	0.06229029	0.2279167	0.08298215
	4	70	0.1073477	0.04863237	0.1001003	0.05842425
Seed Weight Field	ALL	350	0.1575172	0.06517647	0.152492	0.08160069
	1	140	0.1601222	0.0539977	0.1494739	0.05952953
	2	140	0.1802093	0.06308302	0.1808365	0.07707097
	4	70	0.08663223	0.04825522	0.07251977	0.07251977
Seed Area	ALL	340	0.3395272	0.1418648	0.3473739	0.1633647
	1	138	0.3377982	0.1251948	0.3297473	0.1156367
	2	135	0.4164767	0.1063475	0.4046409	0.118042

Continued on next page

Table G.1 – *Continued from previous page*

Phenotype	Group	Count	Mean	SD	Median	IQR
	4	67	0.1880409	0.1116316	0.1960041	0.1150539
Seed Perimeter	ALL	340	2.306538	0.6352842	2.372134	0.6320637
	1	138	2.310844	0.4623427	2.290771	0.4770546
	2	135	2.662174	0.3517457	2.669139	0.5081955
	4	67	1.58109	0.7648919	1.767743	0.7211742
Seed Equivalent Diameter	ALL	340	0.6303272	0.170212	0.6606844	0.1672944
	1	138	0.6385117	0.1239483	0.642587	0.1181776
	2	135	0.7178903	0.0969319	0.7158091	0.1056146
	4	67	0.4370365	0.2099865	0.4964682	0.1546896
Seed Eccentricity	ALL	340	0.4956519	0.1285652	0.4990146	0.1301443
	1	138	0.5060408	0.09407723	0.4918597	0.1012998
	2	135	0.5536698	0.07876767	0.5578927	0.1070021
	4	67	0.3573519	0.1655394	0.4060237	0.08263661
Leaflet length	ALL	223	3.7933	1.010261	3.729015	1.310698
	1	112	4.197308	0.8950867	4.184797	1.211572
	2	81	3.453145	0.9229727	3.400768	1.422633
	4	30	3.203417	1.040659	3.078598	1.453989
Leaflet width	ALL	223	2.539934	0.7811846	2.492383	0.9998918
	1	112	2.894399	0.6968268	2.828536	0.8357433
	2	81	2.290021	0.6932103	2.370641	0.9456458
	4	30	1.891361	0.626525	1.760314	0.804624
Leaflet Perimeter	ALL	223	10.49365	2.874994	10.52767	3.676025
	1	112	11.72799	2.55962	11.55684	3.42958
	2	81	9.543567	2.58432	9.609061	3.569148
	4	30	8.450697	2.655262	8.214596	3.785281
Leaflet Area	ALL	223	7.487108	3.966487	7.062374	4.927222

Continued on next page

Table G.1 – *Continued from previous page*

Phenotype	Group	Count	Mean	SD	Median	IQR
	1	112	9.242943	3.898639	8.738775	4.899303
	2	81	6.122271	3.120154	5.861929	4.109146
	4	30	4.617054	3.126513	3.92825	3.265648
Leaflet Teeth	ALL	223	0.6158445	1.318767	0	0.6666667
	1	112	0.7217262	1.644506	0.3333333	0.6666667
	2	81	0.563786	0.9115646	0	1
	4	30	0.3611111	0.7428888	0	0.3333333
Leaflet Aspect Ratio	ALL	223	1.541383	0.3358005	1.492411	0.2589275
	1	112	1.474724	0.1789601	1.447266	0.1846169
	2	81	1.571256	0.4825011	1.49905	0.2606887
	4	30	1.709588	0.2224387	1.672739	0.2747515
Leaflet Roundness	ALL	223	0.7761109	0.05280355	0.7883052	0.05382625
	1	112	0.7895539	0.03886004	0.7962328	0.03542969
	2	81	0.7720357	0.06280999	0.785823	0.05306773
	4	30	0.7369268	0.04846176	0.7340433	0.0684025
Leaflet Compactness	ALL	223	16.31529	1.793406	15.95721	1.113158
	1	112	15.96744	0.8722301	15.79173	0.7211109
	2	81	16.48801	2.62627	16.00828	1.11888
	4	30	17.14762	1.189571	17.13139	1.611139
Leaflet Rectangularity	ALL	223	0.7116884	0.01826261	0.7139528	0.0238297
	1	112	0.7144389	0.01571116	0.7151364	0.02197071
	2	81	0.7138778	0.01907827	0.715747	0.02078981
	4	30	0.6955081	0.01705442	0.6937923	0.02283242
Leaflet PRL	ALL	223	2.764212	0.1404414	2.770766	0.1797985
	1	112	2.79351	0.1216771	2.791189	0.1392979
	2	81	2.765375	0.1520518	2.778152	0.2096449

Continued on next page

Table G.1 – *Continued from previous page*

Phenotype	Group	Count	Mean	SD	Median	IQR
	4	30	2.651694	0.1188527	2.645486	0.1432468
Leaflet PRLW	ALL	223	1.658486	0.0251776	1.655854	0.02147949
	1	112	1.653985	0.02157672	1.651923	0.01966455
	2	81	1.663179	0.02545828	1.660778	0.01634565
	4	30	1.662619	0.03354697	1.65707	0.03133325
Leaflet Equivalent Diameter	ALL	223	2.94607	0.8297495	2.935871	1.048311
	1	112	3.316336	0.7300337	3.264252	0.9650753
	2	81	2.670838	0.7446453	2.683056	1.004821
	4	30	2.30687	0.7329272	2.231887	0.9382998
Pod Length	ALL	328	6.378448	1.331019	6.354885	1.744202
	1	142	6.220613	1.106887	6.286053	1.318042
	2	124	7.023141	1.328764	7.273054	1.480549
	4	62	5.450558	1.151885	5.475701	1.410483
Pod Width	ALL	328	1.750913	0.2990606	1.756436	0.3778736
	1	142	1.745334	0.2617704	1.734555	0.2636212
	2	124	1.877307	0.2854897	1.880917	0.3096008
	4	62	1.510905	0.2561461	1.515805	0.3446736
Pod Area	ALL	328	7.710642	2.765753	7.483741	3.446418
	1	142	7.418753	2.003441	7.428778	2.398479
	2	124	9.249115	2.931418	9.17176	3.471416
	4	62	5.302216	1.869592	5.091593	2.70084
Pod Perimeter	ALL	328	17.61925	3.688523	17.58571	4.403312
	1	142	17.40041	3.234318	17.36265	3.83328
	2	124	19.12623	3.5701	19.00142	4.151366
	4	62	15.10652	3.450713	14.99187	4.358717
Pod Aspect Ratio	ALL	328	3.622468	0.5271495	3.643844	0.611265

Continued on next page

Table G.1 – *Continued from previous page*

Phenotype	Group	Count	Mean	SD	Median	IQR
	1	142	3.558139	0.5151368	3.575816	0.6467033
	2	124	3.702345	0.5597699	3.802029	0.5125256
	4	62	3.610046	0.4708356	3.566099	0.4874038
	ALL	328	0.6560123	0.06962916	0.6728149	0.07126385
Pod	1	142	0.6596822	0.05659797	0.6702765	0.06451771
Rectangularity	2	124	0.6646704	0.07200275	0.6866435	0.05193345
	4	62	0.630291	0.0852952	0.6390966	0.09896475
Pod	ALL	328	3.044983	0.5953199	3.063194	0.7928261
Equiva-	1	142	3.014299	0.4627505	3.058968	0.5272706
lent	2	124	3.335002	0.6073802	3.405743	0.6668223
Diameter	4	62	2.535223	0.4729872	2.523249	0.6673771

Appendix H

Principle Components Analysis

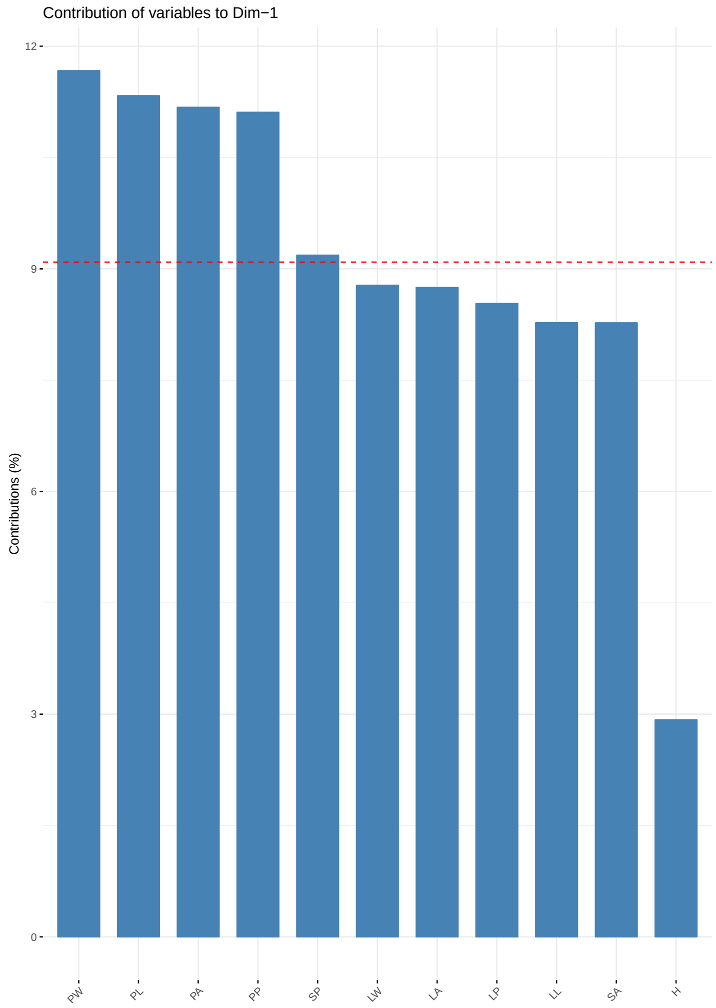


FIGURE H.1: **Contribution of phenotypes to Dimension 1.**

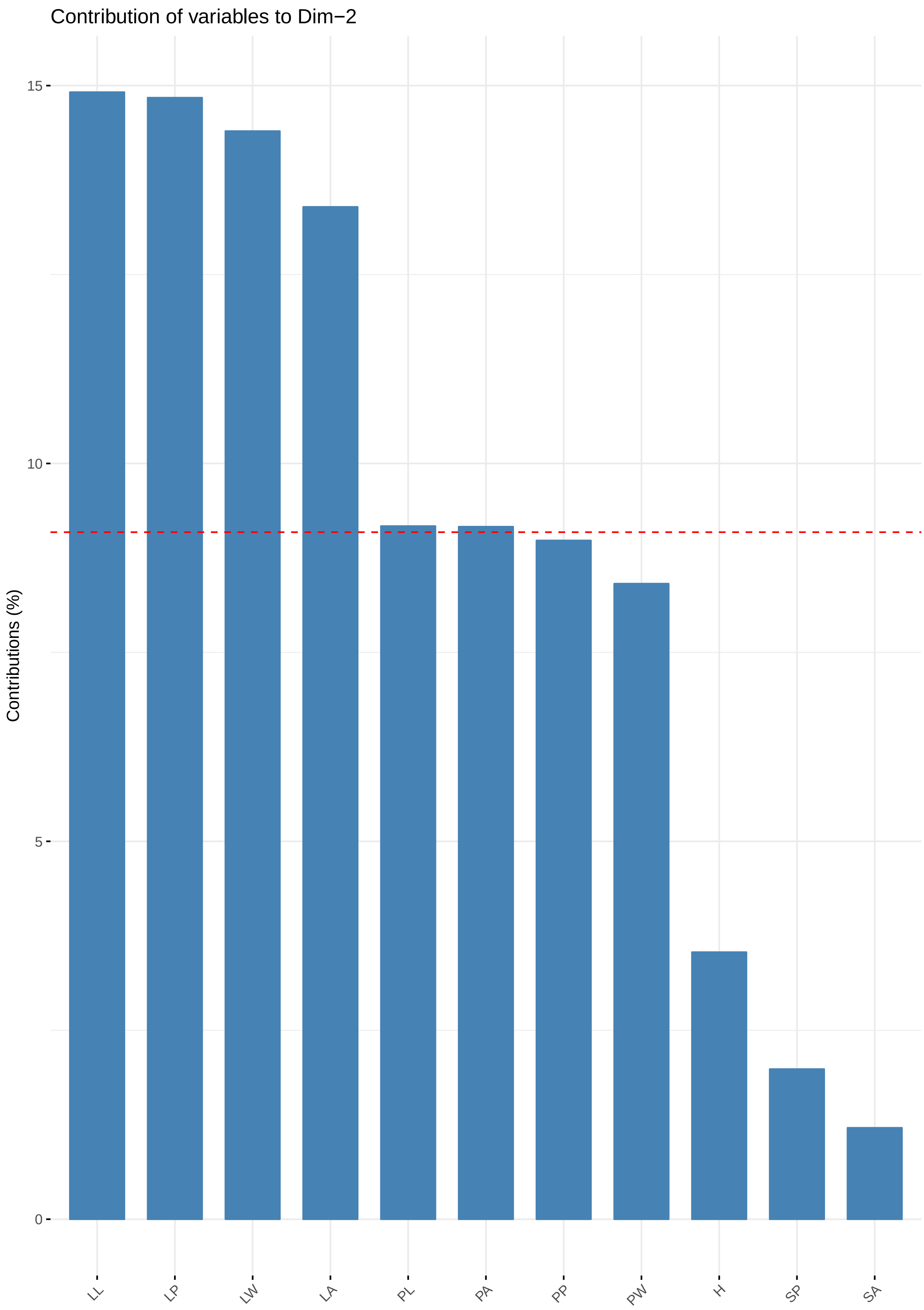


FIGURE H.2: **Contribution of phenotypes to Dimension 2.**

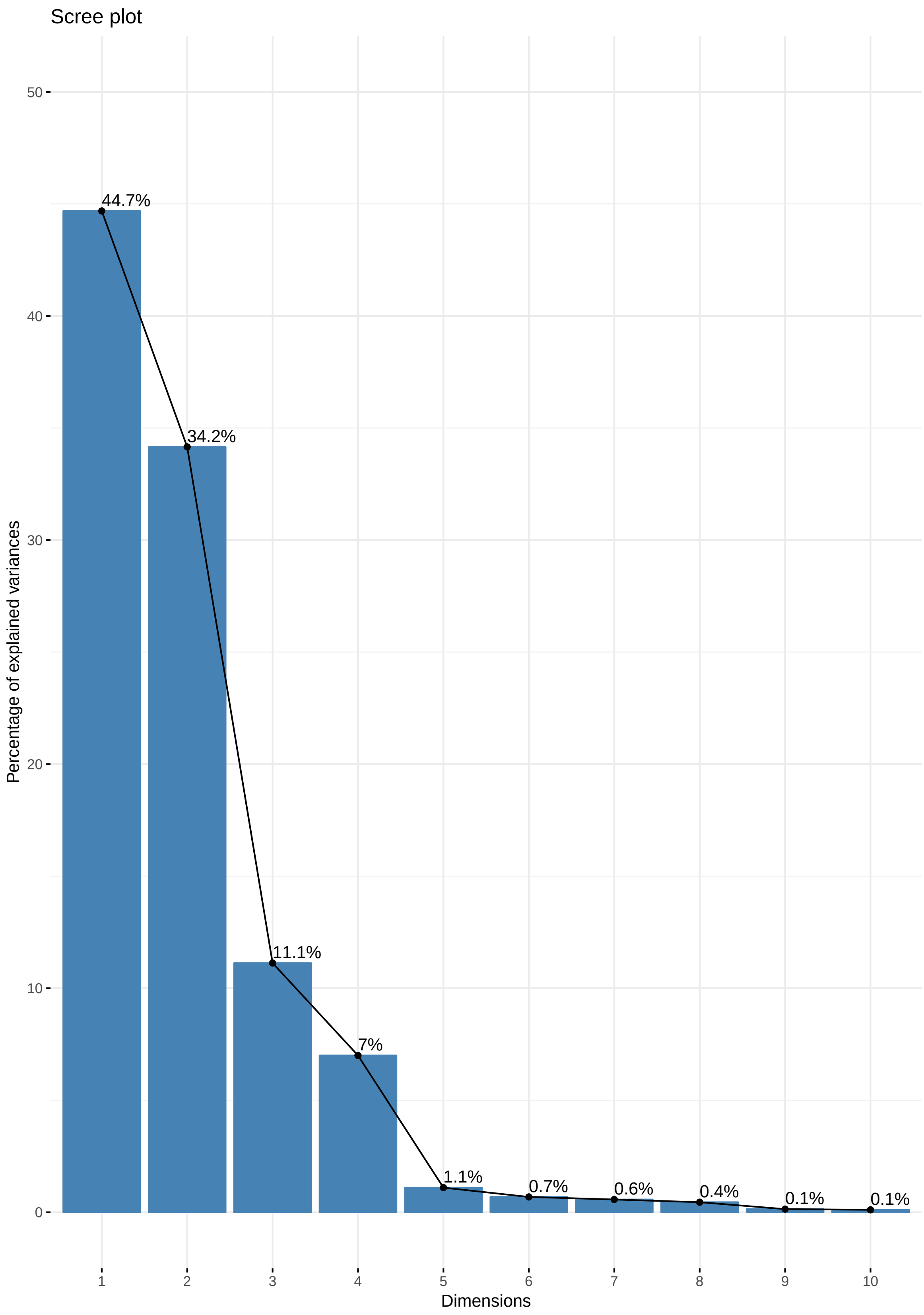


FIGURE H.3: **Scree plot of all dimensions.**

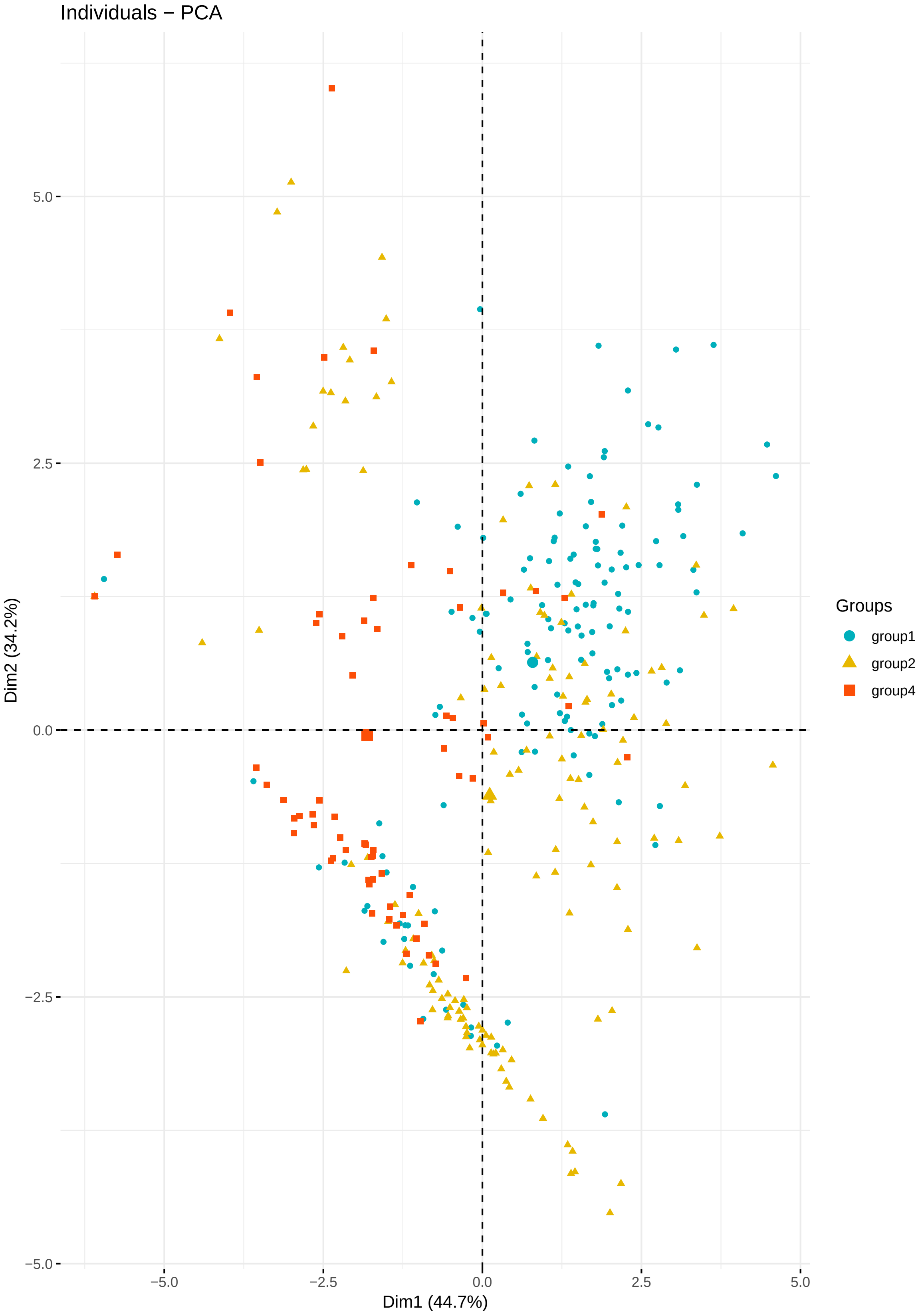


FIGURE H.4: **Principle Components Analysis of all individuals.** Original in colour.

Bibliography

Abawi, GS and TL Widmer (2000). “Impact of soil health management practices on soilborne pathogens, nematodes and root diseases of vegetable crops”. In: *Applied soil ecology* 15.1, pp. 37–47.

Abi-Ghanem, R et al. (2013). “Potential breeding for high nitrogen fixation in Pisum sativum L.: germplasm phenotypic characterization and genetic investigation”. In: *American Journal of Plant Sciences* 2013.

ABySS - Stats. https://github.com/bcgsc/abyss/wiki/ABySS-File-Formats#stats. Accessed: 10th December 2018.

Aho, Alfred V, Brian W Kernighan, and Peter J Weinberger (1979). “Awk—a pattern scanning and processing language”. In: *Software: Practice and Experience* 9.4, pp. 267–279.

Al-Amri, Salem Saleh, Namdeo V Kalyankar, et al. (2010). “Image segmentation by using threshold techniques”. In: *arXiv preprint arXiv:1005.4020*.

Altschul, Stephen F et al. (1990). “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3, pp. 403–410.

Anand, Santosh et al. (2016). “Next generation sequencing of pooled samples: guideline for variants’ filtering”. In: *Scientific reports* 6, p. 33735.

Andrews, Kimberly R et al. (2016). “Harnessing the power of RADseq for ecological and evolutionary genomics”. In: *Nature Reviews Genetics* 17.2, pp. 81–92.

Andrews, Simon. *FastQC: a quality control tool for high throughput sequence data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed: 28th August 2018.

Annicchiarico, Paolo et al. (2017). "GBS-based genomic selection for pea grain yield under severe terminal drought". In: *The Plant Genome*.

Apelt, Federico et al. (2015). "Phytotyping4D: a light-field imaging system for non-invasive and accurate monitoring of spatio-temporal plant growth". In: *The Plant Journal* 82.4, pp. 693–706.

Appleby, Cyril A (1984). "Leghemoglobin and Rhizobium respiration". In: *Annual Review of Plant Physiology* 35.1, pp. 443–478.

Arnaud, Elizabeth et al. (2012). "Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes." In: *KEOD*, pp. 220–225.

Arnold, B et al. (2013). "RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling". In: *Molecular Ecology* 22.11, pp. 3179–3190. ISSN: 1365-294X.

Assess 2.0: Image Analysis Software for Plant Disease Quantification.

Aubert, G et al. (2006). "Functional mapping in pea, as an aid to the candidate gene selection and for investigating synteny with the model legume Medicago truncatula". In: *Theoretical and Applied Genetics* 112.6, pp. 1024–1041. ISSN: 0040-5752.

Aulchenko, Yurii S et al. (2007). "GenABEL: an R library for genome-wide association analysis". In: *Bioinformatics* 23.10, pp. 1294–1296. ISSN: 1367-4803.

Ayling, Sarah C (2012). "Technical appraisal of strategic approaches to large-scale germplasm evaluation". In:

Bacchetta, Gianluigi et al. (2008). "Morpho-colorimetric characterization by image analysis to identify diaspores of wild plant species". In: *Flora-Morphology, Distribution, Functional Ecology of Plants* 203.8, pp. 669–682.

Backhaus, Andreas et al. (2010). "LEAFPROCESSOR: a new leaf phenotyping tool using contour bending energy and shape cluster analysis". In: *New phytologist* 187.1, pp. 251–261.

Bai, Fang and Darleen A DeMason (2006). "Hormone interactions and regulation of Unifoliata, PsPK2, PsPIN1 and LE gene expression in pea (Pisum sativum) shoot tips". In: *Plant and cell physiology* 47.7, pp. 935–948.

Bakr, EM (2005). “A new software for measuring leaf area, and area damaged by Tetranychus urticae Koch”. In: *Journal of applied Entomology* 129.3, pp. 173–175.

Balfourier, François et al. (2007). “A worldwide bread wheat core collection arrayed in a 384-well plate”. In: *Theoretical and Applied Genetics* 114.7, pp. 1265–1275.

Bari, Abdallah et al. (2012). “Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables”. In: *Genetic Resources and Crop Evolution* 59.7, pp. 1465–1481. ISSN: 0925-9864.

Bateson, Mr W (2004). “PROBLEMS OF HEREDITY AS A SUBJECT FOR HORTICULTURAL INVESTIGATION.” In: *A Century of Mendelism in Human Genetics*, p. 153.

BBSRC (2015). *Food Security - Introduction*. Web Page. URL: http://www.bbsrc.ac.uk/research/topical/food/food-security-introduction/.

Begum, Hasina et al. (2015). “Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa)”. In: *PloS one* 10.3, e0119873.

Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Benner, Christian et al. (2016). “FINEMAP: efficient variable selection using summary data from genome-wide association studies”. In: *Bioinformatics* 32.10, pp. 1493–1501.

Bhattacharyya, Madan K et al. (1990). “The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme”. In: *Cell* 60.1, pp. 115–122.

Billiau, Kenny et al. (2012). “Data management pipeline for plant phenotyping in a multisite project”. In: *Functional Plant Biology* 39.11, pp. 948–957. ISSN: 1445-4416.

Biot, Eric et al. (2016). “Multiscale quantification of morphodynamics: MorphoLeaf, software for 2-D shape analysis”. In: *Development*, dev–134619.

Boetzer, Marten et al. (2010). “Scaffolding pre-assembled contigs using SSPACE”. In: *Bioinformatics* 27.4, pp. 578–579.

Bogracheva, T Ya et al. (1999). "The effect of mutant genes at the r, rb, rug3, rug4, rug5 and lam loci on the granular structure and physico-chemical properties of pea seed starch". In: *Carbohydrate Polymers* 39.4, pp. 303–314.

Bordat, Amandine et al. (2011). "Translational genomics in legumes allowed placing in silico 5460 unigenes on the pea functional map and identified candidate genes in Pisum sativum L". In: *G3: Genes, Genomes, Genetics* 1.2, pp. 93–103. ISSN: 2160-1836.

Borianne, Philippe and Guilhem Brunel (2012). "Automated valuation of leaves area for large-scale analysis needing data coupling or petioles deletion". In: *Plant Growth Modeling, Simulation, Visualization and Applications (PMA), 2012 IEEE Fourth International Symposium on*. IEEE, pp. 50–57.

Bourgeois, Michael et al. (2009). "Dissecting the proteome of pea mature seeds reveals the phenotypic plasticity of seed protein composition". In: *Proteomics* 9.2, pp. 254–271.

Boyle, Evan A, Yang I Li, and Jonathan K Pritchard (2017). "An expanded view of complex traits: from polygenic to omnigenic". In: *Cell* 169.7, pp. 1177–1186.

Bradbury, Peter J et al. (2007). "TASSEL: software for association mapping of complex traits in diverse samples". In: *Bioinformatics* 23.19, pp. 2633–2635. ISSN: 1367-4803.

Bradley, Derek and Gerhard Roth (2007). "Adaptive thresholding using the integral image". In: *Journal of graphics tools* 12.2, pp. 13–21.

Brewer, Marin Talbot et al. (2006). "Development of a controlled vocabulary and software application to analyze fruit shape variation in tomato and other plant species". In: *Plant physiology* 141.1, pp. 15–25.

Brewin, Nicholas J (2004). "Plant cell wall remodelling in the Rhizobium–legume symbiosis". In: *Critical Reviews in Plant Sciences* 23.4, pp. 293–316.

Brien, Chris J et al. (2013). "Accounting for variation in designing greenhouse experiments with special reference to greenhouses containing plants on conveyor systems". In: *Plant Methods* 9.1, p. 5.

Brown, TA (2002). *Genomes*. Wiley-Liss.

Brown, Terence A et al. (2009). "The complex origins of domesticated crops in the Fertile Crescent". In: *Trends in Ecology & Evolution* 24.2, pp. 103–109.

Browning, Brian L and Sharon R Browning (2009). “A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals”. In: *The American Journal of Human Genetics* 84.2, pp. 210–223. ISSN: 0002-9297.

Buenrostro, Jason D et al. (2015). “ATAC-seq: a method for assaying chromatin accessibility genome-wide”. In: *Current protocols in molecular biology* 109.1, pp. 21–29.

Bühler, Jonas et al. (2015). “phenoVein-A tool for leaf vein segmentation and analysis”. In: *Plant physiology*, pp–00974.

Burie, Jean-Christophe et al. (2015). “ICDAR2015 competition on smartphone document capture and OCR (SmartDoc)”. In: *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, pp. 1161–1165.

Burstin, Judith et al. (2015). “Genetic diversity and trait genomic prediction in a pea diversity panel”. In: *BMC genomics* 16.1, p. 105. ISSN: 1471-2164.

Bush, William S and Jason H Moore (2012). “Genome-wide association studies”. In: *PLoS computational biology* 8.12, e1002822. ISSN: 1553-7358.

Butler, Jonathan et al. (2008). “ALLPATHS: de novo assembly of whole-genome shotgun microreads”. In: *Genome research*.

Bylesjö, Max et al. (2008). “LAMINA: a tool for rapid quantification of leaf size and shape parameters”. In: *BMC plant biology* 8.1, p. 82.

Caglayan, Ali, Oguzhan Guclu, and Ahmet Burak Can (2013). “A plant recognition approach using shape and color features in leaf images”. In: *International Conference on Image Analysis and Processing*. Springer, pp. 161–170.

Campbell, Michael S et al. (2014). “Genome annotation and curation using MAKER and MAKER-P”. In: *Current Protocols in Bioinformatics* 48.1, pp. 4–11.

Canfield, Donald E, Alexander N Glazer, and Paul G Falkowski (2010). “The evolution and future of Earth’s nitrogen cycle”. In: *science* 330.6001, pp. 192–196.

Canny, John (1986). “A computational approach to edge detection”. In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 679–698.

Cardon, Lon R and Lyle J Palmer (2003). “Population stratification and spurious allelic association”. In: *The Lancet* 361.9357, pp. 598–604.

Carmeliet, Peter and Rakesh K Jain (2000). “Angiogenesis in cancer and other diseases”. In: *nature* 407.6801, p. 249.

Carpenter, Margaret A et al. (2017). “Association mapping of starch chain length distribution and amylose content in pea (Pisum sativum L.) using carbohydrate metabolism candidate genes”. In: *BMC plant biology* 17.1, p. 132.

Carvalho, Miguel AA Pinheiro de et al. (2013). “Cereal landraces genetic resources in worldwide GeneBanks. A review”. In: *Agronomy for sustainable development* 33.1, pp. 177–203. ISSN: 1774-0746.

Casey, Rod et al. (1998). “The effect of modifying carbohydrate metabolism on seed protein gene expression in peas”. In: *Journal of plant physiology* 152.6, pp. 636–640.

Catchen, Julian. *Stacks denovo map*. http://catchenlab.life.illinois.edu/stacks/comp/denovo_map.php. Accessed: 30th September 2018.

Catchen, Julian M et al. (2011). “Stacks: building and genotyping loci de novo from short-read sequences”. In: *G3: Genes, Genomes, Genetics* 1.3, pp. 171–182. ISSN: 2160-1836.

Chang, Wan-Chi et al. (2012). “A bean-free diet increases the risk of all-cause mortality among Taiwanese women: the role of the metabolic syndrome”. In: *Public health nutrition* 15.4, pp. 663–672.

Charmet, Gilles and François Balfourier (1995). “The use of geostatistics for sampling a core collection of perennial ryegrass populations”. In: *Genetic Resources and Crop Evolution* 42.4, pp. 303–309. ISSN: 0925-9864.

Chemistry LibreTexts (2017). https://chem.libretexts.org/Textbook_Maps/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Chemical_Bonding/Fundamentals_of_Chemical_Bonding/Bond_Energies. Accessed: 17th September 2018.

Cheng, P et al. (2014). “Phylogenetic analysis and association mapping for agronomic and quality traits in USDA pea PSP collection”. In: *PHYTOPATHOLOGY*. Vol. 104. 11. AMER PHYTOPATHOLOGICAL SOC 3340 PILOT KNOB ROAD, ST PAUL, MN 55121 USA, pp. 26–26.

Ching, ADA et al. (2002). "SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines". In: *BMC Genetics* 3.1, p. 19. ISSN: 1471-2156.

Chitradevi, B and P Srimathi (2014). "An overview on image processing techniques". In: *International Journal of Innovative Research in Computer and Communication Engineering* 2.11, pp. 6466–6472.

Chong, Zechen, Jue Ruan, and Chung-I Wu (2012). "Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads". In: *Bioinformatics* 28.21, pp. 2732–2737. ISSN: 1367-4803.

Clavijo, Bernardo et al. (2017). "W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data". In: *bioRxiv*, p. 110999.

Coleman, Jonathan RI et al. (2015). "Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray". In: *Briefings in functional genomics* 15.4, pp. 298–304.

Commons, Wikimedia (2010). *HSV color solid cylinder saturation gray*.

Compeau, Phillip EC, Pavel A Pevzner, and Glenn Tesler (2011). "How to apply de Bruijn graphs to genome assembly". In: *Nature biotechnology* 29.11, p. 987.

Cong, Le et al. (2013). "Multiplex genome engineering using CRISPR/Cas systems". In: *Science*, p. 1231143.

Consortium, 1000 Genomes Project et al. (2010). "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319, p. 1061.

Consortium, Plant Ontology (2015). *The cROP (Common Reference Ontologies for Plants) Initiative*. Web Page. URL: http://wiki.plantontology.org/index.php/The_cROP_(Common_Reference_Ontologies_for_Plants)_Initiative.

Cook, Daniel E and Erik C Andersen (2017). "VCF-kit: assorted utilities for the variant call format". In: *Bioinformatics* 33.10, pp. 1581–1582.

Corney, David PA et al. (2012). "Automating digital leaf measurement: the tooth, the whole tooth, and nothing but the tooth". In: *PloS one* 7.8, e42112.

Craig, Josephine et al. (1998). "Mutations in the gene encoding starch synthase II profoundly alter amylopectin structure in pea embryos". In: *The Plant Cell* 10.3, pp. 413–426.

Crowell, Samuel et al. (2014). "High-resolution inflorescence phenotyping using a novel image analysis pipeline, PANorama". In: *Plant physiology*, pp–114.

Culjak, Ivan et al. (2012). "A brief introduction to OpenCV". In: *MIPRO, 2012 proceedings of the 35th international convention*. IEEE, pp. 1725–1730.

Cumbie, Jason S, Sergei A Filichkin, and Molly Megraw (2015). "Improved DNase-seq protocol facilitates high resolution mapping of DNase I hypersensitive sites in roots in Arabidopsis thaliana". In: *Plant Methods* 11.1, p. 42.

Ćwiek-Kupczyńska, Hanna et al. (2016). "Measures for interoperability of phenotypic data: minimum information requirements and formatting". In: *Plant Methods* 12.1, p. 44.

Dahm, Ralf (2008). "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research". In: *Human genetics* 122.6, pp. 565–581.

Danecek, Petr et al. (2011). "The variant call format and VCFtools". In: *Bioinformatics* 27.15, pp. 2156–2158.

Darmadi-Blackberry, Irene et al. (2004). "Legumes: the most important dietary predictor of survival in older people of different ethnicities". In: *Asia Pacific Journal of Clinical Nutrition* 13.2, pp. 217–220.

Darwin, Charles (1859). "On the origin of species by means of natural selection". In: *Murray, London*.

– (1868). *The variation of animals and plants under domestication*. Vol. 2. O. Judd.

Davey, John W et al. (2011). "Genome-wide genetic marker discovery and genotyping using next-generation sequencing". In: *Nature Reviews Genetics* 12.7, pp. 499–510. ISSN: 1471-0056.

De Beukelaer, Herman, Guy F Davenport, and Veerle Fack (2018). "Core Hunter 3: flexible core subset selection". In: *BMC bioinformatics* 19.1, p. 203.

De Beukelaer, Herman et al. (2012). "Core Hunter II: fast core subset selection based on multiple genetic diversity measures using Mixed Replica search". In: *BMC bioinformatics* 13.1, p. 312.

Dell'Acqua, Matteo et al. (2015). "Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays". In: *Genome biology* 16.1, p. 167.

Denisov, Gennady et al. (2008). “Consensus generation and variant detection by Celera Assembler”. In: *Bioinformatics* 24.8, pp. 1035–1040.

Depari, A et al. (2015). “Using smartglasses for utility-meter reading”. In: *Sensors Applications Symposium (SAS), 2015 IEEE*. IEEE, pp. 1–6.

Deulvot, Chrystel et al. (2010). “Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea”. In: *BMC genomics* 11.1, p. 468. ISSN: 1471-2164.

Dijk, Erwin L van et al. (2014). “Ten years of next-generation sequencing technology”. In: *Trends in genetics* 30.9, pp. 418–426. ISSN: 0168-9525.

Dixon, Richard A and Lloyd W Sumner (2003). “Legume natural products: understanding and manipulating complex pathways for human and animal health”. In: *Plant Physiology* 131.3, pp. 878–885.

Doležel, Jaroslav and Johann Greilhuber (2010). “Nuclear genome size: are we getting closer?” In: *Cytometry Part A* 77.7, pp. 635–642.

Dornbusch, Tino and Bruno Andrieu (2010). “Lamina2Shape—An image processing tool for an explicit description of lamina shape tested on winter wheat (Triticum aestivum L.)” In: *Computers and Electronics in Agriculture* 70.1, pp. 217–224.

Doss, Robert P et al. (2000). “Bruchins: insect-derived plant regulators that stimulate neoplasm formation”. In: *Proceedings of the National Academy of Sciences* 97.11, pp. 6218–6223.

Duarte, Jorge et al. (2014). “Transcriptome sequencing for high throughput SNP development and genetic mapping in Pea”. In: *BMC genomics* 15.1, p. 126. ISSN: 1471-2164.

Dumitras, Tudor et al. (2006). “Eye of the Beholder: Phone-based text-recognition for the visually-impaired”. In: *Wearable Computers, 2006 10th IEEE International Symposium on*. IEEE, pp. 145–146.

Dunn, Peter M (2003). “Gregor Mendel, OSA (1822–1884), founder of scientific genetics”. In: *Archives of Disease in Childhood-Fetal and Neonatal Edition* 88.6, F537–F539.

Durso, Neil A and Richard J Cyr (1994). “A calmodulin-sensitive interaction between microtubules and a higher plant homolog of elongation factor-1 alpha.” In: *The Plant Cell* 6.6, pp. 893–905.

Easlon, Hsien Ming and Arnold J Bloom (2014). "Easy Leaf Area: Automated digital image analysis for rapid and accurate measurement of leaf area". In: *Applications in plant sciences* 2.7.

Eaton, Deren AR (2014). "PyRAD: assembly of de novo RADseq loci for phylogenetic analyses". In: *Bioinformatics*, btu121. ISSN: 1367-4803.

Edgar, Robert C (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic acids research* 32.5, pp. 1792–1797.

Ellinghaus, David et al. (2009). "Current software for genotype imputation". In: *Human genomics* 3.4, p. 371. ISSN: 1479-7364.

Ellis, TH et al. (1992). "Linkage maps in pea." In: *Genetics* 130.3, pp. 649–663.

Elshire, Robert J et al. (2011). "A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species". In: *PloS one* 6.5, e19379. ISSN: 1932-6203.

Elzen, Boelie (1986). "Two ultracentrifuges: A comparative study of the social construction of artefacts". In: *Social Studies of Science* 16.4, pp. 621–662.

Endresen, Dag Terje Filip et al. (2012). "Sources of resistance to stem Rust (Ug99) in bread wheat and durum wheat identified using Focused Identification of Germplasm Strategy". In: *Crop Science* 52.2, pp. 764–773. ISSN: 0011-183X.

Ennos, Roland (2007). *Statistical and Data Handling Skills in Biology*. Pearson Education Limited.

Evanno, Guillaume, Sebastien Regnaut, and Jérôme Goudet (2005). "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study". In: *Molecular ecology* 14.8, pp. 2611–2620.

Falster, Daniel S and Mark Westoby (2003). "Plant height and evolutionary games". In: *Trends in Ecology & Evolution* 18.7, pp. 337–343.

Faroq, AL-Tam et al. (2013). "P-TRAP: a panicle trait phenotyping tool". In: *BMC plant biology* 13.1, p. 122.

Faulkner, Christine et al. (2017). "An automated quantitative image analysis tool for the identification of microtubule patterns in plants". In: *Traffic* 18.10, pp. 683–693.

Felsenstein, J (1989). “PHYLIP (Version 3.6) Phylogeny Inference Package”. In: *Cladistics* 5, pp. 164–166.

Finkers, Richard et al. (2014). “Genebanks and genomics: how to interconnect data from both communities?” In: *Plant Genetic Resources*, pp. 1–4. ISSN: 1479-263X.

Fisher, RA (1960). “The Design of Experiments, Hafner Pub”. In: *Co., NY*.

Fitzgibbon, Jessica et al. (2013). “A developmental framework for complex plasmodesmata formation revealed by large-scale imaging of the Arabidopsis leaf epidermis”. In: *The Plant Cell*, tpc–112.

Food and Agricultural Organisation Statistical Databases of the United Nations. http://www.fao.org/faostat/en/#data/QC/metadata. Accessed: 27th May 2018.

Food and Agricultural Organisation Statistical Databases of the United Nations. http://www.fao.org/faostat/en/#data/QC. Accessed: 27th May 2018.

Fotsis, Theodore et al. (1993). “Genistein, a dietary-derived inhibitor of in vitro angiogenesis”. In: *Proceedings of the National Academy of Sciences* 90.7, pp. 2690–2694.

Fougereux, Jean-Albert et al. (1997). “Water stress during reproductive stages affects seed quality and yield of pea (Pisum sativum L.)” In: *Crop science* 37.4, pp. 1247–1252.

Foyer, Christine H et al. (2016). “Neglecting legumes has compromised human health and sustainable food production”. In: *Nature Plants* 2, p. 16112.

Frankel, OH (1984). “Genetic perspectives of germplasm conservation”. In: *Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge*, pp. 161–170.

Franssen, Susanne U et al. (2011). “Comprehensive transcriptome analysis of the highly complex Pisum sativum genome using next generation sequencing”. In: *BMC genomics* 12.1, p. 227. ISSN: 1471-2164.

Freedman, Matthew L et al. (2004). “Assessing the impact of population stratification on genetic association studies”. In: *Nature genetics* 36.4, p. 388.

Galton, Francis (1871). “Pangenesis”. In: *Nature* 4.79, p. 5.

Gao, Song, Niranjan Nagarajan, and Wing-Kin Sung (2011). “Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences”. In: *International Conference on Research in Computational Molecular Biology*. Springer, pp. 437–451.

Garrison, Erik and Gabor Marth (2012). "Haplotype-based variant detection from short-read sequencing". In: *arXiv preprint arXiv:1207.3907*.

Gehan, Malia A et al. (2017). "PlantCV v2: Image analysis software for high-throughput plant phenotyping". In: *PeerJ* 5, e4088.

Ghahramani, Zoubin (2001). "An introduction to hidden Markov models and Bayesian networks". In: *International journal of pattern recognition and artificial intelligence* 15.01, pp. 9–42.

Graham, Peter H and Carroll P Vance (2003). "Legumes: importance and constraints to greater use". In: *Plant physiology* 131.3, pp. 872–877.

Green, Jason M et al. (2012). "PhenoPhyte: a flexible affordable method to quantify 2D phenotypes from imagery". In: *Plant Methods* 8.1, p. 45.

Gresshoff, Peter M et al. (2015). "The value of biodiversity in legume symbiotic nitrogen fixation and nodulation for biofuel and food production". In: *Journal of plant physiology* 172, pp. 128–136.

Groth, JV and AP Roelfs (1987). "The concept and measurement of phenotypic diversity in Puccinia graminis on wheat." In: *Phytopathology* 77.10, pp. 1395–1399.

Hammer, P Allen and Douglass A Hopper (1997). "Experimental design". In: *Plant growth chamber handbook?.(Eds RW Langhans, TW Tibbitts) pp*, pp. 177–187.

Hanahan, Douglas and Robert A Weinberg (2000). "The hallmarks of cancer". In: *cell* 100.1, pp. 57–70.

Hanmandlu, Madasu, Devendra Jha, and Rochak Sharma (2003). "Color image enhancement by fuzzy intensification". In: *Pattern recognition letters* 24.1-3, pp. 81–87.

Harjes, Carlos E et al. (2008). "Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification". In: *Science* 319.5861, pp. 330–333.

Harrison, Christopher J et al. (2000). "The rug3 locus of pea encodes plastidial phosphoglucomutase". In: *Plant Physiology* 122.4, pp. 1187–1192.

Harrison, Richard G and Erica L Larson (2014). "Hybridization, introgression, and the nature of species boundaries". In: *Journal of Heredity* 105.S1, pp. 795–809.

Hashemi, Zohre et al. (2014). "Cooking enhances beneficial effects of pea seed coat consumption on glucose tolerance, incretin, and pancreatic hormones in high-fat-diet–fed rats". In: *Applied Physiology, Nutrition, and Metabolism* 40.4, pp. 323–333.

Heath, MC and PD Hebblethwaite (1985). "Solar radiation interception by leafless, semi-leafless and leafed peas (Pisum sativum) under contrasting field conditions". In: *Annals of Applied Biology* 107.2, pp. 309–318.

Heinz, Sven et al. (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities". In: *Molecular cell* 38.4, pp. 576–589.

Henry, Amanda G, Alison S Brooks, and Dolores R Piperno (2010). "Microfossils in calculus demonstrate consumption of plants and cooked foods in Neanderthal diets (Shanidar III, Iraq; Spy I and II, Belgium)". In: *Proceedings of the National Academy of Sciences*, p. 201016868.

Herten, Koen et al. (2015). "GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments". In: *BMC bioinformatics* 16.1, p. 73.

Hill, GN et al. (2011). "Image-assisted gap estimation, a simple method for measuring grapevine leaf canopy density". In: *Crop science* 51.6, pp. 2801–2808.

Hintum, TJL van et al. (2000). "Core collections of plant genetic resources. IPGRI Technical Bulletin No. 3". In: *International Plant Genetic Resources Institute, RomeItaly*.

Hofer, Julie MI and TH Noel Ellis (1998). "The genetic control of patterning in pea leaves". In: *Trends in Plant Science* 3.11, pp. 439–444.

Holdsworth, William L et al. (2017). "A community resource for exploring and utilizing genetic diversity in the USDA pea single plant plus collection". In: *Horticulture research* 4, p. 17017.

Horler, RSP et al. (2017). "SeedStor: a germplasm information management system and public database". In: *Plant and Cell Physiology* 59.1, e5–e5.

Huang, Xiaoqiu and Anup Madan (1999). "CAP3: A DNA sequence assembly program". In: *Genome research* 9.9, pp. 868–877.

Huang, Xiaoqiu et al. (2003). "PCAP: a whole-genome assembly program". In: *Genome research* 13.9, pp. 2164–2170.

Huang, Xuehui et al. (2010). "Genome-wide association studies of 14 agronomic traits in rice landraces". In: *Nature genetics* 42.11, pp. 961–967. ISSN: 1061-4036.

Hylton, Christopher and Alison M Smith (1992). "The rb mutation of peas causes structural and regulatory changes in ADP glucose pyrophosphorylase from developing embryos". In: *Plant Physiology* 99.4, pp. 1626–1634.

Hyten, David L et al. (2006). "Impacts of genetic bottlenecks on soybean genome diversity". In: *Proceedings of the National Academy of Sciences* 103.45, pp. 16666–16671.

IBM Corp (2013). "IBM SPSS statistics for windows, version 22.0". In: *Armonk, NY: IBM Corp*.

Inc., Illumina. *HiSeq SBS Kit V2*. https://emea.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/hiseq-rapid-sbs-kit.html. Accessed: 30th September 2018.

– *HiSeq SBS Kit V4*. https://emea.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/hiseq-sbs-v4.html. Accessed: 30th September 2018.

Iqbal, Amjad et al. (2006). "Nutritional quality of important food legumes". In: *Food chemistry* 97.2, pp. 331–335.

Iva, Smykalova et al. (2013). "Phenotypic evaluation of flax seeds by image analysis". In: *Industrial Crops and Products* 47, pp. 232–238.

Jackman, Shaun D et al. (2017). "ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter". In: *Genome research*, gr–214346.

James Hutton Institute. *Potatosize*. https://www.hutton.ac.uk/research/groups/information-and-computational-sciences/potatosize. Accessed: 26th August 2018.

Jensen, Erik Steen and Henrik Hauggaard-Nielsen (2003). "How can increased use of biological N 2 fixation in agriculture benefit the environment?" In: *Plant and Soil* 252.1, pp. 177–186.

Jeong, Seongmun et al. (2017). "GenoCore: A simple and fast algorithm for core subset selection from large genotype datasets". In: *PloS one* 12.7, e0181420.

Jing, Runchun et al. (2007). "Gene-based sequence diversity analysis of field pea (Pisum)". In: *Genetics* 177.4, pp. 2263–2275.

Jing, Runchun et al. (2010). "The genetic diversity and evolution of field pea (Pisum) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis". In: *BMC evolutionary biology* 10.1, p. 44. ISSN: 1471-2148.

Johnson, Randall C et al. (2010). "Accounting for multiple comparisons in a genome-wide association study (GWAS)". In: *BMC genomics* 11.1, p. 724.

Joly, Alexis et al. (2014). "Interactive plant identification based on social image data". In: *Ecological Informatics* 23, pp. 22–34.

Jombart, Thibaut (2008). "adegenet: a R package for the multivariate analysis of genetic markers". In: *Bioinformatics* 24.11, pp. 1403–1405.

Joosen, Ronny VL et al. (2010). "GERMINATOR: a software package for high-throughput scoring and curve fitting of Arabidopsis seed germination". In: *The Plant Journal* 62.1, pp. 148–159.

Kagale, Sateesh et al. (2016). "Analysis of Genotyping-by-Sequencing (GBS) Data". In: *Plant Bioinformatics: Methods and Protocols*, pp. 269–284.

Kalo, P et al. (2004). "Comparative mapping between Medicago sativa and Pisum sativum". In: *Molecular Genetics and Genomics* 272.3, pp. 235–246. ISSN: 1617-4615.

Kaló, Péter et al. (2005). "Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators". In: *Science* 308.5729, pp. 1786–1789.

Kalyoncu, Cem and Önsen Toygar (2015). "Geometric leaf classification". In: *Computer Vision and Image Understanding* 133, pp. 102–109.

Kaur, Er Kavneet and Vijay Kumar Banga (2013). "Number plate recognition using OCR technique". In: *International Journal of Research in Engineering and Technology* 2.09.

Kaur, Sukhjiwan et al. (2012). "Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers". In: *BMC genomics* 13.1, p. 104. ISSN: 1471-2164.

Keilwagen, Jens et al. (2014). "Separating the wheat from the chaff–a strategy to utilize plant genetic resources from ex situ genebanks". In: *Scientific reports* 4.

Kichaev, Gleb et al. (2014). "Integrating functional data to prioritize causal variants in statistical fine-mapping studies". In: *PLoS genetics* 10.10, e1004722.

Kidner, Catherine Anne and Saima Umbreen (2010). "Why is leaf shape so variable". In: *International Journal of Plant Developmental Biology* 4, pp. 64–75.

Kilian, Benjamin and Andreas Graner (2012). "NGS technologies for analyzing germplasm diversity in genebanks". In: *Briefings in functional genomics*, elr046. ISSN: 2041-2649.

Kim, Kyu-Won et al. (2007). "PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets". In: *Bioinformatics* 23.16, pp. 2155–2162.

Knaus, Brian J and Niklaus J Grünwald (2017). "VCFR: a package to manipulate and visualize variant call format data in R". In: *Molecular Ecology Resources* 17.1, pp. 44–53.

Koboldt, Daniel C et al. (2012). "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing". In: *Genome research*.

Korte, Arthur and Ashley Farlow (2013). "The advantages and limitations of trait analysis with GWAS: a review". In: *Plant methods* 9.1, p. 29. ISSN: 1746-4811.

Kouris-Blazos, Antigone and Regina Belski (2016). "Health benefits of legumes and pulses with a focus on Australian sweet lupins". In: *Asia Pacific journal of clinical nutrition* 25.1, pp. 1–17.

Krajewski, P et al. (2012). "QTL for yield components and protein content: a multienvironment study of two pea (Pisum sativum L.) populations". In: *Euphytica* 183.3, pp. 323–336.

Krishnakumar, Vivek et al. (2014). "MTGD: The Medicago truncatula genome database". In: *Plant and Cell Physiology* 56.1, e1–e1.

Kumagai, Hirotaka et al. (2007). "A novel ankyrin-repeat membrane protein, IGN1, is required for persistence of nitrogen-fixing symbiosis in root nodules of Lotus japonicus". In: *Plant Physiology* 143.3, pp. 1293–1305.

Kumar, Neeraj et al. (2012). "Leafsnap: A computer vision system for automatic plant species identification". In: *Computer vision–ECCV 2012*. Springer, pp. 502–516.

Kwon, Soon-Jae et al. (2012). "Genetic diversity, population structure and genome-wide marker-trait association analysis emphasizing seed nutrients of the USDA pea (Pisum sativum L.) core collection". In: *Genes & Genomics* 34.3, pp. 305–320.

Lafond, G, LE Evans, and ST Ali-Khan (1981). "Comparison of near-isogenic leafed, leafless, semi-leafless, and reduced stipule lines of peas for yield and associated traits". In: *Canadian journal of plant science* 61.2, pp. 463–465.

Langmead, Ben and Steven L Salzberg (2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4, pp. 357–359. ISSN: 1548-7091.

Larrainzar, Estíbaliz and Stefanie Wienkoop (2017). "A Proteomic View on the Role of Legume Symbiotic Interactions". In: *Frontiers in plant science* 8, p. 1267.

Laurie, Cathy C et al. (2010). "Quality control and quality assurance in genotypic data for genome-wide association studies". In: *Genetic epidemiology* 34.6, pp. 591–602.

Lee, Chia-Ling and Shu-Yuan Chen (2006). "Classification of leaf images". In: *International Journal of Imaging Systems and Technology* 16.1, pp. 15–23.

Lee, THOMAS D et al. (1988). "Patterns of fruit and seed production". In: *Plant reproductive ecology: patterns and strategies. Oxford University Press, New York*, pp. 179–202.

Leigh, John A, Ethan R Signer, and Graham C Walker (1985). "Exopolysaccharide-deficient mutants of Rhizobium meliloti that form ineffective nodules". In: *Proceedings of the National Academy of Sciences* 82.18, pp. 6231–6235.

LemnaTec (2018). *LemnaTec-OS Phenotyping Software*. URL: https://www.lemnatec.com/products/software/.

Lester, Diane R et al. (1997). "Mendel's stem length gene (Le) encodes a gibberellin 3 beta-hydroxylase." In: *The Plant Cell* 9.8, pp. 1435–1443.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14, pp. 1754–1760. ISSN: 1367-4803.

Li, Heng et al. (2009). "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–2079.

Li, Yue and Manolis Kellis (2016). "Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases". In: *Nucleic acids research* 44.18, e144–e144.

Li, Yun et al. (2010). "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes". In: *Genetic epidemiology* 34.8, pp. 816–834. ISSN: 1098-2272. URL:

https://deepblue.lib.umich.edu/bitstream/handle/2027.42/78318/gepi_20533_sm_Supplfig2.pdf?sequence=1.

Li, Zhenyu et al. (2012). "Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph". In: *Briefings in functional genomics* 11.1, pp. 25–37.

Liebman, Matt and Elizabeth Dyck (1993). "Crop rotation and intercropping strategies for weed management". In: *Ecological applications* 3.1, pp. 92–122.

Limited, Petiole (2018). *Petiole*. URL: http://petioleapp.com/.

Lipka, Alexander E et al. (2012). "GAPIT: genome association and prediction integrated tool". In: *Bioinformatics* 28.18, pp. 2397–2399. ISSN: 1367-4803.

Lischer, Heidi EL and Laurent Excoffier (2011). "PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs". In: *Bioinformatics* 28.2, pp. 298–299.

Liu, Lin et al. (2012). "Comparison of next-generation sequencing systems". In: *BioMed Research International* 2012. ISSN: 1110-7243.

Liu, Yan-Li et al. (2008). "A robust and fast non-local means algorithm for image denoising". In: *Journal of computer science and technology* 23.2, pp. 270–279.

Lobet, Guillaume (2017). "Image analysis in plant sciences: publish then perish". In: *Trends in Plant Science* 22.7, pp. 559–566.

Lobet, Guillaume, Xavier Draye, and Claire Périlleux (2013). "An online database for plant image analysis software tools". In: *Plant methods* 9.1, p. 38.

Lu, Fei et al. *TASSEL 3.0 Universal Network Enabled Analysis Kit (UNEAK) pipeline documentation*. https://bytebucket.org/tasseladmin/tassel-5-source/wiki/docs/TasselPipelineUNEAK.pdf?rev=4e497b7d0e44fba851fce921e30540eccb95bda9. Accessed: 30th September 2018.

Lu, Fei et al. (2013). "Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol". In: *PLoS genetics* 9.1, e1003215. ISSN: 1553-7404.

Luo, Ruibang et al. (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler". In: *Gigascience* 1.1, p. 18.

Mabee, Paula M et al. (2007). "Phenotype ontologies: the bridge between genomics and evolution". In: *Trends in ecology & evolution* 22.7, pp. 345–350. ISSN: 0169-5347.

Macdiarmid, Jennie I et al. (2018). "Assessing national nutrition security: The UK reliance on imports to meet population energy and nutrient recommendations". In: *PloS one* 13.2, e0192649.

Machado, Bruno Brandoli et al. (2016). "BioLeaf: A professional mobile application to measure foliar damage caused by insect herbivory". In: *Computers and Electronics in Agriculture* 129, pp. 44–55.

Mamanova, Lira et al. (2010). "Target-enrichment strategies for next-generation sequencing". In: *Nature methods* 7.2, p. 111.

Mapleson, Daniel et al. (2016). "KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies". In: *Bioinformatics* 33.4, pp. 574–576.

Marchini, Jonathan and Bryan Howie (2010). "Genotype imputation for genome-wide association studies". In: *Nature Reviews Genetics* 11.7, pp. 499–511. ISSN: 1471-0056.

Mastretta-Yanes, A et al. (2015). "Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference". In: *Molecular Ecology Resources* 15.1, pp. 28–41.

Matplotlib - pyplot. https://matplotlib.org/api/pyplot_api.html. Accessed: 1st December 2018.

Max Planck Institute (2018). *Network Extraction From Images 2.0*. URL: http://nefi.mpi-inf.mpg.de/index.php.

Maxted, Nigel and Mike Ambrose (2001). "Peas (Pisum L.)" In: *Plant genetic resources of legumes in the Mediterranean*. Springer, pp. 181–190.

McCouch, Susan R et al. (2012). "Genomics of gene banks: A case study in rice". In: *American journal of botany* 99.2, pp. 407–423. ISSN: 0002-9122.

McDonald, GK and GM Paulsen (1997). "High temperature effects on photosynthesis and water relations of grain legumes". In: *Plant and Soil* 196.1, pp. 47–58.

McGee, Rebecca J and James R Baggett (1992). “Inheritance of Stringless Pod in Pisum sativum L.” In: *Journal of the American Society for Horticultural Science* 117.4, pp. 628–632.

McKenna, Aaron et al. (2010). “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data”. In: *Genome research*.

McLennan, Deborah A (2010). “How to read a phylogenetic tree”. In: *Evolution: Education and Outreach* 3.4, p. 506.

McPhee, Kevin (2005). “Variation for seedling root architecture in the core collection of pea germplasm”. In: *Crop Science* 45.5, pp. 1758–1763.

Mendel, Gregor (1996). “Experiments in plant hybridization (1865)”. In: *Verhandlungen des naturforschenden Vereins Brünn.) Available online: www.mendelweb.org/Mendel.html (accessed on 1 January 2013)*.

Menge, Duncan NL, Amelia A Wolf, and Jennifer L Funk (2015). “Diversity of nitrogen fixation strategies in Mediterranean legumes”. In: *Nature plants* 1, p. 15064.

Miescher-Rüsch, Friedrich (1871). *Ueber die chemische Zusammensetzung der Eiterzellen*.

Migicovsky, Zoë et al. (2018). “Morphometrics reveals complex and heritable apple leaf shapes”. In: *Frontiers in Plant Science* 8, p. 2185.

Miko, I (2008). “Gregor Mendel and the principles of inheritance”. In: *Nature Education* 1.1, p. 134.

Miller, Lewis E (1954). *The chemistry and vertical distribution of the oxides of nitrogen in the atmosphere*. Tech. rep. AIR FORCE CAMBRIDGE RESEARCH LABS HANSCOM AFB MA.

Miller, Nathan et al. (2018). *Seed size analyser Bisque*. URL: https://bisque.cyverse.org/module_service/SeedSize/.

Miller, Nathan D et al. (2017). “A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images”. In: *The Plant Journal* 89.1, pp. 169–178.

Moore, Candace R et al. (2013). “Mapping quantitative trait loci affecting Arabidopsis thaliana seed morphology features extracted computationally from images”. In: *G3: Genes, Genomes, Genetics* 3.1, pp. 109–118.

Mungall, Christopher J et al. (2010). "Integrating phenotype ontologies across multiple species". In: *Genome biology* 11.1, R2. ISSN: 1465-6906.

Muñoz-Amatriaín, María et al. (2014). "The USDA barley core collection: genetic diversity, population structure, and potential for genome-wide association studies". In: *PloS one* 9.4, e94688.

Murray, Michael G, Debra L Peters, and William F Thompson (1981). "Ancient repeated sequences in the pea and mung bean genomes and implications for genome evolution". In: *Journal of Molecular Evolution* 17.1, pp. 31–42.

Mustroph, Angelika, Uwe Sonnewald, and Sophia Biemelt (2007). "Characterisation of the ATP-dependent phosphofructokinase gene family from Arabidopsis thaliana". In: *FEBS letters* 581.13, pp. 2401–2410.

Nadalin, Francesca, Francesco Vezzi, and Alberto Policriti (2012). "GapFiller: a de novo assembly approach to fill the gap within paired reads". In: *BMC bioinformatics* 13.14, S8.

Nawab, Nausherwan Nobel et al. (2008). "Genetic variability, correlation and path analysis studies in garden pea (Pisum sativum L.)" In: *J. Agric. Res* 46.4, pp. 333–340.

Neumann, Pavel et al. (2002). "Chromosome sorting and PCR-based physical mapping in pea (Pisum sativum L.)" In: *Chromosome Research* 10.1, pp. 63–71.

Newton, Glen and Dwight Deugo (2007). "Takaka: Eclipse Image Processing Plug-in." In: *IPCV*, pp. 213–219.

Next Instruments (2016). *SeedCount*. URL: `http://www.nextinstruments.net/products/seedcount`.

Novotný, Petr and Tomáš Suk (2013). "Leaf recognition of woody species in Central Europe". In: *Biosystems Engineering* 115.4, pp. 444–452.

Odong, TL et al. (2013). "Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation". In: *Theoretical and Applied Genetics* 126.2, pp. 289–305. ISSN: 0040-5752.

Oghbaei, Morteza and Jamuna Prakash (2016). "Effect of primary processing of cereals and legumes on its nutritional quality: A comprehensive review". In: *Cogent Food & Agriculture* 2.1, p. 1136015.

Okçu, Gamze, Mehmet Demir Kaya, and Mehmet Atak (2005). "Effects of salt and drought stresses on germination and seedling growth of pea (Pisum sativum L.)" In: *Turkish journal of agriculture and forestry* 29.4, pp. 237–242.

Olby, Robert (1987). "William Bateson's introduction of Mendelism to England: a reassessment". In: *The British Journal for the History of Science* 20.4, pp. 399–420.

Oldoni, Tatiane Luiza C et al. (2011). "Isolation and analysis of bioactive isoflavonoids and chalcone from a new type of Brazilian propolis". In: *Separation and purification Technology* 77.2, pp. 208–213.

Oliveira, Marcelo F et al. (2010). "Establishing a soybean germplasm core collection". In: *Field crops research* 119.2-3, pp. 277–289.

Ono, Hanako et al. (2015). "Removal of redundant contigs from de novo RNA-Seq assemblies via homology search improves accurate detection of differentially expressed genes". In: *BMC genomics* 16.1, p. 1031.

Otsu, Nobuyuki (1979). "A threshold selection method from gray-level histograms". In: *IEEE transactions on systems, man, and cybernetics* 9.1, pp. 62–66.

P. sativum CSFL RefTrans V1. https://www.coolseasonfoodlegume.org/analysis/143. Accessed: 30th May 2018.

P. sativum CSFL RefTrans V2. https://www.coolseasonfoodlegume.org/analysis/158. Accessed: 30th May 2018.

Pan, Qingchun et al. (2017). "The genetic basis of plant architecture in 10 maize recombinant inbred line populations". In: *Plant physiology* 175.2, pp. 858–873.

Paris, Josephine R, Jamie R Stevens, and Julian M Catchen (2017). "Lost in parameter space: A road map for Stacks". In: *Methods in Ecology and Evolution*. ISSN: 2041-210X.

Peksen, Erkut et al. (2004). "Some seed traits and their relationships to seed germination and field emergence in pea (Pisum sativum L.)" In: *Journal of Agronomy* 3.4, pp. 243–246.

Peppe, Daniel J et al. (2011). "Sensitivity of leaf size and shape to climate: global patterns and paleoclimatic applications". In: *New Phytologist* 190.3, pp. 724–739.

Perabo, FGE et al. (2008). "Soy isoflavone genistein in prevention and treatment of prostate cancer". In: *Prostate cancer and prostatic diseases* 11.1, p. 6.

Peterson, Brant K et al. (2012). “Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species”. In: *PloS one* 7.5, e37135. ISSN: 1932-6203.

Pethybridge, Sarah J and Scot C Nelson (2015). “Leaf Doctor: A new portable application for quantifying plant disease severity”. In: *Plant Disease* 99.10, pp. 1310–1316.

PhenoSpex. https://phenospex.com/. Accessed: 12th September 2018.

Phillippy, Adam M, Michael C Schatz, and Mihai Pop (2008). “Genome assembly forensics: finding the elusive mis-assembly”. In: *Genome biology* 9.3, R55.

Pieruschka, Roland and Hendrik Poorter (2012). “Phenotyping plants: genes, phenes and machines”. In: *Functional Plant Biology* 39.11, pp. 813–820. ISSN: 1445-4416.

Pisum sativum unigene v1. https://www.coolseasonfoodlegume.org/sativum_unigene_v1. Accessed: 30th May 2018.

Pisum sativum unigene v2. https://www.coolseasonfoodlegume.org/sativum_unigene_v2. Accessed: 30th May 2018.

Pisum sativum unigene wa1. https://www.coolseasonfoodlegume.org/sativum_unigene_wa1. Accessed: 30th May 2018.

Poland, Jesse A et al. (2012). “Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach”. In: *PloS one* 7.2, e32253. ISSN: 1932-6203.

Polder, G, G Blokker, and GWAM van der Heijden (2012). “An ImageJ plugin for plant variety testing”. In: *Proceedings of the ImageJ User and Developer Conference 2012, 24-26 October 2012, Mondorf-les-Bains, Luxembourg*, pp. 168–173.

Pop, Mihai (2009). “Genome assembly reborn: recent computational challenges”. In: *Briefings in bioinformatics* 10.4, pp. 354–366.

Prasad, R, KK Mukherjee, and G Gangopadhyay (2014). “Image-Analysis Based on Seed Phenomics in Sesame”. In: *Plant Breeding and Seed Science* 68.1, pp. 119–136.

Prep, SPEX Sample. *Geno/Grinder® - Automated Tissue Homogenizer and Cell Lyser*. https://www.spexsampleprep.com/2010genogrinder. Accessed: 30th September 2018.

Price, Alkes L et al. (2006). “Principal components analysis corrects for stratification in genome-wide association studies”. In: *Nature genetics* 38.8, pp. 904–909.

Price, Charles A et al. (2010). “LEAF GUI: segmenting and analyzing the structure of leaf veins and areoles”. In: *Plant Physiology*, pp–110.

Prioul, S et al. (2004). “Mapping of quantitative trait loci for partial resistance to Mycosphaerella pinodes in pea (Pisum sativum L.), at the seedling and adult plant stages”. In: *Theoretical and Applied Genetics* 108.7, pp. 1322–1334.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly (2000). “Inference of population structure using multilocus genotype data”. In: *Genetics* 155.2, pp. 945–959. ISSN: 0016-6731.

Puchta, Holger (2002). “Gene replacement by homologous recombination in plants”. In: *Functional Genomics*. Springer, pp. 173–182.

Purugganan, Michael D and Dorian Q Fuller (2009). “The nature of selection during plant domestication”. In: *Nature* 457.7231, p. 843.

Qaisrani, SN et al. (2014). “Protein source and dietary structure influence growth performance, gut morphology, and hindgut fermentation characteristics in broilers”. In: *Poultry science* 93.12, pp. 3053–3064.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rambaut, A (2007). *FigTree, a graphical viewer of phylogenetic trees*. http://tree.bio.ed.ac.uk/software/figtree. Accessed: 28th August 2018.

Rasmussen, KJ (1999). “Impact of ploughless soil tillage on yield and soil quality: a Scandinavian review”. In: *Soil and Tillage Research* 53.1, pp. 3–14.

Rayner, Tracey et al. (2017). “Genetic Variation Controlling Wrinkled Seed Phenotypes in Pisum: How Lucky Was Mendel?” In: *International journal of molecular sciences* 18.6, p. 1205.

Remmler, Lauren and Anne-Gaëlle Rolland-Lagan (2012). “Computational method for quantifying growth patterns at the adaxial leaf surface in three dimensions”. In: *Plant Physiology*, pp–112.

Rocca-Serra, Philippe et al. (2010). “ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level”. In: *Bioinformatics* 26.18, pp. 2354–2356. ISSN: 1367-4803.

Ross-Ibarra, Jeffrey, Peter L Morrell, and Brandon S Gaut (2007). “Plant domestication, a unique opportunity to identify the genetic basis of adaptation”. In: *Proceedings of the National Academy of Sciences* 104.suppl 1, pp. 8641–8648.

Rousseau, Céline et al. (2013). “High throughput quantitative phenotyping of plant resistance using chlorophyll fluorescence image analysis”. In: *Plant Methods* 9.1, p. 17.

Royer, Dana L et al. (2005). “Correlations of climate and plant ecology to leaf size and shape: potential proxies for the fossil record”. In: *American Journal of Botany* 92.7, pp. 1141–1151.

Ruysschaert, Greet et al. (2004). “Soil loss due to crop harvesting: significance and determining factors”. In: *Progress in Physical Geography* 28.4, pp. 467–501.

Ryan, Thomas P and JP Morgan (2007). “Modern experimental design”. In: *Journal of Statistical Theory and Practice* 1.3-4, pp. 501–506.

Salinas, Julio et al. (1988). “Compositional compartmentalization and compositional patterns in the nuclear genomes of plants”. In: *Nucleic Acids Research* 16.10, pp. 4269–4285.

Sandeep, VV, DK Kanaka, and K Keshavulu (2013). “Seed image analysis: its applications in seed science research”. In: *International Research Journal of Agricultural Science* 1.2, pp. 30–36.

Schaid, Daniel J, Wenan Chen, and Nicholas B Larson (2018). “From genome-wide associations to candidate causal variants by statistical fine-mapping”. In: *Nature Reviews Genetics* 19.8, p. 491.

Schindelin, Johannes et al. (2012). “Fiji: an open-source platform for biological-image analysis”. In: *Nature methods* 9.7, p. 676.

Schmidt, Dominic et al. (2009). “ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions”. In: *Methods* 48.3, pp. 240–248.

Schneider, Caroline A, Wayne S Rasband, and Kevin W Eliceiri (2012). “NIH Image to ImageJ: 25 years of image analysis”. In: *Nature methods* 9.7, p. 671.

Sciences, Beckman Coulter Life. *Agencourt GenFind V2*. https://www.mybeckman.uk/reagents/genomic/dna-isolation/from-blood. Accessed: 30th September 2018.

SciPy. https://docs.scipy.org/doc/scipy/reference/. Accessed: 1st December 2018.

SeedStor JI1153. https://www.seedstor.ac.uk/search-infoaccession.php?idPlant=24549. Accessed: 30th September 2018.

Shafiekhani, A et al. (2017). "Automated classification of wrinkle levels in seed coat using relevance vector machine". In:

Shapely - Manipulation and analysis of geometric objects. https://github.com/Toblerity/Shapely. Accessed: 1st December 2018.

Shendure, Jay et al. (2017). "DNA sequencing at 40: past, present and future". In: *Nature* 550.7676, p. 345.

Sherriff, Leanne J et al. (1994). "Decapitation reduces the metabolism of gibberellin A20 to A1 in Pisum sativum L., decreasing the Le/le difference". In: *Plant physiology* 104.1, pp. 277–280.

Shewry, Peter R (2007). "Improving the protein content and composition of cereal grain". In: *Journal of cereal science* 46.3, pp. 239–250.

Sievenpiper, JL et al. (2009). *Effect of non-oil-seed pulses on glycaemic control: a systematic review and meta-analysis of randomised controlled experimental trials in people with and without diabetes*.

Simão, Felipe A et al. (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". In: *Bioinformatics* 31.19, pp. 3210–3212.

Simon, CJ and FJ Muehlbauer (1997). "Construction of a chickpea linkage map and its comparison with maps of pea and lentil". In: *Journal of Heredity* 88.2, pp. 115–119.

Simpson, Jared T et al. (2009). "ABySS: a parallel assembler for short read sequence data". In: *Genome research*, gr–089532.

Sindhu, Anoop et al. (2014). "Gene-based SNP discovery and genetic mapping in pea". In: *Theoretical and applied genetics* 127.10, pp. 2225–2241.

Singh, Akansha, Shalini Singh, and J Dayal Prasad Babu (2011). "Heritability, character association and path analysis studies in early segregating population of field pea (Pisum sativum L. var. arvense)". In: *International Journal of Plant Breeding and Genetics* 5.1, pp. 86–92.

Skinner, Mitchell E et al. (2009). "JBrowse: a next-generation genome browser". In: *Genome research* 19.9, pp. 1630–1638. ISSN: 1088-9051.

Skye Instruments Limited (2018). *Skye*. URL: `http://www.skyeinstruments.com/products/plant-analysis-systems/leaf-arearoot-length-systems/`.

Slatkin, Montgomery (2008). "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future". In: *Nature Reviews Genetics* 9.6, p. 477.

Smart Imaging Technologies (2013). *Simagis*. URL: `http://smartimtech.com/analysis.htm`.

Smartt, J (1984). "Evolution of grain legumes. I. Mediterranean pulses". In: *Experimental Agriculture* 20.4, pp. 275–296.

Smit, Patrick et al. (2005). "NSP1 of the GRAS protein family is essential for rhizobial Nod factor-induced transcription". In: *Science* 308.5729, pp. 1789–1791.

Smith, Barry et al. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration". In: *Nature biotechnology* 25.11, pp. 1251–1255. ISSN: 1087-0156.

Smith, Ray (2007). "An overview of the Tesseract OCR engine". In: *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*. Vol. 2. IEEE, pp. 629–633.

Smýkal, P et al. (2008). "Effort towards a world pea (Pisum sativum L.) germplasm core collection: the case for common markers and data compatibility". In: *Pisum Genet* 40, pp. 11–14.

Smýkal, Petr et al. (2011). "Phylogeny, phylogeography and genetic diversity of the Pisum genus". In: *Plant Genetic Resources* 9.1, pp. 4–18.

Smýkal, Petr et al. (2015). "Legume crops phylogeny and genetic diversity for science and breeding". In: *Critical Reviews in Plant Sciences* 34.1-3, pp. 43–104.

Smykalova, I et al. (2011). "Morpho-colorimetric traits of Pisum seeds measured by an image analysis system". In: *Seed Science and Technology* 39.3, pp. 612–626.

Smýkal, Petr et al. (2012). "Pea (Pisum sativum L.) in the genomic era". In: *Agronomy* 2.2, pp. 74–115.

Snoad, B (1974). "A preliminary assessment of ?leafless peas?" In: *Euphytica* 23.2, pp. 257–265.

Spindel, Jennifer et al. (2015). "Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines". In: *PLoS genetics* 11.2, e1004982.

Steele, KELLY P and MARTIN F Wojciechowski (2003). "Phylogenetic analyses of tribes Trifolieae and Vicieae, based on sequences of the plastid gene matK (Papilionoideae: Leguminosae)". In: *Advances in legume systematics, part* 10, pp. 355–370.

Sturtevant, Alfred H (1913). "The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association". In: *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology* 14.1, pp. 43–59.

Sun, Han, Jingyu Yang, and Mingwu Ren (2005). "A fast watershed algorithm based on chain code and its application in image segmentation". In: *Pattern Recognition Letters* 26.9, pp. 1266–1274.

Suwarno, Willy B et al. (2015). "Genome-wide association analysis reveals new targets for carotenoid biofortification in maize". In: *Theoretical and Applied Genetics* 128.5, pp. 851–864.

Suzuki, Marina et al. (2018). "OLIGOCELLULA1/HIGH EXPRESSION OF OSMOTICALLY RESPONSIVE GENES15 promotes cell proliferation with HISTONE DEACETYLASE9 and POWERDRESS during leaf development in Arabidopsis thaliana". In: *Frontiers in plant science* 9.

Tanabata, Takanari et al. (2012). "SmartGrain: high-throughput phenotyping software for measuring seed shape through image analysis". In: *Plant physiology*, pp–112.

Tang, Haibao et al. (2014). "An improved genome release (version Mt4. 0) for the model legume Medicago truncatula". In: *BMC genomics* 15.1, p. 312.

Tar'an, B et al. (2003). "Quantitative trait loci for lodging resistance, plant height and partial resistance to mycosphaerella blight in field pea (Pisum sativum L.)" In: *Theoretical and Applied Genetics* 107.8, pp. 1482–1491.

Tessari, Paolo, Anna Lante, and Giuliano Mosca (2016). "Essential amino acids: master regulators of nutrition and environmental footprint?" In: *Scientific reports* 6, p. 26074.

Thachuk, Chris et al. (2009). "Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures". In: *BMC bioinformatics* 10.1, p. 243.

The Arabidopsis Information Resource (TAIR). https://www.arabidopsis.org/Blast/index.jsp. Accessed: 22nd November 2018.

Thorisson, Gudmundur A, Juha Muilu, and Anthony J Brookes (2009). "Genotype–phenotype databases: challenges and solutions for the post-genomic era". In: *Nature Reviews Genetics* 10.1, pp. 9–18. ISSN: 1471-0056.

Tian, Feng et al. (2011). "Genome-wide association study of leaf architecture in the maize nested association mapping population". In: *Nature genetics* 43.2, p. 159.

Tilman, David et al. (2002). "Agricultural sustainability and intensive production practices". In: *Nature* 418.6898, p. 671.

Torkamaneh, Davoud, Jérôme Laroche, and François Belzile (2016). "Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies". In: *PloS one* 11.8, e0161333. ISSN: 1932-6203.

Tran, Lam-Son Phan et al. (2004). "Isolation and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter". In: *The Plant Cell* 16.9, pp. 2481–2498.

Treuren, Rob van and Theo JL van Hintum (2014). "Next-generation genebanking: plant genetic resources management and utilization in the sequencing era". In: *Plant Genetic Resources*, pp. 1–10. ISSN: 1479-263X.

Trichopoulou, Antonia et al. (1995). "Diet and overall survival in elderly people". In: *Bmj* 311.7018, pp. 1457–1460.

Turner, Stephen D (2014). "qqman: an R package for visualizing GWAS results using QQ and manhattan plots". In: *Biorxiv*, p. 005165.

United Nations, Department of Economic and Population Division Social Affairs (2017). “World Population Prospects: The 2017 Revision, Key Findings and Advance Tables.” In: *ESA/P/WP/248*. DOI: https://esa.un.org/unpd/wpp/Publications/Files/WPP2017_KeyFindings.pdf.

US National Plant Germplasm System PI 210639. https://npgsweb.ars-grin.gov/gringlobal/accessiondetail.aspx?id=1174807. Accessed: 30th September 2018.

USDA National Nutrient Database for Standard Reference. https://ndb.nal.usda.gov/ndb/. Accessed: 17-07-2017.

Van Orsouw, Nathalie J et al. (2007). “Complexity reduction of polymorphic sequences (CRoPS$^{\mathrm{TM}}$): a novel approach for large-scale polymorphism discovery in complex genomes”. In: *PloS one* 2.11, e1172. ISSN: 1932-6203.

Vanclay, Jerome K (2006). “Experiment designs to evaluate inter-and intra-specific interactions in mixed plantings of forest trees”. In: *Forest Ecology and Management* 233.2-3, pp. 366–374.

Varma, Varun and Anand M Osuri (2013). “Black Spot: a platform for automated and rapid estimation of leaf area from scanned images”. In: *Plant ecology* 214.12, pp. 1529–1534.

Varshney, Rajeev K, Andreas Graner, and Mark E Sorrells (2005). “Genomics-assisted breeding for crop improvement”. In: *Trends in plant science* 10.12, pp. 621–630. ISSN: 1360-1385. URL: http://ac.els-cdn.com/S1360138505002554/1-s2.0-S1360138505002554-main.pdf?_tid=20e81d6a-b830-11e4-8f13-00000aab0f6b&acdnat=1424347937_4e4a3e109b81c2d480c7c8adb11778cb.

Varshney, Rajeev K et al. (2015). “Translational genomics in agriculture: some examples in grain legumes”. In: *Critical Reviews in Plant Sciences* 34.1-3, pp. 169–194. ISSN: 0735-2689. URL: http://www.tandfonline.com/doi/pdf/10.1080/07352689.2014.897909.

Venkateshwaran, Muthusubramanian and Jean-Michel Ané (2011). “Legumes and nitrogen fixation: physiological, molecular, evolutionary perspective and applications”. In: *The Molecular Basis of Nutrient Use Efficiency in Crops*, pp. 457–489.

Venora, G, Oscar Grillo, and R Saccone (2009). “Quality assessment of durum wheat storage centres in Sicily: evaluation of vitreous, starchy and shrunken kernels using an image analysis system”. In: *Journal of cereal science* 49.3, pp. 429–440.

Vincent, Luc and Pierre Soille (1991). “Watersheds in digital spaces: an efficient algorithm based on immersion simulations”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6, pp. 583–598.

Visscher, Peter M et al. (2012). “Five years of GWAS discovery”. In: *The American Journal of Human Genetics* 90.1, pp. 7–24.

Voight, Benjamin F and Jonathan K Pritchard (2005). “Confounding from cryptic relatedness in case-control association studies”. In: *PLoS genetics* 1.3, e32.

Vooren, JG Van de and GWAM Van Der Heijden (1993). “Measuring the size of french beans with image analysis”. In: *Plant Varieties and Seeds* 6, pp. 47–47.

Wagner, SC (2011). *Biological Nitrogen Fixation. Nature Education Knowledge, 3, 15.*

Wäldchen, Jana and Patrick Mäder (2017). “Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review”. In: *Archives of Computational Methods in Engineering*, pp. 1–37.

Walt, Stéfan van der et al. (June 2014). “scikit-image: image processing in Python”. In: *PeerJ* 2, e453. ISSN: 2167-8359. DOI: `10.7717/peerj.453`. URL: `http://dx.doi.org/10.7717/peerj.453`.

Walt, Stéfan van der, S Chris Colbert, and Gael Varoquaux (2011). “The NumPy array: a structure for efficient numerical computation”. In: *Computing in Science & Engineering* 13.2, pp. 22–30.

Walter, Achim and Ulrich Schurr (2005). “Dynamics of leaf and root growth: endogenous control versus environmental impact”. In: *Annals of botany* 95.6, pp. 891–900.

Wang, Anqi et al. (2018). “BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach”. In: *Bioinformatics* 34.12, pp. 2019–2028.

Wang, Shi et al. (2012). “2b-RAD: a simple and flexible method for genome-wide genotyping”. In: *Nature methods* 9.8, pp. 808–810. ISSN: 1548-7091.

Wang, TL and CL Hedley (1991). “Seed development in peas: knowing your three ?r’s?(or four, or five)”. In: *Seed Science Research* 1.1, pp. 3–14.

Warr, Amanda et al. (2015). “Exome sequencing: current and future perspectives”. In: *G3: Genes, Genomes, Genetics*, g3–115.

Warren, René L et al. (2006). “Assembling millions of short DNA sequences using SSAKE”. In: *Bioinformatics* 23.4, pp. 500–501.

Weeden, NF, MJ Ambrose, et al. (2004). “Ser appears to be the serrate leaflet locus mapped on linkage group III.” In: *Pisum Genetics* 36.

Weeden, Norman F (2007). “Genetic changes accompanying the domestication of Pisum sativum: is there a common genetic basis to the ‘domestication syndrome’for legumes?” In: *Annals of Botany* 100.5, pp. 1017–1025. ISSN: 0305-7364.

Weeden, NORMAN F, SOREN Brauner, and Jerzy A Przyborowski (2002). “Genetic analysis of pod dehiscence in pea (Pisum sativum L.)” In: *Cellular and Molecular Biology Letters* 7.2B, pp. 657–664.

Wehner, TC and ET Gritton (1981). “Effect of the n gene on pea pod characteristics”. In: *J. Amer. Soc. Hort. Sci* 106.2, pp. 181–183.

Weight, Caroline, Daniel Parnham, and Richard Waites (2008). “TECHNICAL ADVANCE: LeafAnalyser: a computational method for rapid and large-scale analyses of leaf shape variation”. In: *The Plant Journal* 53.3, pp. 578–586.

Wellcome Collection - Johann Friedrich Miescher. Photograph. https://wellcomecollection.org/works/vsxpqn8m. Accessed: 16th June 2018.

Wellcome Collection - Portrait of Charles Darwin. https://wellcomecollection.org/works/t7yqd4uq. Accessed: 16th June 2018.

Wellcome Collection - Portrait of Gregor Johann Mendel, Garrison. https://wellcomecollection.org/works/tc5xq5ad. Accessed: 16th June 2018.

Wen, Weiwei et al. (2011). “Detection of genetic integrity of conserved maize (Zea mays L.) germplasm in genebanks using SNP markers”. In: *Genetic Resources and Crop Evolution* 58.2, pp. 189–207. ISSN: 0925-9864.

Westphal, Egbert (1974). *Pulses in Ethiopia, their taxonomy and agricultural significance*. Pudoc.

Whan, Alex P et al. (2014). “GrainScan: a low cost, fast method for grain size and colour measurements”. In: *Plant methods* 10.1, p. 23.

Widawsky, David et al. (1998). “Pesticide productivity, host-plant resistance and productivity in China”. In: *Agricultural economics* 19.1-2, pp. 203–217.

WinFolia, TM (2001). “Regent Instruments Inc”. In: *Quebec, Canada, version PRO*.

WinSeedle, TM (2005). “Regent Instruments Inc”. In: *Quebec, Canada, version PRO*.

Wouw, Mark van de et al. (2010). “Genetic erosion in crops: concept, research results and challenges”. In: *Plant Genetic Resources* 8.01, pp. 1–15. ISSN: 1479-263X.

Wrzaczek, Michael et al. (2010). “Transcriptional regulation of the CRK/DUF26 group of receptor-like protein kinases by ozone and plant hormones in Arabidopsis”. In: *BMC Plant Biology* 10.1, p. 95.

Wu, Congling et al. (2015). “Fine phenotyping of pod and seed traits in Arachis germplasm accessions using digital image analysis”. In: *Peanut Science* 42.2, pp. 65–73.

Wu, Yang et al. (2017). “Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data”. In: *Genome biology* 18.1, p. 86.

Xiao, Yingjie et al. (2016). “Genome-wide dissection of the maize ear genetic architecture using multiple populations”. In: *New Phytologist* 210.3, pp. 1095–1106.

Xu, Fei et al. (2009). “Leaf morphology correlates with water and light availability: what consequences for simple and compound leaves”. In: *Progress in Natural Science* 19.12, pp. 1789–1798.

Yan, Jianbing et al. (2010). “Rare genetic variation at Zea mays crtRB1 increases β-carotene in maize grain”. In: *Nature genetics* 42.4, p. 322.

Yan, WenGui et al. (2007). “Development and evaluation of a core subset of the USDA rice germplasm collection”. In: *Crop Science* 47.2, pp. 869–876.

Yandell, Mark and Daniel Ence (2012). “A beginner’s guide to eukaryotic genome annotation”. In: *Nature Reviews Genetics* 13.5, p. 329.

Yang, Jian et al. (2011). “Genomic inflation factors under polygenic inheritance”. In: *European Journal of Human Genetics* 19.7, p. 807.

Yang, Wanneng et al. (2014). “Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice”. In: *Nature communications* 5.

Zair, Wathek, Nigel Maxted, and Ahmed Amri (2018). “Setting conservation priorities for crop wild relatives in the Fertile Crescent”. In: *Genetic Resources and Crop Evolution* 65.3, pp. 855–863.

Zargar, Sajad Majeed et al. (2015). “Recent advances in molecular marker techniques: insight into QTL mapping, GWAS and genomic selection in plants”. In: *Journal of crop science and biotechnology* 18.5, pp. 293–308.

Zeng, Ping et al. (2015). “Statistical analysis for genome-wide association study”. In: *Journal of biomedical research* 29.4, p. 285.

Zerbino, Daniel and Ewan Birney (2008). “Velvet: algorithms for de novo short read assembly using de Bruijn graphs”. In: *Genome research*, gr–074492.

Zeuli, PL Spagnoletti and CO Qualset (1993). “Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat”. In: *Theoretical and Applied Genetics* 87.3, pp. 295–304.

Zhang, Wenyu et al. (2011). “A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies”. In: *PloS one* 6.3, e17915.

Zhang, Yong et al. (2008). “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome biology* 9.9, R137.

Zhang, Zhiwu et al. (2010). “Mixed linear model approach adapted for genome-wide association studies”. In: *Nature genetics* 42.4, p. 355.

Zhao, Keyan et al. (2011). “Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa”. In: *Nature communications* 2, p. 467.

Zhu, Beibei et al. (2015). “Dietary legume consumption reduces risk of colorectal cancer: evidence from a meta-analysis of cohort studies”. In: *Scientific reports* 5, p. 8797.

Zhu, Hongyan et al. (2005). “Bridging model and crop legumes through comparative genomics”. In: *Plant physiology* 137.4, pp. 1189–1196. ISSN: 1532-2548.

Zohary, Daniel and Maria Hopf (1973). "Domestication of Pulses in the Old World Legumes were companions of wheat and barley when agriculture began in the Near East". In: *Science* 182.4115, pp. 887–894. ISSN: 0036-8075.

Zohary, Daniel, Maria Hopf, and Ehud Weiss (2012). *Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*. Oxford University Press on Demand.