Methods for Marker Assisted Breeding in Octoploid Strawberry (*Fragaria* × *ananassa*)

Joe He

A thesis for submitted the degree of
Doctor of Philosophy

Department of Agriculture, Policy and Development,
University of Reading
March 2021

Declaration

Declaration: I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Joe He

Acknowledgements

Firstly, I would like to thank my supervisors Richard Harrison and Michael Shaw for advice and guidance throughout my degree. I would also like to thank all the researchers at EMR for their assistance and input into my project. I am particularly grateful to Bo Li for assistance in strawberry imaging, Robert Vickerstaff and Andrew Armitage for assistance in bioinformatics, Helen Cockerton and Charlotte Nellist for advice on strawberry genotyping and Helen Bates for practical laboratory advice. I would also like to thank Amanda Karlstrom for aid in phenotyping and Sarah Hatcher and Alex White for assistance in data input and analysis. Finally, I would like to thank my fellow students who have made my experience at EMR all the more interesting and enjoyable.

Abstract

Strawberry is an economically important crop with global and UK production on an upwards trend. Strawberry breeding efforts attempt to generate novel varieties that have increased yields, resistance to pathogens, good eating quality and high nutritional content. Genomic prediction (GP) is an advanced marker assisted prediction (MAP) technique that makes predictions about agronomically important traits in crops. Three areas of improvement were identified to assist commercial deployment of GP for strawberry.

Breeding efforts currently rely primarily on visual and mechanical measurement of plant phenotypes, which is slow, imprecise and liable to human biases. A strawberry phenotyping platform was developed that captured images from 360° around the strawberry fruit to generate 3D representations. Seven fruit quality traits were calculated from the representations, which showed good concordance with manual measurements. Deployment of the system could lower phenotyping costs, increase throughput, increase precision and thus improve GP accuracy.

Current genotyping approaches for dense marker panels in strawberry are too expensive for commercial deployment. A rational design process was implemented to generate amplicon sets that would genotype a panel of markers to maximise information for GP, with scalability to accommodate resources available to different breeding programmes. The design process failed to generate marker information due to unexpected interaction in the multiplexed PCR reaction.

The relative effectiveness of phenotypic prediction and MAP in strawberries is unclear. Moreover, existing models of GP in strawberry do not represent all the traits of interest to breeders. Between years predictions of 15 fruit quality traits were implemented using phenotype only, traditional MAP (tMAP) and GP models. GP had similar selection accuracy compared to phenotype only prediction, but tMAP performed significantly worse than the other models. It was concluded that GP would likely yield benefits to strawberry breeding in the context of speed breeding.

Contents

Chapter 1

General Introduction

1.1 Strawberry Biology

1.1.1 Flower and Fruit

The primary role of the flowers is to attract pollinators. Flowers comprise 5 white petals, possibly tinged with pink or purple, surrounding 20 - 35 stamens, each covered with golden pollen. The stamen surround a conic receptacle, which is covered by 60 - 600 pistils arranged in a spiral pattern, each of which contains an ovule. At the base of the flower is a circle of 10 sepals, which develop into the calyx of the fruit [Heide et al., 2013, Kirsten, 2014]. The primary flower is the first to develop and mature, followed by secondary, tertiary, and additional flowers developing terminally from two or three branches formed beneath the primary flower on the main floral axis [Heide et al., 2013]. The primary fruit are the largest and most prized by growers; secondary fruit are smaller and only 80% of tertiary and 50% of quaternary flower buds develop to anthesis [Poling, 2012, Heide et al., 2013].

Flowers are typically hermaphrodites as strong selection has been applied to commercially grown varieties to allow ease of crossing for favourable varieties. Fertilisation occurs when pollen is brushed onto the receptacle and the ovules at the base of each pistil [Kirsten, 2014]. Both self-fertilisation and cross-fertilisation can occur, with cross-fertilisation conducted by insects, primarily wild bees, honey bees, and hoverfiles [Rader et al., 2016, Wietzke et al., 2018]. However, self-fertilisation results in smaller fruit, increased deformed fruit, increased sugar:acid ratios and reduced shelf life when inoculated with *Botrytis cinerea*. Overall, open fertilisation increases the value of the crop by 92% compared with self fertilisation [Wietzke et al., 2018]. Upon fertilisation, the receptacle swells and enlarges, becoming green and then red as the fruit ripens. Biologically, the achenes are the true fruit of the strawberry, with the surrounding red fleshy part being a modified shoot tip [Shulaev et al., 2011].

Strawberries are considered to be non-climacteric as they do not produce an ethylene burst upon ripening and are insensitive to ethylene [Symons et al., 2012, Villarreal et al., 2016]. During the ripening process, four plant hormones change drastically in profile, but it remains unclear which hormones are necessary and sufficient for the ripening process [Symons et al., 2012]. At present, there is no known procedure for artificially ripening unripe strawberries, so strawberries cannot be harvested unripe, stored until needed and

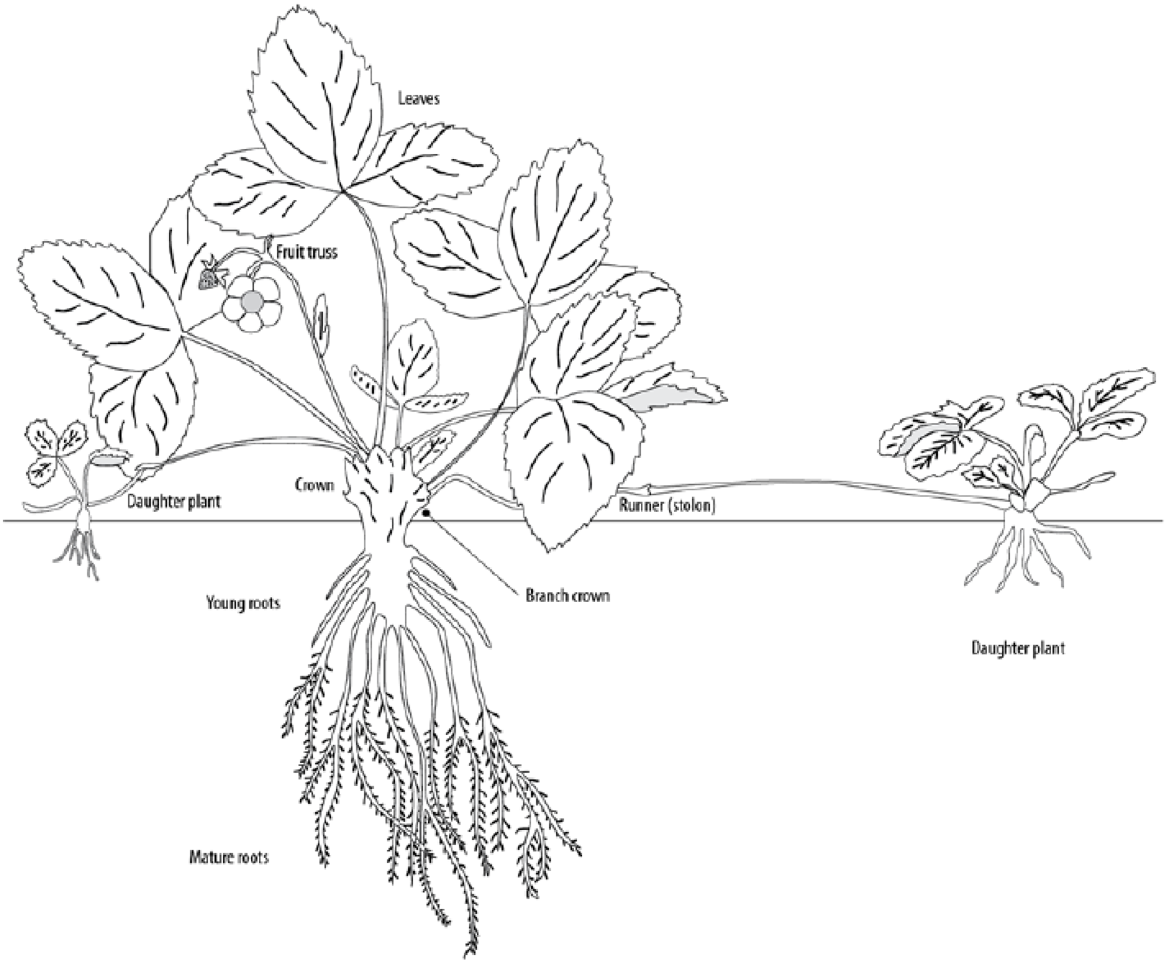


Figure 1.1: Diagram of a mature strawberry plant. Inflorescences bearing flowers, leaves and stolons emerge above ground from the crown and roots emerge below ground [Trejo-téllez and Gómez-merino, 2014].

artificially ripened.

1.1.2 Plant Habit and Life Cycle

A diagram of the mature dessert strawberry plant is presented (Figure 1.1). They have a shrub-like stature, with variable height dependant on environmental conditions and variety. The central crown is a non-woody modified stem, from which flower inflorescences, leaves and stolons emerge above ground and roots emerge below ground. Branch crowns split from the main crown and can have additional inflorescences [Kirsten, 2014].

The leaves are compound trifoliates and the primary site for photosynthesis and transpiration. The leaves are arranged in a spiral pattern around the central crown and have a phyllotaxy of 3-5 [Darrow, 1966]. They are evergreen in all known species of *Fragaria*, except *F. iinumae* [Liston et al., 2014]. Leaves will live for 1-3 months with older leaves typically dying during the winter and younger leaves grow to replace them [Poling, 2012]. Leaves are produced all seasons, but growth is particularly rapid in long days and is slowed by extremes of temperature [Kirsten, 2014]. Growth of leaves is primarily through cell enlargement and full growth of the leaf takes approximately 2-3 weeks [Darrow, 1966].

The roots of strawberries are fairly shallow and serve to gather water and nutrients for the plant as well as acting as a physical anchor. Strawberries have two types of roots. Primary roots are deeper and can persist for years, whist feeder roots typically have a lifetime of days to weeks [Kirsten, 2014]. The roots also serve as a storage for starch in winter [Poling, 2012].

Stolons are the asexual reproductive organs of the strawberry. They are a long shoot produced from the leaf axil during phases of vigorous growth and grow laterally along the soil. The daughter plants develop roots where they are in contact with the soil, making it a useful method used by breeders to propagate plants with identical genetic advantages of the mother plant. They contain developed xylem and phloem systems that allows transport of nutrients to developing ramets until their roots are able to support the plant [Savini et al., 2008]. Each plant can produce 10 - 15 runner chains per plant per year with three to five ramets per chain, allowing the production of up to 50 daughter plants in a season [Savini et al., 2008]. Runnering is controlled through hormonal and environmental cues, with activation by gibberellins and suppression by 8 to 12 short day cycles [Hytönen, 2009]. After several weeks, the stolon deteriorates and the new daughter clone continues in its life cycle [Rubinstein, 2015].

Strawberries can persist for several years. There are two main varieties of strawberries grown commercially, characterised by their growing habits. 'Short day' strawberries develop buds in the shorter days of late summer, overwinter and produce fruit around June in the Northern hemisphere or December in the Southern Hemisphere [Rubinstein, 2015]. 'Day Neutral' varieties do not have a photoperiodic requirement for flower bud initiation, with the first buds appearing approximately 3 months after planting [Rubinstein, 2015].

1.2 History and Geography

1.2.1 Taxonomy

There are 25 described species of the genus *Fragaria*, with *F. ananassa* being the most widely cultivated for human consumption [Johnson et al., 2014]. *Fragaria* is a member of the sub-tribe *Fragariinae*, which also includes 10 genera of over 400 species. *Fragariinae* species typically are non-woody, have pistils with lateral to basal styles, a basal chromosome number of x=7 and are polyploids [Lundberg et al., 2009]. *Fragariinae* is a member of *Potentilleae*, which contains several shrub-like genera, native to the arctic and alpine regions, and used as ornamentals and a source of pharmacologically active compounds [Tomczyk and Latté, 2009]. *Potentilleae* is a sister tribe of *Sanguisorbeae* and *Colurieae*, which, together with the commercially important genera of *Rosa* (ornamental roses) and *Rubus* (raspberries and blackberries) form the sub-family of *Rosoideae* [Longhi et al., 2014]. Phylogenetic analysis of nucleotide sequences across ten nuclear and chloroplastic regions suggests that *Rosoideae* along with *Dryadoideae* and *Amygdaloideae* are the three monophyletic sub-families that comprise *Rosaceae* [Potter et al., 2007]. *Dryadoideae* is a small sub-family comprising of just four known actinorhizal genera. *Amygdaloideae* (renamed from *Spiraeoideae* under recommendation from the International Code of Nomenclature) consists of a range of shrubs and trees and include the fruit crops of *Malus* (apples), *Pyrus* (pears), and *Prunus* (cherries, plums, apricots and almonds). *Rosaceae* is a major family of the Rosales order and comprises over 90 known genera and around 3000 known species, including a range of nutritionally, economically and ornamentally important crops [Christenhusz and Byng, 2016]. The Rosales order includes cannabis *Cannabaceae*, nettles *Urticaceae* and buckthorn *Rhamnaceae* [Zhang et al., 2011]. The overall nomenclature of strawberry is Plantae Magnoliophyta Magnoliopsida Rosidae Rosales *Rosaceae Rosoideae Potentilleae Fragariinae Fragaria.* The relatively close relationship between strawberry and several other economically and nutritionally important species makes strawberries a useful model organism.

1.2.2 Evolution

Phylogenetic analysis of chloroplastic genomes from 21 *Fragaria* species indicates that the last common ancestor (LCA) to *Fragaria* existed 1.0 - 4.1 million years ago and the LCA to the octoploid clade existed 0.37 - 2.0 million years ago [Njuguna et al., 2013]. This is in agreement with a previous study showing the LCA of *Fragaria* aged 2.7 million years ago and the octoploid clade 0.45 million years ago [Njuguna, 2010]. Analysis of the closely related genome of apple as well as molecular markers in *Prunus* and *Fragaria* for haploblocks suggests that the ancestral species to these nutritionally important members of Rosaceae had 9 chromosomes [Illa et al., 2011].

It is clear from early experiments in cytology that strawberries have a base chromosome number of 7, with the dessert strawberry being an octoploid [Ichijima, 1926]. Early studies into the origins of the polyploid were mostly based on observations of meiotic chromosome pairing [Badenes and Byrne, 2012]. With the advent of novel molecular tools, 43 accessions representing 14 species were genotyped at the nuclear ITS region and the chloroplastic *trnL* region. Maximum parsimony was used to cluster the accessions, suggesting that *F. vesca* and *F. iinumae* were likely donors of genomic material [Potter et al., 2000].

More recently, multiple studies have identified *F. vesca* and *F. iinumae* as donors of genomic material to the octoploid strawberry. A linkage group based study identified *F. vesca* and *F. iinumae* as donors with two unknown but closely related diploids being donors of the other genomes [Tennessen et al., 2014]. Analysis of the ADH1 gene from 19 *Fragaria* species identified two clades which suggested *F. vesca*, *F. mandshurica*, and possibly *F. bucharica* as donors to one and *F. iinumae* as the other [DiMeglio et al., 2014]. Analysis using SNPs from three biparental strawberry populations found *F. vesca* as a donor to each chromosome and *F. iinumae* as a donor to chromosome 4 [Vining et al., 2017]. This is consistent with studies indicating the mitochondrial donor for *F. ananassa* was derived from *F. iinumae* [Mahoney et al., 2010] and the chloroplastic genome was derived from *F. vesca* [Njuguna et al., 2013].

Recent sequencing and assembly of the strawberry genome has attempted to clarify the evolutionary origins of the other subgenomes in strawberry. Utilising 31 *de novo* assembled transcriptomes of every described diploid *Fragaria* species indicated *F. iinumae* and *F. nipponica* as contributors to two of the four subgenomes. The other contributors were likely *F. viridis* and *F. vesca* with *F. moschata* being a hexaploid intermediate [Edger et al., 2019]. However, further analysis found no evidence of *F. nipponica* and *F. viridis* ancestry in the octoploid, with the authors suggesting that the algorithm used by

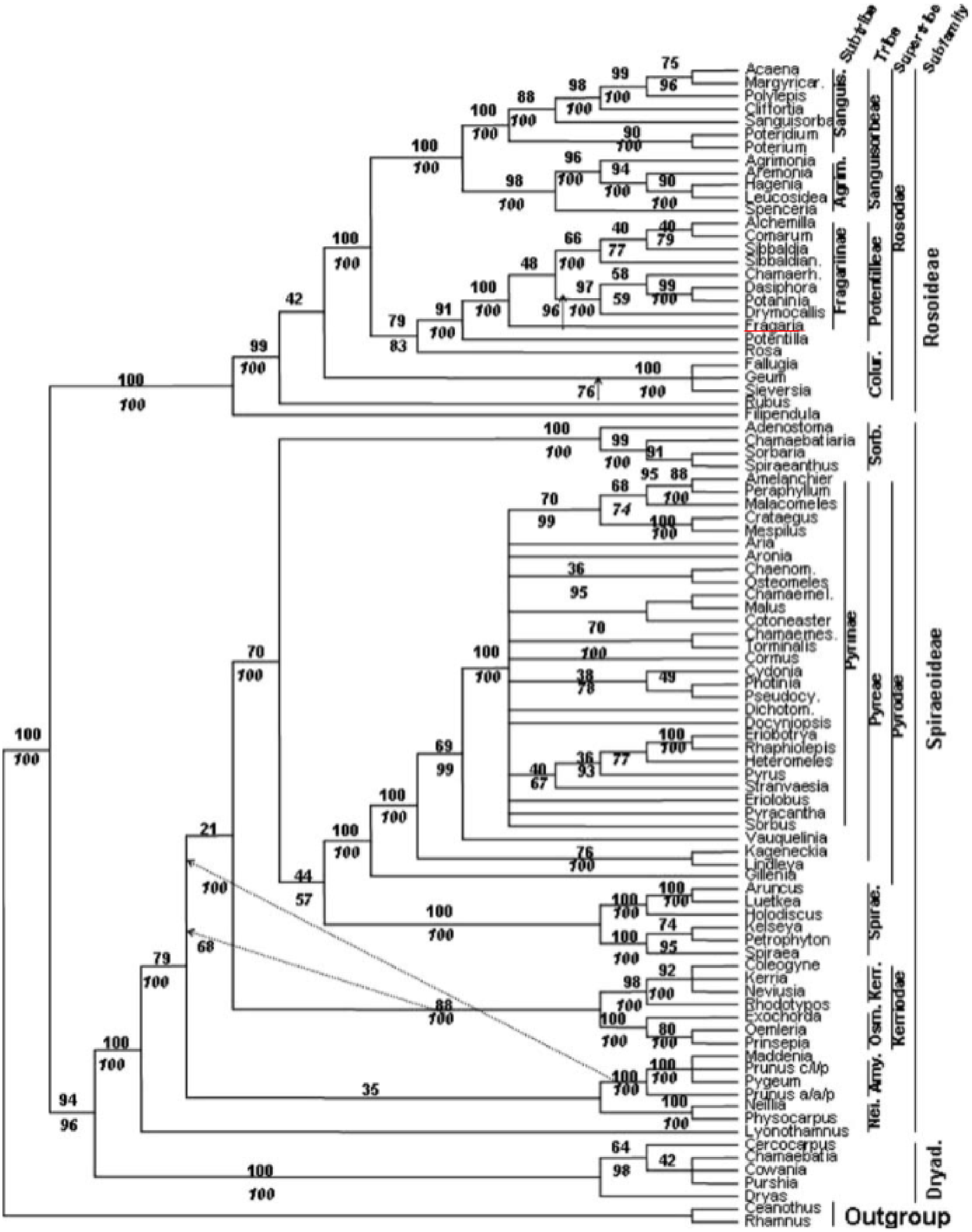


Figure 1.2: Greatest parsimony phylogenetic tree of 88 genera, based on nucleotide sequences from ten nuclear and chloroplastic regions. Fragaria genus underlined. [Potter et al., 2007]

Edger et al. to derive the ancestors ignored data suggesting multiple subgenomes to be derived from the same ancestor [Liston et al., 2020].

1.2.3 Distribution and Cultivation

Historically, members of the *Fragaria* genus have been consumed in Europe for hundreds of years. In Europe, the French cultivated *F. vesca* in gardens in the 1300s with the French king having over 1200 plants in the Royal Gardens. In England, numerous references were made by apothecaries and herbalists to its supposed medicinal properties in the 1500s [Darrow, 1966].

Due to the relatively recent emergence of the dessert strawberry as a cultivated species, we have historical records regarding its discovery. *F. ananassa* was the result of a natural hybridisation of *F. virginiana* and *F. chiloensis*, probably in northern Europe. It was first characterised by Antoine Nicholas Duschene in 1766 in the Royal Gardens in Brittany, France, who named it after the pineapple, after noting similar fragrances. *F. virginiana* originated from North America, though it was not cultivated there. We are unsure how the first clones arrived in Europe, but it appeared widespread by the late 1500s and new importations occurred frequently [Finn et al., 2013]. *F. chiloensis*, also known as the beach strawberry, are native to the beaches and mountains of central and southern Chile, Hawaii, and the coast of western North America from the middle of California to the Aleutian Islands [Hancock et al., 1999]. Five plants were brought to France by the spy Captain Amedee Frezier from the coast of Chile in the early 1700, where he was sent to identify Spanish colonial fortifications. These plants were presented to the king and grown in the King's Gardens [Darrow, 1966, Hancock et al., 1999].

Geographically, *Fragaria* is distributed mostly in the northern hemisphere in North America, Europe and Asia (Figure 1.3). They comprise of a range of ploidies ranging from diploid to decaploid. They occupy a range of elevations from sea level sand dunes to mountain meadows [Liston et al., 2014]. Within a bioclimatic niche, *Fragaria* display high levels of divergence, suggesting continuous evolution and a wide gene pool. The physical size of the geographic range that a species occupies does not appear to correlate with ploidy or phylogeny, but correlates with self-compatibility and the age of the species [Johnson et al., 2014]. These observations can be rationalised as self-compatibility likely allows the frontiers of a species range to expand in the absence of mates, whilst a new species often experiences a lag in range expansion due to population size, environmental and genetic factors [Crooks and Soule, 1999].

1.3 Consumption

1.3.1 Nutrition

Strawberries are popularly eaten across the world both fresh and processed. Fresh, they are consumed by themselves, as cakes, salads, in breakfast cereals, and dipped in chocolate. Processed, they are consumed as jam, drinks, ice cream, yoghurt, and sweets [Chandler et al., 2012, Siles et al., 2013, Kurotobi et al., 2010]. Moreover, strawberry derivatives are utilised in the manufacture of cosmetics, perfumes, air fresheners and pharmaceuticals [Siles et al., 2013].

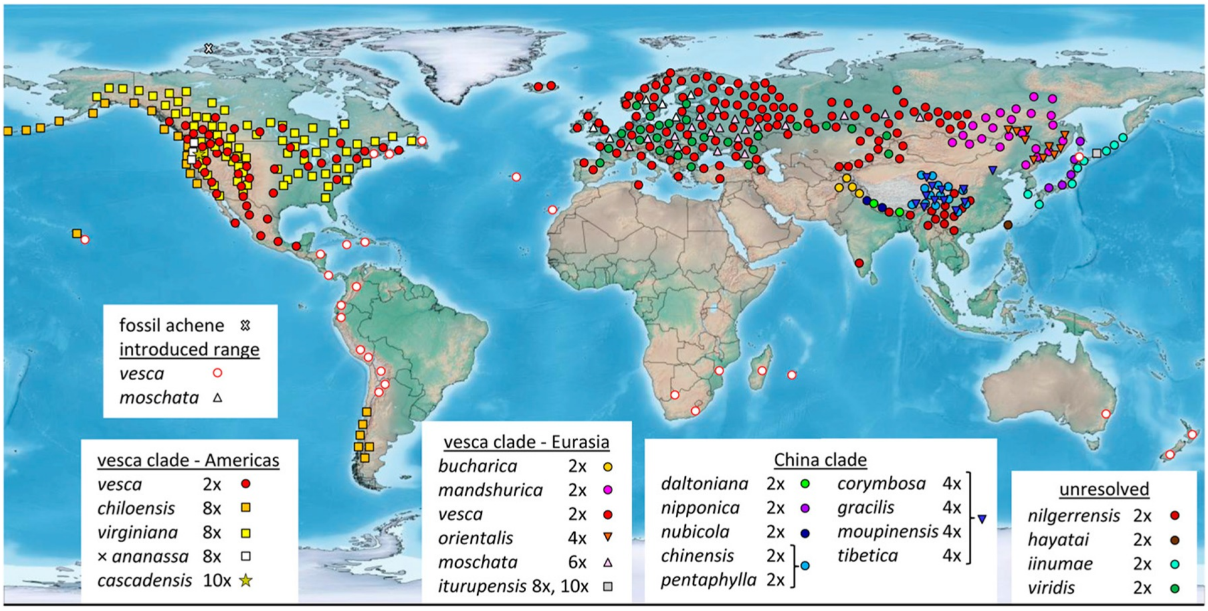


Figure 1.3: Global distribution of *Fragaria* species and their ploidies [Liston et al., 2014]
.

The flavour of food is the sum of the multiple senses that inform the brain what we are eating, including texture and the chemical profile that contribute to the taste and aroma of the food [Klee and Tieman, 2018]. Strawberries have complicated chemical profiles with mass spectrometry and chromatography studies identifying hundreds of volatile compounds with high dependence on the extraction method [Zorrilla-Fontanesi et al., 2012, Williams et al., 2005, Määttä-Riihinen et al., 2004], making identification of flavour profiles with high consumer acceptance challenging. Studies of consumer preference have consistently identified sweetness, and strawberry flavour as preferred, but other traits have been inconsistently identified [Schwieterman et al., 2014].

In a survey of 1137 consumers, consumers prefer strawberries with ideal red internal colour, with intense flavour, ideal external colour, and longer shelf-life [Choi, 2015]. Regression of metabolites to consumer acceptance identified 2 metabolite profiles with high consumer acceptance: high contents of straight chain ester and methyl anthranilate in combination with a balanced sugar-acid ratio and absence of methyl anthranilate with a sweet impression caused by sugars and the enhancing effect of some volatile organic compounds [Ulrich and Olbricht, 2016].

Strawberries are a valuable source of nutritionally important chemicals (Table 1.1). Notably, strawberries contains high concentrations of vitamin C, manganese, riboflavin, folate and dietary fibre. Vitamin C plays a role in the protection of cells through antioxidant activity [Duarte et al., 2009] as well as maintenance of bone mass [Gabbay et al., 2010] and support of collagen biosynthesis [Boyera et al., 1998]. Manganese functions as a cofactor for a variety of enzymes and is involved in development, digestion, reproduction, antioxidant defence, energy biosynthesis, immunity, neuronal activity and increasing bone mineral density [Bae and Kim, 2008, Chen et al., 2018]. Riboflavin is an important cofactor in a range of enzymes and is associated with iron absorption, maintenance of the cardiovascular system, development of the gastrointestinal tract, maintenance of the cornea and possibly protection from certain cancers [Powers, 2003]. Folate is required as a cofactor for DNA and RNA biosynthesis enzymes with clinical evidence showing protection from various cancers, in particular colorectal cancer [Moll and Davis, 2017, Mason and Tang, 2016]. Dietary fibre contributes to easing of constipation, reduction of blood sugar and cholesterol, reduction of calorie intake through satiation and the prevention of a range of cancers [Giampieri et al., 2012, Yangilar, 2013].

Nutrient	**RDI**	**per 100g Strawberries**	**Percentage of RDI**
Water/g		90.95	
Energy/kcal	3155	32	1.0 %
Protein/g	60	0.67	1.1 %
Carbohydrate/g	*	7.68	
Dietary Fibre/g	30	2.0	6.7 %
Sugars/g	*	4.89	
Calcium/mg	1000	16	1.6 %
Iron/mg	11.3	0.41	3.6 %
Magnesium/mg	300	13	1.3 %
Phosphorus/mg	775	24	3.1 %
Potassium/mg	3500	153	4.4 %
Sodium/mg	1600	1	0.06 %
Zinc/mg	9.5	0.14	1.5 %
Manganese/mg	2.3**	0.386	17 %
Copper/mg	1.0	0.048	4.8 %
Vitamin C/mg	40	58.8	150 %
Thiamine/mg	1.1	0.024	2.2 %
Riboflavin/mg	1.34	0.022	16 %
Niacin/mg	18	0.386	2.1 %
Vitamin B6/mg	1.5	0.047	3.1 %
Folate/µg	200	24	12 %

Table 1.1: Nutritional composition of selected compounds in strawberry [Giampieri et al., 2012]. The recommended daily intake (RDI) of nutrients is provided to indicate the potential of strawberries as a source of these nutrients. RDI data is calculated based on an 18 year old male weighing 80kg (British Nutrition Foundation). *RDI for carbohydrates and sugars is based on its contribution to energy rather than mass and is thus not comparable. ** Estimated average requirement is provided as RDI is not available [Food and Nutrition Board, 2002]

Strawberries also contain non-nutritive phytochemicals, which are bioavailable, bioactive, and possibly responsible for some of the health benefits of strawberry consumption [Afrin et al., 2016]. Polyphenols are the main group of phytochemicals and are associated with antimicrobial, antiallergenic and antihypertensive properties [Giampieri et al., 2012]. Anthocyanins are another major group of chemicals that may be beneficial to human health, but metabolism within the human body is complex [Felgines et al., 2003]. The mechanism of non-nutritive compounds from strawberries in human health remains an area of active research, but clinical and *in vitro* experiments clearly show the range of benefits associated with strawberry consumption.

1.3.2 Health

Significant research has been conducted on the effects of strawberries on the cardiovascular system, with clinical evidence suggesting that strawberry consumption has cardioprotective activities [Hooper et al., 2008]. Healthy volunteers supplemented with 500g strawberries per day over a month were found to have reduced biomarkers for cardiovascular disease risk (reduced low-density-lipoproteins, cholesterol and triglyceride levels in serum). Additionally, they were observed to have reduced haemolysis and increased active platelets, though the effect size was small [Alvarez-Suarez et al., 2014]. Processed strawberries retain benefits for the cardiovascular system. For example, twenty overweight subjects had their diets supplemented with freeze-dried strawberry powder over 3 weeks. The intervention benefited the volunteers' blood biomarker profile by reducing cholesterol and low-density lipoproteins and increasing high-density lipoproteins [Zunino et al., 2011]. In a longer term and larger scale study, 93600 women between the ages of 25 and 42 were tracked over 18 years through a combination of questionnaires and medical records. It was found that myocardial infarction rates were decreased when three or more servings of anthocyanin rich blueberries and strawberries were consumed per week [Cassidy et al., 2013].

Strawberry consumption has also been investigated for its effects on cancer, and has generally been found to have anticancer properties. Strawberry and strawberry extract have been demonstrated to inhibit transformation of cancer cells [Xue et al., 2001], induce apoptosis in leukaemia and breast cancer cells *in vitro*, and prolong lifespan of mice bearing breast adenocarcinoma [Somasagara et al., 2012]. In the A17 breast cancer cell line, strawberry extract modulates expression of various genes known to be associated with cell migration, adhesion and invasion and also greatly reduces viability in a dose dependant fashion. In mice challenged with A17 cells, the volume and weight of retrieved tumours were found to be significantly smaller in individuals on a diet enriched with 15% strawberry extract [Amatori et al., 2016]. Larger scale epidemiological studies further support the anticancer properties of strawberries. In a study of oesophageal cancer in the United States, there was a significant protective effect from consumption of members of the *Rosaceae* family [Freedman et al., 2007]. Consistent with this, another study found that consumption of members of *Rosaceae* was associated with reduced risks of head and neck cancer [Freedman et al., 2008].

Other studies show a range of additional potential benefits of strawberry consumption including anti-inflammation, neuroprotection, protection from oxidative stress, and short term amelioration of risk factors for diabetics [Afrin et al., 2016]. Human dermal fibroblasts generate a nuclear factor kappa-light-chain-enhancer of activated B cells (NFϰB) mediated inflammation response when exposed to *E. coli* lipopolysaccharide, which may lead to tissue damage and chronic inflammation diseases. Strawberry extract pre-treated cells exposed to lipopolysaccharide had lower inflammation biomarkers including NFϰB, increased enzymatic antioxidant activity, and reduced apoptosis compared to untreated cells [Gasparrini et al., 2018]. In a study of over 16000 women, intake of anthocyanidins was mostly attributed to strawberry and blueberry consumption. Greater intake of strawberries and anthocyanidins was associated to slower mental decline of cognitive ability after adjustment for a variety of confounding variables [Devore et al., 2013]. Strawberry consumption increases the serum antioxidant concentration significantly and decreases oxidative stress and DNA damage in cells [Cao et al., 1998, Pajk et al., 2006]. A significant reduction of blood pressure was observed in type II diabetics after freeze dried strawberry beverage was supplemented into their over six weeks [Amani, 2014].

There appears to be little evidence of direct negative health impacts of strawberry consumption for the general population. Cytotoxical studies indicate that human cell viability is only affected by exposure to high concentrations of strawberry extract for extended durations [Forbes et al., 2016]. Measurements of postprandial blood glucose levels in healthy volunteers show that high sugar concentrations in some strawberry jams were associated with high glycemic indexes, but no significant differences in blood glucose ratios were observed after 30 minutes. [Kurotobi et al., 2010]. Nonetheless, there are several health issues and perceived health issues surrounding the consumption of strawberries.

Firstly, as the fruit is frequently consumed unprocessed or minimally processed, there is the risk of contamination by human pathogens. Contamination can occur during growing, harvesting, post-harvest handling or distribution and occur through contact with contaminated animals, water, soil, equipment or human handling [Lafarga et al., 2018]. Characterisation of field grown strawberry microbial communities indicated the presence of known opportunistic human pathogens and fungal populations potentially capable of mycotoxin biosynthesis [Jensen et al., 2012]. Additionally, cases of disease outbreaks have been traced to contaminated strawberries, such as in east Germany in 2012. Over 11000 cases of gastroenteritis was reported in the largest food-borne outbreak in the country, despite a timely recall of more than half the contaminated batch. The outbreak was traced to frozen strawberries from China carrying norovirus [Bernard et al., 2014].

Secondly, there is concern associated with the consumption of ectopic pesticides associated with strawberries. Based on analysis of data from the US Department of Agriculture, strawberries are amongst the most pesticide contaminated crops in America (Environmental Working Group, 2016). Washing strawberries with water seems to be relatively ineffective at removing common pesticide residues, reducing pesticide concentrations by just 10% - 20% [Mee Kin and Guan Huat, 2010]. However estimates on human toxicity of strawberries suggests that the average reduction in disability adjusted life years from pesticide consumption is minimal. The nutritional benefits of fruit and vegetable consumption outweigh this considerably [Juraske et al., 2009].

Finally, there have been some documented cases of allergies to strawberries. Several allergens have been identified as binding strongly to immunoglobulins in putative allergics and elicits release of histamine associated with the allergic response. Some allergic reactions can be severe and include anaphylaxis and death [Zuidmeer et al., 2006, Karlsson et al., 2004].

1.4 Growth and Supply

1.4.1 Production, Economics and Distribution

Strawberries are grown across the world for human consumption, with *F. ananassa* being the dominant species [Johnson et al., 2014]. *F. chiloensis* was widely grown in Chile until the until the late 1800s and now is only grown to a small extent there. *F. vesca* is grown all across Europe, but primarily on a small scale as a garden variety [Hancock et al., 2008]. Additionally, there is a small market in Europe for the intensely flavoured hexaploid *F. moschata* [Darrow, 1966, Pet'ka et al., 2012].

Global production and consumption of strawberry is on an upwards trend (Figure 1.4), with China being the dominant producer and consumer. Within Europe, Spain and Turkey are the major producers of strawberry, producing 344679 tonnes and 440968 tonnes in 2018 respectively. In the UK, 131639 tonnes were produced in 2018, more than tripling production since 2000 (`www.FAO.org`). There are at least 384 different distinct named varieties of strawberries (patent search `https://worldwide.espacenet.com/` accessed 21/05/2020). The most grown variety in the world is 'Camarosa', a variety developed by the University of California, a publicly funded American breeding programme [Hancock et al., 2008]. In the UK, the most popular variety is 'Elsanta', a plant with high yields and firm berries with a long shelf life [Warner et al., 2010].

The average wholesale price of strawberries in the UK is £2.55 per kilogram (based on data 27/04/2018 - 13/07/2018) [DEFRA, 2018]. The UK produced £273 million worth of strawberries in 2018 (`https://www.gov.uk/government/statistics/latest-horticulture-statistics`). In 2015, the average strawberry yield in the UK was 22.3 tonnes per hectare, an increase of 125% over 20 years [Pelham, 2017]. On average, strawberries require 160 and 120 hours of labour per tonne for soil and substrate systems respectively, with strawberry picking amounting to approximately half the labour required to grow strawberries [Pelham, 2017].

In the UK, 85% of the demand for strawberries is during the peak summer months of May - September [Pelham, 2017]. Due to year-round demand and the non-climacteric nature of strawberry, the supply chain has adapted to produce and import strawberries from locations suitable for production throughout the year [Mezzetti et al., 2018]. During the summer months, the UK consumes its own supply of strawberry, while during the off season, strawberries are imported. In Europe, between January and March, strawberry production is primarily from Spain, Greece and Turkey, whilst from October to December, production comes from Belgium, the Netherlands and Italy [Mezzetti et al., 2018].

1.4.2 Growing Systems

Strawberries are suitable for growth on many soil types, ranging from sandy soil to heavier clays providing that nutrients, water and drainage are available. The matted row system is the simplest and requires the least capital investment and preparation. Plants are set in rows with a spacing of 45cm within soil rows and 1.5m between rows and allowed to establish for a year. Irrigation and application of straw may be deployed for winter protection and fertiliser application and weeding may be performed. Runners are allowed to establish and pinned down by hand to generate a 'mat' of plants over the establishment year and harvesting occurs from the second year onwards [Stevens, 2005]. Under this system, the daughter plants produce the majority of the crops, but a major drawback is that the establishment year is time inefficient.

In the plasticulture system, raised beds are formed, fumigated, and covered in plastic to prevent erosion and weed growth. Using black plastic causes the plants to bloom earlier by approximately two weeks and may increase cropping season duration. Strawberries

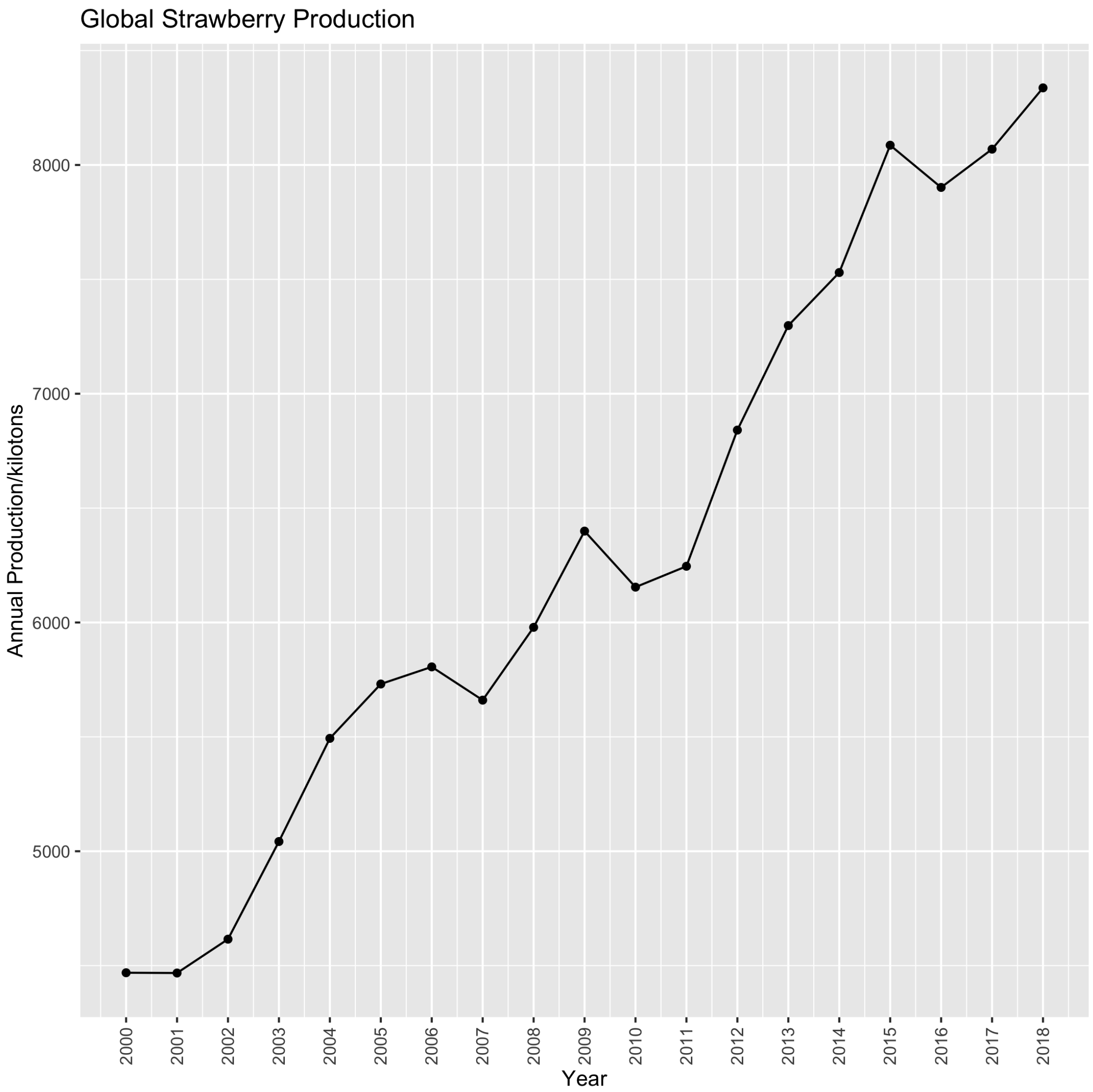


Figure 1.4: Global strawberry production by year (`https://http://www.fao.org/faostat/en/data/QC/`).

are planted in offset double rows in holes in the plastic with 1.5m between rows and 30cm between plants within rows [Fernandez et al., 2001]. Drip irrigation is provided through underground pipes and straw may be placed around the plants in Autumn for cold protection [Black et al., 2002]. This system is more popular in warmer areas and requires larger investment for the equipment [Fernandez et al., 2001, Kirsten, 2014], but can increase yields by 2.5 times.

In recent years, many farmers in the UK have moved towards annual crops and growth under protective cover and in soil free substrate, with 85% of the strawberry production in the UK under polytunnels (https://allmanhall.co.uk/blog/overview-of-the-strawberry-industry-in-the-uk). The substrate used is typically coconut coir for mitigation of soil borne fungal pathogens. Advantages of the soil free system include extension of the growing season, increased ease of picking and better control of the fertigation and pollination. These inputs are estimated to be more than doubled compared to a field grown crop [Boyer et al., 2016].

In order to minimise labour costs, automated strawberry picking machines are being developed. Strawberry picking robots should combine four subsystems: vision for detection, an arm for motion delivery, an end effector for picking, and finally a platform to increase the workspace [Xiong et al., 2018]. Algorithms have been developed for identifying the picking point, with up to 84% accuracy from single images, including berries obscured by leaves, but difficulties remain with identification of overlapping strawberries, which were not addressed [Huang et al., 2017]. The end effector needs to harvest the ripe fruit gently, but without slipping, and be tolerant of misidentification of the picking point [Dimeas et al., 2014]. Due to the difficulty of accurate detection of overlapping berries under a range of lighting conditions and orientations, a harvester that grasps groups of berries may be more suitable [Xiong et al., 2018]. However, trials of this system have a low accuracy of just 53 % in field, and a long duration of 10.6 seconds per fruit picking cycle [Xiong et al., 2019]. Including humans in checking can increase accuracy of a robot by providing feedback for further learning [Huang et al., 2020].

1.4.3 Environmental Impacts and Sustainability

Strawberry farming, like all agriculture, generates a range of environmental impacts. Strawberry production generates different amounts of carbon dioxide equivalent (CDE) greenhouse gasses, depending on the production system and the location and scale of the farm. Life cycle analysis indicates that strawberries produced in the USA in under the plasticulture system generated 1.75kg - 5.48kg of CDE per kilogram of strawberries produced, depending on the state [Tabatabaie and Murthy, 2016]. In the UK, analysis of 14 systems indicated that 0.13kg - 1.14kg CDE per kilogram of class 1 fruit generated, up to the harvest stage [Warner et al., 2010]. In Germany, strawberry generates approximately three kilograms CDE was generated per kilogram of strawberry produced, though there is considerable variation between best and worst case scenarios [Gunady et al., 2012]. Interestingly, farm vehicles contributed little to the CDE per unit strawberry as they had long lifespans and were used heavily; the most significant source of CDE production was associated with plastics, such as in the tunnel coverings [Tabatabaie and Murthy, 2016] and temperature maintenance of the greenhouse in these production systems [Gunady et al., 2012, Soode et al., 2015]. CDE production is lower per unit strawberry produced in coir than in soil as yields are higher in soil-less systems [Mordini et al., 2009] and is lower than other agricultural products such as asparagus, roses and orchids [Soode et al., 2015].

Life cycle analyses have identified eutrophication, ozone depletion, petrochemical ozone formation, acidification, carcinogen synthesis and water toxicity as additional environmental impacts [Romero-Gámez and Suárez-Rey, 2020, Tabatabaie and Murthy, 2016]. Synthesis of potassium fertiliser was primarily responsible for petrochemical ozone creation in its manufacture and eutrophication and acidification in its runoff. Carcinogen synthesis and ozone depletion was primarily due to plastic synthesis for coverings [Khoshnevisan et al., 2013, Tabatabaie and Murthy, 2016, Soode-Schimonsky et al., 2017].

Efforts have been made to mitigate the environmental impacts of strawberry production by utilising waste products from other industries, whilst possibly improving plant performance. For example, peat is often used as a substrate for production in nurseries, but there is concern about the environmental impacts of the destruction of peat bogs. Addition of up to 75% olive mill waste as a surrogate is an effective substrate providing the plant is fertigated with a nitrogen source [Altieri et al., 2010]. Addition of 3% biochar, a solid coproduct of biomass pyrolysis typically produced from energy generation, to peat increased the pH and significantly increased phosphorous, potassium, calcium, magnesium and total organic carbon concentrations. It was found that strawberries grown with biochar had a different rhizosphere microbiology, increased mass, upregulated defense related genes and increased resistance to grey mould, powdery mildew and anthracnose [De Tender et al., 2016, Meller Haral et al., 2012]. In the UK, strawberry is often grown in coconut coir, an inexpensive side product of the coconut industry. Addition of coconut coir dust to the growing media of strawberry significantly increases the sugar content and ascorbic acid content berry as well as the yield of the plant [Ayesha et al., 2011].

1.4.4 Pathogens and Symbionts

Significant costs are associated with pests and pathogens due to crop losses and the expense of treatments to manage the disease [Nellist, 2018]. Diseases can be classified by their causative agents as fungal, oomycete, bacterial and viral.

Fungal pathogens are the most significant in terms of loss of yield for strawberries [Nellist, 2018]. Powdery mildew (*Podosphaera aphanis*) is an obligate biotrophic airborne fungus, which spreads rapidly in high humidity and relatively cool temperatures, along with low light intensity. It has been an increasing problem recently due to developing fungicide resistance and movement towards tunnel production, which favours the pathogen [Asalf et al., 2014]. *Verticillium dahliae* damages the vascular system of strawberries, resulting in the characteristic wilt of the plant. Efforts to identify resistance have found multiple markers associated with resistance, which may be useful in marker assisted selection (MAS) and genomic selection (GS) [Cockerton et al., 2019]. Fusarium wilt is caused by *Fusarium oxysporum*, with symptoms being stunting of the young leaves, necrosis of the roots, wilting of the plant and eventually death [Nellist, 2018].

The only significant bacterial strawberry disease is angular leaf spot, caused by *Xanthomonas fragariae*. Symptoms begin with waterlogged lesions on the underside of leaves, and in severe cases the calyx, which enlarge and ooze with bacterial inoculum [Nellist, 2018]. Oomycete pathogens of strawberry include *Phytophthora* spp. and *Pythium* spp..

Phytophthora cactorum causes crown rot, and *Phytophthora fragariae* causes red core root rot [Nellist et al., 2019]. Viruses do not generally have clear symptoms in commercial plants, but cause stunted growth, decreased vigour and decreased yields [Nellist, 2018]. Strawberry mild yellow edge virus occurs worldwide and reduces yields by up to 30% [Thompson and Jelkmann, 2004]. Strawberry crinkle virus is transmitted through aphid vectors and results in reduced yield [Klerks et al., 2004].

There is the potential to use symbionts to improve strawberry production. Strawberry yield and first class berries were increased with the addition of arbuscular mycorrhizal fungi to its growth substrate, particularly under water and nitrogen limited conditions [Boyer et al., 2016]. Arbuscular mycorrhizal fungi were also observed to improve water use efficiency.

1.5 Breeding

1.5.1 Breeding Programmes

Artificial selection is the modification of a species by human intervention so that certain desirable traits are represented in successive generations (`https://www.dictionary.com/browse/artificial-selection`). Breeders have crossed plants exhibiting favourable characteristics for thousands of years; the resulting sexual recombination generates variation, upon which the breeder makes selections for individuals exhibiting combinations of desirable traits. Breeding is based on crossing germplasm material with agronomically important traits, such as high yield, and selecting for the most favourable offspring. These offspring are then trialled over several years, usually under different environmental conditions to confirm these traits before release. Usually, a new cultivar takes seven years to develop from breeding to commercial release, but may take up to 20 years [Badenes and Byrne, 2012]. Organisation of breeders into breeding programmes allows for sharing of resources and an increase of labour to achieve specific goals [Gallardo et al., 2014]. They share broadly similar aims of improving fruit yield and quality, pathogen resistance and productivity [Gallardo et al., 2014, Rubinstein, 2015].

Strawberry breeding programmes can be found all over the world, but are primarily concentrated in North America and Europe [Rubinstein, 2015]. For example, in 2010, there were 18 strawberry breeding programmes in USA and Canada [Gallardo et al., 2014]. Between 2006 and 2013, at least 76 new strawberry varieties have been released by breeding programmes in the US [Choi, 2015]. Funding for breeding programmes comes from a mixture of sources including governmental, private and royalties on intellectual property, with different programmes having different outlooks on the future of funding. Programmes have 1 - 9 full time equivalent workers, performs tens or low hundreds of crosses and typically screen tens of thousands of plants per year. Breeding programmes release 1 - 15 new varieties over a 5 year period, typically targeted for their local market [Knight et al., 2005, Gallardo et al., 2014]. However, recently there has been growth of larger breeders and private enterprises to dominate the supply chains for strawberries in some markets. For example Driscoll's breeds their own cultivars, produces the starts and grows the strawberries, holding a 34% share of the conventional strawberry market in the USA [Rubinstein, 2015].

Due to the difficulty in identifying a chemical profile that has high consumer acceptance, difficulty and expense in quantifying a strawberry chemical profile [Schwieterman et al., 2014], difficulty in identifying hereditary markers associated with chemical profiles and the low value that consumers preference has in determining strawberry market value[Gallardo et al., 2014], improvement in flavour is often not the highest priority in breeding programmes. A survey of 8 strawberry breeding programmes in North America indicated the most important trait was post-harvest quality, followed by yield and texture [Gallardo et al., 2014]. In another survey of 86 producers and 1137 consumers, it was identified that producers prefer to grow firm strawberries with intense flavour, and ideal external and internal red colour, while consumers prefer strawberries with ideal red internal colour, with intense flavour, ideal external colour, and longer shelf-life [Choi, 2015].

For breeding strategies, it is desirable to understand the method of inheritance for strawberry. Low resolution genetic maps have provided some evidence of polysomic inheritance, but more sensitive experiments to date have only found evidence of disomic inheritance. Analysis of 4 microsatallites in *F. virginiana* was found to be consistent with disomic Mendelian inheritance [Ashley et al., 2003]. A genetic map for *F. ananassa* was generated using 148 molecular markers. In the 42 linkage groups where markers in both coupling and repulsion phase was found, there was a 1:1 ratio of coupling and repulsion phase markers in resulting recombinant progeny, consistent with disomic inheritance.

1.5.2 Genetic Engineering

The development of modern gene theory and transformation technologies has allowed for the insertion of favourable genes from any domain of life into strawberries as a means to generate variation. This targeted insertion is particularly attractive for the pyramiding of monogenic traits or traits with a small number of quantitative trait loci (QTLs) rapidly and without compromising existing characteristics [Passey et al., 2003].

Transformation of strawberries to resist a range of pests and diseases has seen success and shows promise in reducing chemical controls required [Qin et al., 2008]. Insertion of the cowpea trypsin inhibitor into strawberries resulted in plants having up to 362 % greater root weight compared to the control when exposed to the vine weevil *Otiorhynchus sulcatus* [Graham et al., 1997]. Transformation of a chintinase into strawberry resulted in plants that were significantly less susceptible to *V. dahliae* [Chalavi et al., 2003]. More recently, strawberries transformed with the *Arabidopsis thaliana* NPR1 gene showed increased resistance to anthracnose, powdery mildew and angular leaf spot [Silva et al., 2015].

Strawberries have also been genetically engineered to be resistant to abiotic stresses. Using *Agrobacterium* mediated transformation, antifreeze protein from fish has been inserted into strawberry, though no experimental evidence of cold resistance was presented [Khammuang et al., 2005]. Transformation of the wheat acidic dehydrin gene into strawberry resulted in plants that were able to resist ion leakage at temperatures 5°C lower than the untransformed control [Houde et al., 2004].

Despite the promise of utilising genetic modification in strawberry breeding, there are no known large scale commercially available genetically engineered strawberries. One major obstacle is the reluctance of the public to accept consumption of transgenic strawberries [Schaart et al., 2011]. Additionally, there are issues with low efficiencies of transformation, difficulty identifying and isolating genes for transferring into strawberry and variable expression after transformation [Qin et al., 2008]. Some of these issues have been alleviated in strawberry, with the assembly of the *F. vesca* genome [Shulaev et al., 2011] and more recently the *F. ananassa* genome [Edger et al., 2019]. Transformation has been implemented with the CRISPR transformation system allowing precise transformations with up to 80% efficiency [Wilson et al., 2019]. CRISPR has been used to edit the TM6 gene responsible for petal and anther development in strawberries [Martín-Pizarro and Posé, 2018].

1.6 Aims of this Investigation

The output from strawberry breeding programs are primarily novel cultivars that are superior to currently available varieties. In order to develop novel cultivars, specific alleles of genes controlling agronomically important traits must be selected. Marker assisted breeding (MAB) generates statistical models that associate traits of interest with genetic markers and subsequent selection on these markers allows for selection of traits. Genomic prediction (GP) is a novel approach to MAB, which has the potential to increase the genetic gain per unit time for strawberry breeding over existing traditional methods of marker assisted breeding (tMAS).

Three areas were identified for improvement to assist in deployment of GP for strawberry, which comprise the three results chapters of this thesis:

1. Strawberry phenotyping in commercial breeding is currently slow, imprecise and liable to human biases [He et al., 2017]. Assessment typically relies on the human eye to quantify traits such as achene density, shape and colour [Mathey et al., 2013], which could be more precisely measured using automated electronic remote sensing instruments. In this section, an automated 3D phenotyping system for measuring strawberry fruit quality traits will be developed to overcome some of these limitations.

2. Cost-benefit models of strawberry breeding programmes indicate that currently available marker panels are too expensive for deployment in genomic selection (GS) breeding programmes [Wannemuehler, 2018]. Studies indicate that reduction of markers from tens of thousands to thousands and potentially hundreds still allows modestly accurate selection under the GS breeding scheme [e Sousa et al., 2019]. Imputation of missing markers is effective for marker reduction in GS, but little evidence is available for durability of the markers when the genetics of the breeding population changes (such as through introgression of novel material). In this section, a custom algorithm will be designed to maximise the information likely to be gained from markers for GS by optimising the biological properties of markersets. Experimental evidence for efficacy of the markerset for GS will be generated using an open genotype resolution by sequencing approach.

3. There is limited evidence of the efficacy of GS in strawberries. Existing studies only model five strawberry traits [Gezan et al., 2017, Osorio et al., 2021] of the scores of

traits of agronomic import [Mathey et al., 2013]. In this section, GP models will be generated for 15 strawberry fruit quality traits of agronomic importance using data gathered 2013 - 2016. The accuracy of selection will be compared to tMAS and a genotype data free linear prediction model in a biparental mapping population.

Chapter 2

A High-throughput 3D Phenotyping System for Strawberry

2.1 Preface

The contents of this chapter (with minor edits) have been published in the peer-reviewed journal *Plant Methods*, under the title ‘A novel 3D imaging system for strawberry phenotyping’ [He et al., 2017]. The software developed and datasets generated are available from the Image-processing repository at the East Malling Github site (`www.github.com/organizations/eastmallingresearch/`). My contribution to this manuscript was to assist in experimental design, conduct the manual assessment of strawberries, perform the statistical analysis and preparation of the manuscript. Bo Li devised the experiments, developed software and guided data analysis. Richard Harrison assisted in experimental design and editorial oversight.

2.2 Abstract

Accurate and quantitative phenotypic data in plant breeding programmes is vital in breeding to assess the performance of genotypes and to make selections. Traditional strawberry phenotyping relies on the human eye to assess most external fruit quality attributes, which is time-consuming and subjective. 3D imaging is a promising high-throughput technique that allows multiple external fruit quality attributes to be measured simultaneously.

A low cost multi-view stereo (MVS) imaging system was developed, which captured data from 360° around a target strawberry fruit. A 3D point cloud of the sample was derived and analysed with custom developed software to estimate berry height, length, width, volume, calyx size, colour and achene number. Analysis of these traits in 100 fruits showed good concordance with manual assessment methods.

This study demonstrates the feasibility of an MVS based 3D imaging system for the rapid and quantitative phenotyping of seven agronomically important external strawberry traits. With further improvement, this method could be applied in strawberry breeding programmes as a cost-effective phenotyping technique.

2.3 Background

A successful strawberry breeding programme generates and selects genotypes with traits suitable for the industry in its target geographic region [Mathey et al., 2013]. As often genotypes cannot be directly observed, traditional breeding selects on the basis of a weighted selection index of phenotypes [Badenes and Byrne, 2012]. In order to maximise the accuracy of selection, heritable traits of interest must be measured precisely and accurately. Currently, most external fruit quality phenotyping approaches in strawberry breeding relies on the human eye to make assessments [Mathey et al., 2013]. This approach is labour-intensive, prone to human bias and generates ordinal data less suitable for the most powerful quantitative statistical models [Goddard and Hayes, 2007].

Use of image analysis has the potential to overcome some of these limitations, with previous studies showing success in utilising 2D high-throughput imaging systems to assess external fruit quality [Dadwal and Banga, 2012]. Most studies were focussed on colour analysis of fruits, including apple [Throop et al., 2005], citrus, [Blasco et al., 2007], mango [Kang et al., 2008] and banana [Mendoza and Aguilera, 2004], but some systems have assessed morphological attributes, including the size of apples [Blasco et al., 2003] and the shape of oranges [Costa et al., 2009]. For strawberry, an automated grading system was developed that assesses colour, size and four degrees of shape [Xu and Zhao, 2010]. In another 2D strawberry imaging system, the maximum fruit diameter could be derived by automatically identifying the axis from the top of the calyx to the tip of the nose [Nagata et al., 2000]. However, 2D image analysis is not always a reliable fruit phenotyping method due to uneven colour distribution and occlusion of morphology from different viewing perspectives [Paulus et al., 2014].

Recently, 3D imaging has been increasingly explored as cost of hardware decreases and reassembly techniques improve [Vázquez-Arellano et al., 2016], with a range of sensors deployed for plant phenotyping. Light detection and ranging (LiDAR) was used to generate detailed 3D models of plants [Kjaer and Ottosen, 2015, Paulus et al., 2013], but is currently expensive, time consuming and complex to implement [Zhang et al., 2016]. Binocular stereovision is a low-cost solution for 3D plant canopy reconstruction [Klodt et al., 2015], but with only two viewing perspectives, is insufficient to model the entire target. Other techniques, including time of flight [Klose et al., 2011, Alenyà et al., 2011] and structured light [Chéné et al., 2012] have similar limitations in gathering 360° data from the target.

Studies of 3D based phenotyping of fruit are limited. A 3D model of mango was generated using four cameras and the shape-from-silhouettes reconstruction method, but it did not encompass 360°of the fruit. Five parameters were extracted from the 3D model including length, width, thickness, volume and surface area in order to sort the mangoes by size. Image based sorting accuracy was comparable to manual sorting, but no comparison was of individual trait data to a 'gold standard' was presented [Chalidabhongse et al., 2006]. Shapes-from-silhouettes was also successfully applied to the 3D reconstruction of tomato seedlings with ten calibrated cameras. Stem height and leaf area were accurately measured after geometry segmentation [Golbach et al., 2015].

Multi-view stereovision (MVS), which originated from binocular stereovision, is a promising approach for fruit phenotyping by capturing images from multiple overlapping viewpoints [hristian Rose et al., 2015]. For the determination of the intrinsic camera parameters and the positions of uncalibrated cameras, Structure from Motion (SfM) is a widely used technique (Figure 2.1). SfM detects feature points, called keypoints, from all the input 2D images using Scale-invariant Feature Transform (SIFT) algorithm [Lindeberg, 2012]. The number of keypoints is determined by image quality, including resolution and texture. The relative pose and camera locations are determined by matching keypoints across all images and iteratively refined by bundle adjustment, resulting in a point cloud [hristian Rose et al., 2015]. The coordinates generated by SfM is in an arbitrary image space, making it necessary to transform the coordinate system into an object space using a known standard [Westoby et al., 2012]. This method has been demonstrated to be low cost, highly precise, easy to implement, generate 360° colour information, and require no camera calibration. MVS and SfM have been successfully utilised to generate estimates of leaf and stem dimensions of paprika [Zhang et al., 2016].

In this study, a novel 3D imaging based approach for phenotyping fruit was explored. MVS and SfM were applied to generate 3D models of strawberry and software was developed to measure seven agronomically important external strawberry traits. This method is promising to facilitate strawberry breeding by providing a high-throughput, objective and low-cost phenotyping system.

2.4 Methods

2.4.1 Fruit Material

100 strawberries were purchased from local supermarkets, including 10 different varieties, to represent the diverse range of commercially available phenotypes. Fruits would likely have been subject to chilling to 4°C within four hours of harvest and kept at that temperature throughout the supply chain until sale. Fruits were stored at 4°C until assessment and were assessed before their 'use by' dates.

2.4.2 Manual Assessment

In order to validate the results of the 3D analysis, phenotypic data were collected manually immediately after imaging. Measurements of dimensions were performed using a pair of digital callipers and measurement of volume was performed using an overflow can and measuring cylinder (Table 2.1).

2.4.3 Image Capture

The sample was pinned onto a dark blue holder (38mm ×19mm × 19mm) fixed in the centre of a turntable rotating at 0.02Hz. A single lens reflex (SLR) camera (Canon EOS 1200D, Canon Inc., Tokyo, Japan) was placed facing the sample with a focal length of 55mm so that the field of view is large enough to accommodate the largest strawberry sample. The distance between the lens and the sample was set to 50cm with a viewing angle of 35°to the horizontal, which allows for maximum visualisation of the strawberry body without occlusion of the calyx. The relative positions of the camera and holder was fixed for all samples. The sample was illuminated with two white LED light sources

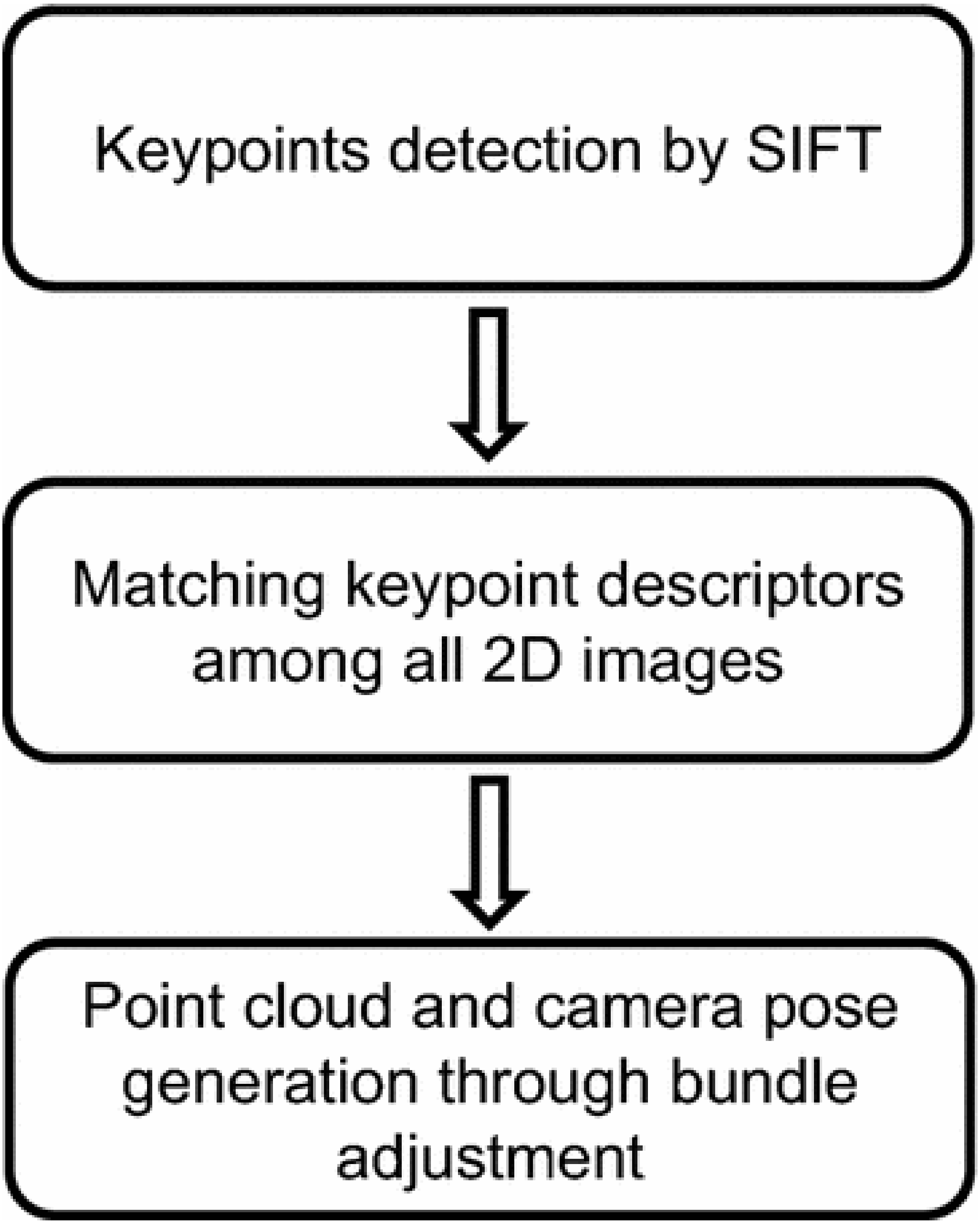

Keypoints detection by SIFT
Matching keypoint descriptors among all 2D images
Point cloud and camera pose generation through bundle adjustment

Figure 2.1: a flowchart of the Structure from Motion algorithm [He et al., 2017]

.

External Quality Parameter	Scoring Metric
Achene Number	Number of achenes visible, without disturbing calyx
Calyx size	Maximum Euclidean distance between any pair of points on the calyx
Colour	Scale 1 - 8 (Strawberry colour chart for experimental ends, Ctifl, France)
Height	Dimension of fruit from centre of calyx to tip of nose
Length	Greatest dimension of fruit orthogonal to the height
Width	Greatest dimension of the fruit orthogonal to both height and length
Volume	Volume of displaced water when fruit was completely submerged

Table 2.1: Manual assessment of seven external strawberry quality parameters [He et al., 2017]

against a white background (Figure 2.3). In total, 146 images were captured per sample over 50 seconds (ISO speed rating at 800, shutter speed at 1/125 seconds, aperture value at 5.38 EV). With this configuration, no blurring was found in any image.

3D Point Cloud Reconstruction

A Dell desktop computer (CPU Xeon X5560 @2.80 GHz × 16, Intel Co., Santa Clara, CA, USA) with a graphics card (Quadro K2200 GPU, NVIDIA Co., Santa Clara, CA, USA) operating with Linux Ubuntu 14.04 was used in this study, for both software development and point cloud processing.

The point cloud reconstruction was implemented with commercial software (Agisoft Photoscan, Agisoft, LLC, St. Petersburg, Russia; licence required), utilising the Structure from Motion (SfM) algorithm [Zhang et al., 2016]. Due to the high overlap between adjacent images and high resolution (5148 × 3456) of each image, pre-processing software was developed to automatically reduce the number of images by discarding three frames from every four. This was found to be the minimum number to reconstruct all the 3D models successfully. Additionally, each image was rescaled to the resolution of 1000 × 1450, which greatly increased the processing speed with satisfactory point cloud quality (Figure 2.3).

Automated 3D Image Analysis

The automated point cloud analysis software was developed in C++ with Point Cloud Library (`http://pointclouds.org/`). The software was programmed to automatically load all point cloud files in order and process them in a batch by implementing the point cloud segmentation and external quality attributes measurement algorithms.

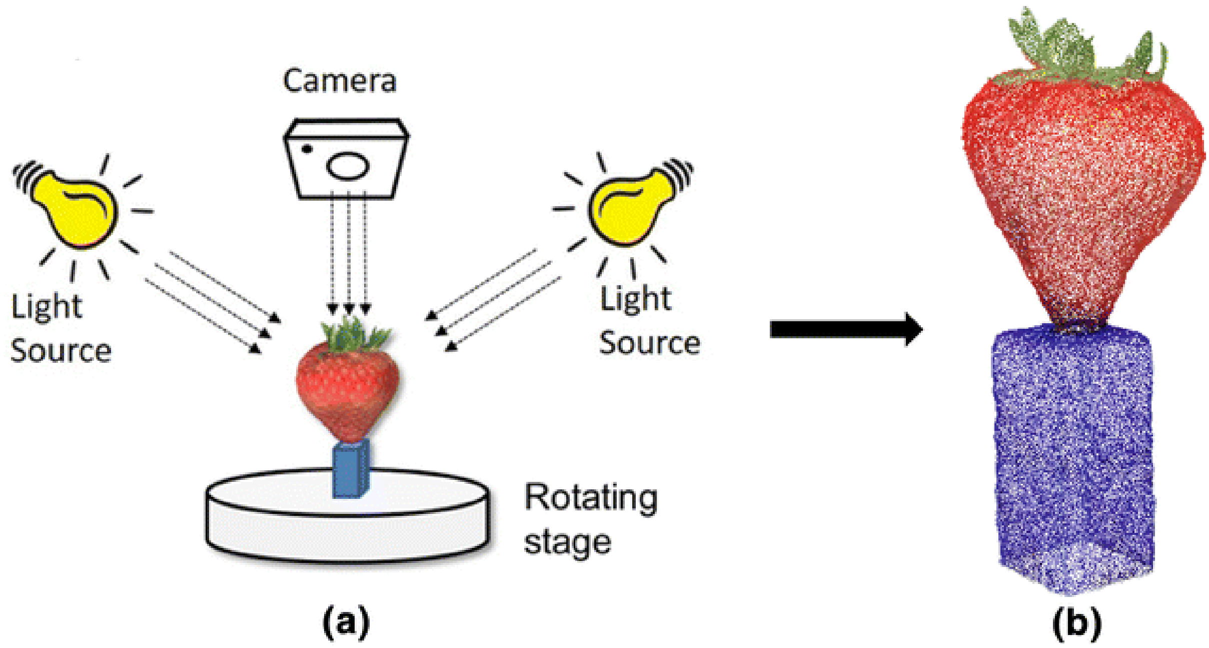


Figure 2.2: (a) schematic of image capture (b) 3D point cloud generated using captured images [He et al., 2017]
.

Each point cloud was first converted from Red Green Blue (RGB) space to Hue Saturation Value (HSV) space. Using an arbitrary threshold on the hue channel, which is defined as the attribute of a visual sensation to one of the perceived colours [Wu and Sun, 2013], the point cloud was segmented into calyx, body, achenes and holder.

The orienting Bounding Boxes (OBBs) is the box with the smallest volume that encloses all the points in the point cloud. The major eigenvectors of the covariance matrix of points in a point cloud define the major axis of its OBB [Ding et al., 2004]. The second axis was determined by calculating the maximum Euclidean distance of the points in the point cloud orthogonal to the major axis. The final axis was orthogonal to both other axes.

Height, Length and Width: An OBB was fitted to the point cloud of the combination of the fruit body and holder segments. The OBB was not fitted directly to the body, as its irregular shape often resulted in misidentification of the vertical axis. The height of the combination of fruit body and holder was always the largest dimension, so the magnitude of the OBB major axis was assumed to be equivalent to the height of the fruit body and holder model. As the fruit body was always longer and wider than the holder, the second and third dimensions of the OBB represented length and width respectively. The height of the holder was estimated by fitting an OBB to its point cloud and the difference in height between it and the combination of fruit body and holder OBB was assumed to be the height of the fruit. Ratios between the three fruit body dimensions and the height of the holder were multiplied by the true height of the holder to derive the strawberry height, width and length.

Volume: The mesh of the strawberry body was constructed from the point cloud using Poisson Surface Reconstruction [Kazhdan et al., 2006], which produces an enclosed

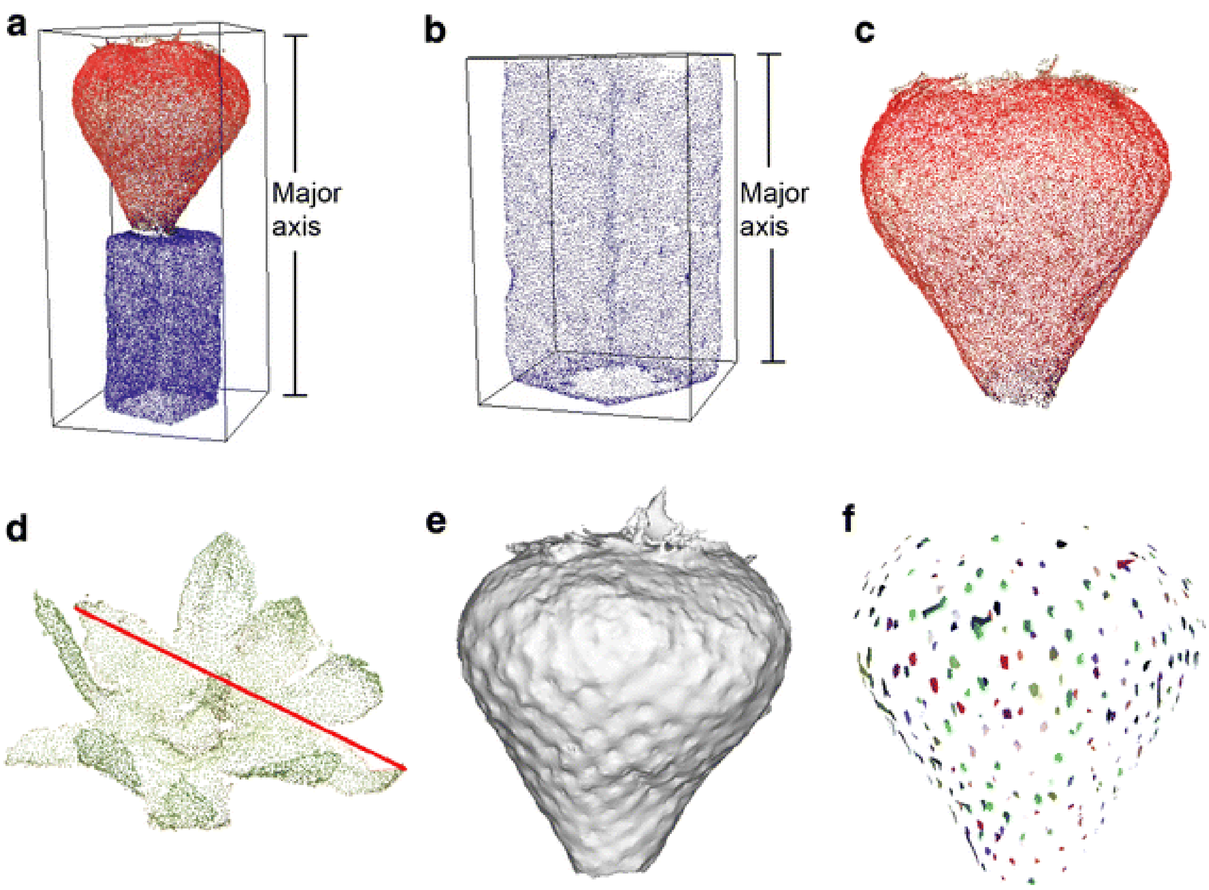


Figure 2.3: (a) OBB fitted to 3D point cloud with determination of major axis (b), (c), (d) Segmentation of holder, fruit body and calyx respectively. Red line indicates maximum Euclidean distance between points in segment (e) mesh of strawberry for volume calculation (f) estimation of achene locations [He et al., 2017]

.

mesh without and edges or large holes. The mesh volume was calculated by summing the volume of every triangle based pyramid formed from each face of the mesh and the origin of the point cloud [Zhang and Chen, 2001].

Calyx Size: The edges of the calyx segment were identified by applying convex hull [Cupec et al., 2011], enabling rapid calculation of maximum Euclidean distance. The ratio between the calyx maximum distance and the height of the holder OBB was multiplied by the true height of the holder to estimate calyx size.

Achene Number: The segmentation of achenes from the point cloud was based on identifying points in the body segment with an arbitrary range in the hue channel of HSV space. A clustering algorithm based on Euclidean distances between points was implemented to group points corresponding to the same achene [Dixon and Brereton, 2009] and the number of clusters was counted automatically.

Colour: As hue value in HSV space represents visual sensation of perceived colour [Wu and Sun, 2013], the mean intensity of the hue channel of every point in the body segment was calculated for the assessment of the strawberry colour.

Statistics

In this study the concordance correlation coefficient (CCC) was used to measure the concordance between the manually derived and 3D image derived external fruit quality traits [Lin, 1989]. Additionally, the coefficient of determination (r^2) was calculated to estimate correlation between the sets of values. Statistical analysis was performed using R [R Core Team, 2017]. Linear models and associated coefficents were derived using the 'lm' function; the root mean square error (RMSE) was derived using the 'Metrics' package [Hamner, 2012] and the CCC was derived using the 'Agreement' package [Yu and Lawrence, 2012].

2.5 Results

In order to evaluate the measurement of seven external strawberry fruit quality parameters using 3D imaging (hereafter referred to as automated assessment), 100 berries were automatically and manually assessed. Reliable reconstruction could be achieved by taking a minimum of 37 images per berry with 100% successful reconstruction, though the nose of the fruit was often missing due to occlusion from the shooting angle. With the described setup, data capture took approximately 60 seconds, including 10 seconds of operator action per sample. Model reconstruction took approximately 15 minutes and parameter derivation took approximately 50 seconds. Both these operations were fully automated.

In order to validate the 3D reconstruction, the point cloud of the holder segment was manually measured in Meshlab [Cignoni et al., 2008], an open source software for 3D mesh visualisation. Although their absolute sizes in image space were inconsistent (range: 0.36 1.73; mean: 0.78; standard deviation 0.27), ratios among the height, width and length were the similar to the true ratios. As there was no evidence of distortion, the absolute height of the holder was used as a standard for fruit dimension measurements. Moreover, incorporation of the holder point cloud ensured that the vertical axis was always greater than any other axis, allowing the major eigenvector of the point cloud covariance matrix to consistently define the vertical axis.

To validate the measurements, the seven traits were measured on a sample of 100 fruits using both manual and automated assessment (Figure 2.4). Concordance and correlation were assessed using CCC and r^2 respectively. Good concordance (CCC >0.9) and correlation ($r^2 > 0.9$) were found between the measurements of fruit dimensions and volume. Weaker concordance (CCC = 0.86) and correlation ($r^2 = 0.87$) was found between the measurements of calyx size, which was possibly due to the soft calyx being moved during assessment. Weak concordance (CCC = 0.67) and correlation ($r^2 = 0.77$) were found between the measurements of achene number, which is possibly due to the lack of information gathered regarding the nose of the fruit. Weak correlation ($r^2 = 0.68$) was found between the measurements of colour, with high variance in the manual scores. This was likely due to the variability of colour on each fruit and the subjective nature of the score.

2.6 Discussion

Good concordance between manual and automated measurements of calyx size, height, length, width and volume, and promising results for achene number and colour were achieved. It is suggested that the qualitative traits of strawberry currently used in breeding can be represented using the measurements generated from this study. For instance, a 'long conic' [Mathey et al., 2013] fruit has a large ratio of height to width and measurement of 'cap size' [Mathey et al., 2013] can be defined by the ratio of calyx size to fruit width and length.

With further development, automated assessment could be suitable for integration into existing strawberry breeding programmes, bringing a range of advantages. Firstly, the quantitative, accurate and unbiased measurements would increase the accuracy of the selection in strawberry breeding. The precise measurements would be particularly suitable for input into models of genomic selection, which attempt to quantify small effect quantitative trait loci (QTLs) associated with polygenic traits [Meuwissen et al., 2001, Gezan et al., 2017]. Secondly, automated assessment has the potential to improve the speed of assessment. The described setup requires approximately 10 seconds of human operator time per sample, approximately 20-fold faster than making the equivalent manual measurements. Thirdly, the low cost and wide availability of hardware mean that this approach can be easily introduced into existing breeding programmes with minimal capital expenditure.

Measurement error may have arisen from a range of sources. During manual assessment, the axis of measurement was determined by eye, potentially resulting in non-maximal distances or non-orthogonal axes. As the calyx is soft, errors may have been induced in the operation of the callipers. Correlation between the measurements of colour may be weak as manual assessment is subjective and it is difficult to assess fruit with uneven colour distributions.

This imaging system can be developed to reduce the duration of data capture by using alternative imagers such as scientific cameras or webcams with programmable shutter

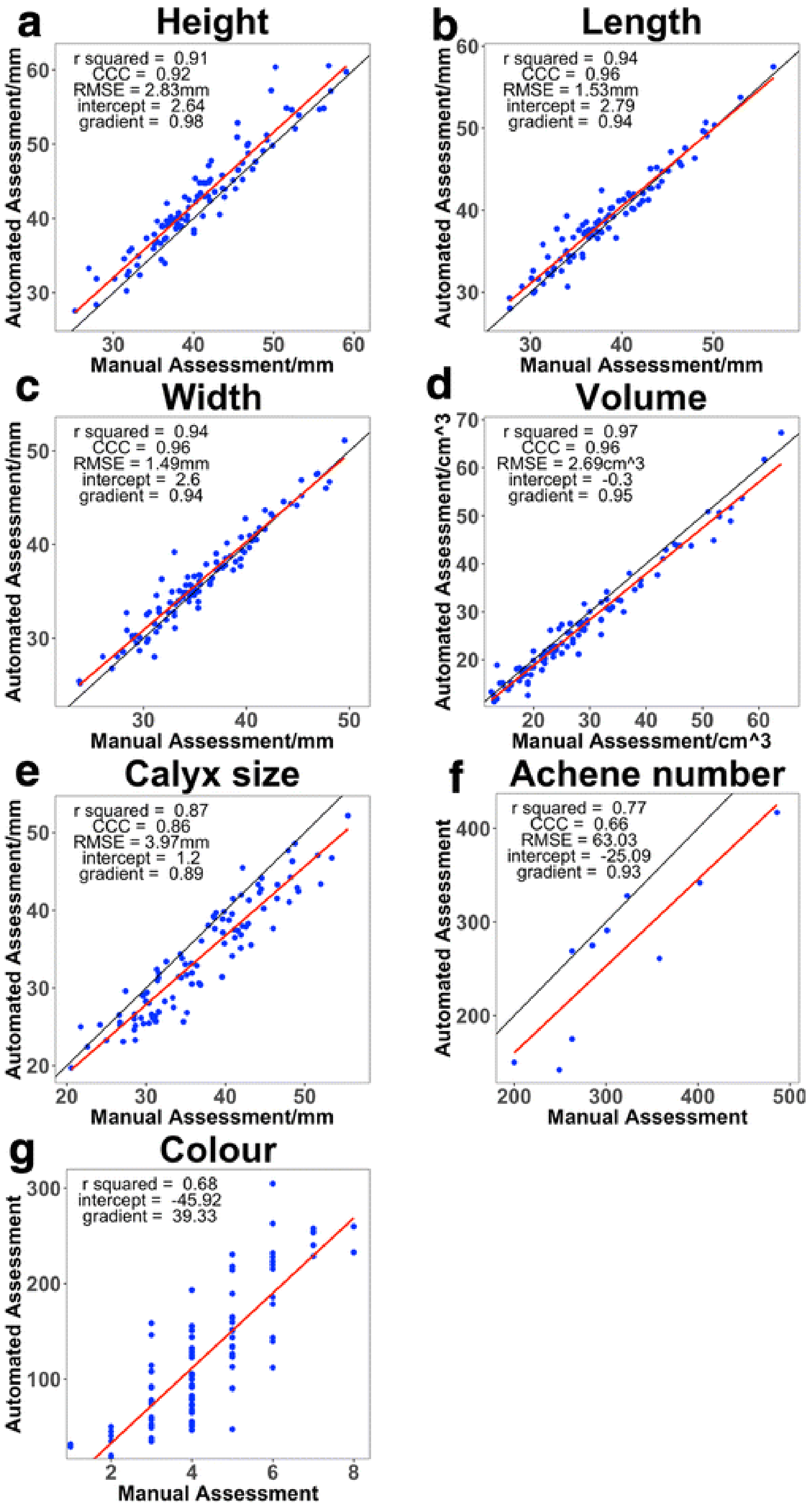


Figure 2.4: Regression analysis for 7 traits as measured by automated manual assessment. Sample size = 100 for all measurements, except achene number, where sample size = 10. Red lines are least squares linear regression curves and black lines are idealized regression curves (y = x) [He et al., 2017]

.

speeds and resolutions. Reducing resolution to 1000 x 1450 greatly increases the processing speed compared to the original images, but further investigation is needed to identify the minimum resolution to generate satisfactory point clouds. Use of multiple calibrated cameras to capture information from different viewpoints simultaneously could also be explored to further improve the quality and efficiency of 3D reconstruction, particularly from the nose of the fruit and the data capture speed.

As both fruit body and achenes have a range of colours, our current algorithm of arbitrary hue thresholding is unlikely to be reliable in identifying achenes from a range of cultivars. More sophisticated adaptive or texture based thresholding algorithms would likely improve the cluster identification.

It is believed that more traits could be derived from the gathered dataset. Firstly, algorithms exist that can calculate the surface area of a 3D mesh [Zhang and Chen, 2001], which together with reliable achene counts could be used to quantify achene density. Secondly, rotational symmetry could be quantified by considering the distribution of the Euclidean distance of points to the principal axis in 2D slices of the point cloud orthogonal to the principle axis. Thirdly, it may be possible to quantify the morphology of the fruit body at the neck of the fruit to derive information regarding the neck line.

2.7 Conclusion

In this study, a MVS based 3D reconstruction pipeline was developed and utilised to generate *in silico* model of strawberries. Automated 3D image point cloud analysis software was developed in house to derive berry achene counts, calyx size, colour, height, length, width and volume of the model. This study found good correlation between the automated and manual assessment techniques for dimension measurements and volume, suggesting that automated assessment is a promising technique to be utilised in place of manual assessment for these traits.

The focus of this study was the investigation of the use of 3D imaging to phenotype strawberries for commercial breeding. This system, with further improvement, can be quantitative, accurate, rapid and require little capital investment to be integrated into existing strawberry breeding programmes. This approach can also be further developed for strawberry quality control as its high precision is particularly suited for assessing differences within single cultivars, a situation frequently encountered in pack houses.

Chapter 3

Rational Design of Marker sets for Genomic Prediction in Strawberry

The research described in this chapter was mostly conducted before publication of the reference strawberry genome [Edger et al., 2019] and some of the methodology rests on assumptions made in the absence of the reference genome. A brief discussion of steps to improve the rational design process is provided, given the availability of the strawberry genome (subsection 3.4.2). Code and datasets for this chapter are available upon request.

3.1 Introduction

3.1.1 Cost Effective Marker Assisted Breeding for Strawberry

Markers are measurable differences between individuals that can be used to distinguish between them and fall under three groups: morphological, biochemical and genetic. With the advent of high-throughput genotyping, the number of genetic markers that can be measured has grown rapidly, making them a good target for prediction of plant performance [Collard et al., 2005]. Restriction fragment length polymorphisms (RFLPs) are polymorphisms due to differential presence of restriction enzyme sites within the genome and have been exploited for the construction of genetic maps [Chang et al., 1988]. Simple sequence repeats (SSRs) can be detected with targeted PCR amplifications, followed by gel electrophoresis [Kantartzi, 2013]. In recent times, single nucleotide polymorphisms (SNPs) have been identified as particularly useful for plant sciences as SNPs occur at high frequencies and can be simultaneously and cost effectively probed using microarray technology [Khlestkina and Salina, 2006, Verma et al., 2017].

Marker assisted selection (MAS) allows identification of agronomically important traits (such as disease resistance) at the seedling stage. If it is unacceptable for a plant to be of a particular genotype (such as susceptibility markers for diseases endemic in a target region), then the individuals can be culled, saving on costs of raising the individual to maturity. Based on data from the University of Minnesota breeding program, the break even cost for MAS in strawberry would be \$ 0.69 - \$ 1.08 if no additional cost was required to assess the trait phenotypically at maturity, increasing to \$ 9.70 - \$ 15.73 if the cost of assessment was \$ 20.00 [Luby and Shaw, 2000].

Expanded decision support analysis of a similar system modelling marker assisted seedling selection (MASS) as a three year process split into six sections was implemented, with variables based on historic data from Washington State University. For the four scenarios tested, strawberry had negative savings ranging from −70% to −145% due to the low cost of traditional selection in strawberry relative to the fixed costs of implementing MASS [Edge-Garza et al., 2015]. Interestingly, the model suggests the cull level of each stage of selection (in other crops) rather than the DNA testing cost chiefly governed the most cost effective stage to deploy MASS. Consistent with this analysis, another model shows, with data based on 2017 data from Washington State University, a 95% reduction of genotyping cost would be needed before it would be cost efficient for deployment in a strawberry breeding programme [Wannemuehler, 2018, Wannemuehler et al., 2020].

3.1.2 Cost Effective Genotyping Methods

In general, there are two methods to reduce costs of genotyping: reduction of the number of markers genotyped and development of techniques that use cheaper equipment/reagents or reduced labour per marker genotyped. Both approaches are likely needed to achieve cost effectiveness for deployment of MAS in strawberry. Approaches to reducing markers genotyped include imputation, where higher density marker profiles are imputed from low density profiles. For example, cost reduction could be achieved in wheat by marker imputation, though it was noted that GS benefited only marginally [He et al., 2015]. Imputation was dependant on rate of linkage disequilibrium (LD) decay and locations of regions of the genome that shows little evidence of recombination, termed haplotype blocks (haploblocks). Analysis of genomic variants across individuals in a population allows experimental identification of haploblocks with Practical Haplotype Graphs [Paten et al., 2017]. Selective genotyping of one (or few) markers per haploblock can be performed with imputation of the remaining markers based on the haplotype the marker is part of, and the assumption that there are few recombination events per generation. In sorghum, this approach showed similar accuracies in GP for seven traits, even when the number of markers was downsampled by 99% [Jensen et al., 2020]. Other approaches of marker selection experimentally identify markers with the largest effect in GP, then redeploy the reduced markers for cost effective GP. In rice and maize, post-hoc marker selection based on this approach could maintain selection accuracy [e Sousa et al., 2019].

Little effort has been made to select markers on the basis of genomic features such as exon locations, but this method of selection may overcome some of the shortcomings of haploblock selection and post-hoc marker selection. Imputation and post-hoc selection approaches are dependent on the discovery population, but as the population in a real breeding population changes over time, with introgression of new germplasm, for instance, the selected markers may not remain informative. It was reasoned that markers physically close to pertinent genomic features would be more useful for MAP as they were more likely to be in LD with genes controlling plant traits, regardless of the population. Moreover, it was reasoned that targeting conserved regions of the genome was less likely to result in off-target amplification or detection when using PCR based genotyping approaches. It is known that purifying selection is disproportionately active in the coding regions of genomes. Retrotransposons in *Arabidopsis* are disproportionately found in the heterochromatic pericentromeric regions of the genome [Pereira, 2004]. Analysis of the bZIP family of genes within strawberries shows that the ratio of synonymous to nonsynonymous mutations (K_s/K_a) is typically < 0.4, indicating strong purifying selection [Wang et al., 2015]. Reduction in accuracy of GP when deploying limited numbers of markers in other crops are generally small until marker numbers are less than $\sim$ 1000 - 5000 [Kriaridou et al., 2020, Zhang et al., 2019].

Genotyping can be classified into open and closed systems. In a closed system, such as a SNP array, the same markers are assessed across all individuals, while an open system genotypes individuals dependant on changeable reagents (e.g. primers) [Darrier et al., 2019]. Open systems are generally preferable as they allow for scaling of genotyping effort dependant on resources available to a breeding program, and replacement of markers that are not informative for particular populations. Currently, there are two SNP arrays available for strawberry; the IStraw90K Axiom SNP Array (90K SNP array) [Bassil et al., 2015] and the Axiom IStraw35 384HT array (35K SNP array) [Verma et al., 2017]. The latter represents the subset of markers from the former that were found to experimentally segregate in various strawberry populations in seven locations around the world and reduced costs from \$ 80 - \$ 105 to \$ 50 per sample [Verma et al., 2017]. Moreover, marker development costs for SNP arrays are not trivial and should be taken into consideration [Wannemuehler et al., 2020].

Genotyping-by-sequencing (GBS) utilises short read sequencing technologies to sequence the genome of a target plant, and in conjunction with bioinformatic identification of a set of known polymorphisms, allows resolution of markers. This approach allows genotyping of all markers across the genome and can inform marker discovery for future experiments [Deschamps et al., 2012]. GBS costs are associated with the depth of sequencing, and the implementation of genome simplification by digestion with restriction enzymes and fractionation prior to sequencing [Rowan et al., 2017]. Reduction of read depth allows for low costs, but when sequencing depth is < 1, not all markers will be genotyped, with limited ability to select the most useful markers. Restriction enzymes reduce genome complexity and thus costs, but targeting of specific loci is limited by the presence of restriction enzyme sites. Moreover, GBS typically requires an assembled genome of the target organism for bioinformatic identification of genotype [Kim et al., 2016]. In barley, the commercially available 50K SNP array costs approximately $\frac{1}{3}$ of the price of GBS per informative biallelic marker resolved [Darrier et al., 2019].

Repeat Amplification Sequencing (rAmpSeq) utilises a single primer expected to bind to multiple repetitive regions throughout the genome. The amplicons are treated as markers and their genotype is resolved by sequencing. One advantage of rAmpSeq is that it does not require a well assembled genome to be effective and is an order of magnitude cheaper than the currently available strawberry SNP arrays at $\sim$ \$ 5 per sample [Buckler et al., 2016]. However, rAmpSeq also does not allow for targeting of specific markers and repeat rich regions of the genome may be less informative on average as they contain a dearth of coding regions. rAmpSeq has been demonstrated in maize for GP for kernel zinc concentration, but prediction accuracies were significantly lower than genoypic information obtained by GBS [Guo et al., 2020].

Genotyping-in-Thousands by sequencing (GT-seq) is a potential cost-effective technique to genotype large populations of individuals for a small (50 - 500) panels of SNPs [Campbell et al., 2014]. It utilises two PCR steps to add Illumina sequencing primer sites, Illumina capture sites and unique barcodes to each individual, whilst also amplifying the targeted region. Then the DNA from each individual is normalised and pooled into a single test tube. Next, the sample is sequenced using Illumina short read technology to identify genotypes. Finally, a bioinformatics pipeline is employed to resolve the reads into individuals and ratios of genotypes taken to determine the genotype at each locus (Figure 3.1). GT-seq allows for custom sequencing of targeted regions (providing suitable primers can be found), which allows for rational selection of markers likely to be useful for genomic prediction (GP), with costs comparable to rAmpSeq at around $\sim$ \$ 5 per sample.

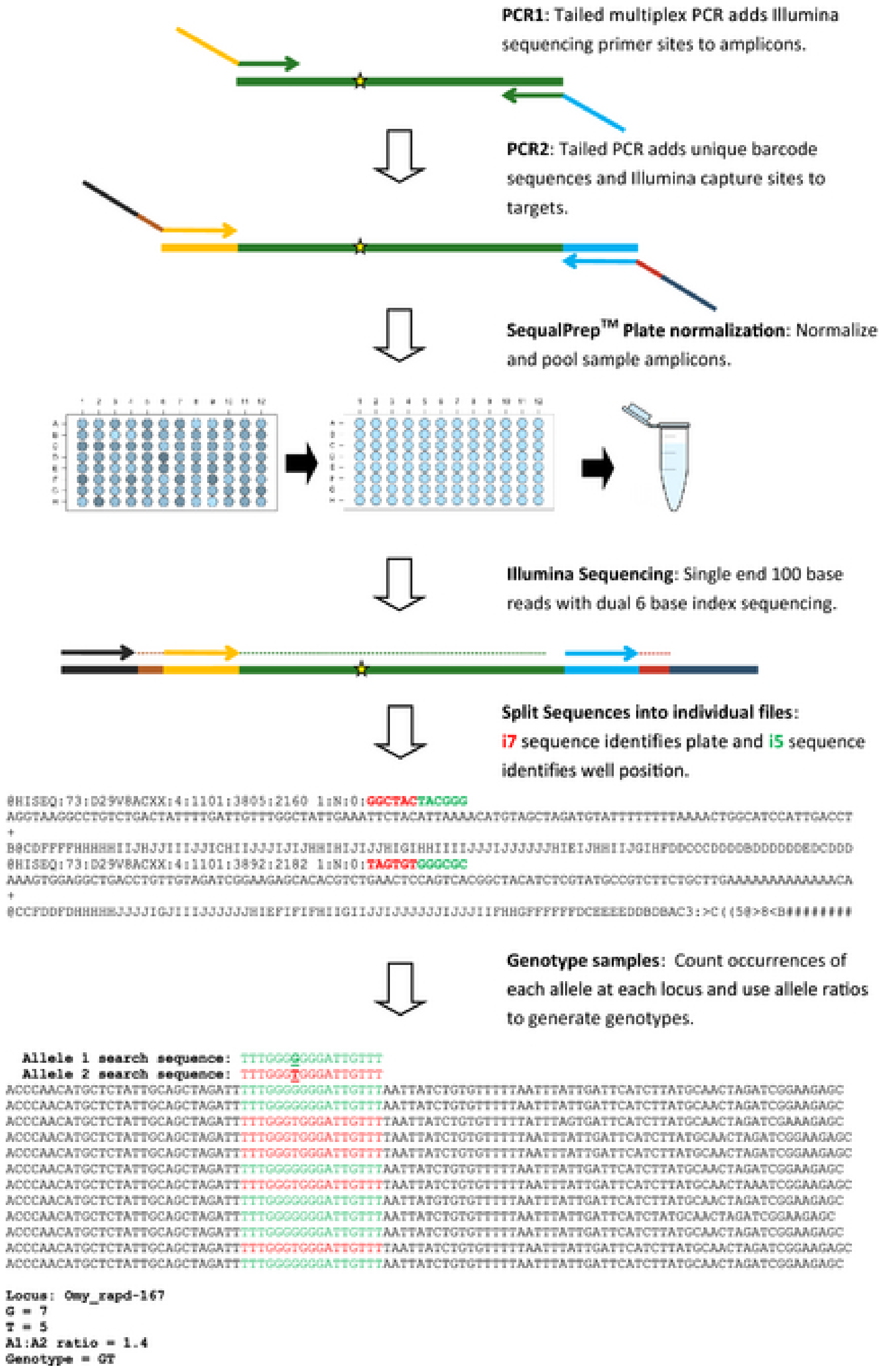
3.1.3 Genome Annotation

Genome annotation ascribes additional information about the location, function and other pertinent information regarding genes in a genome. Features of interest include genes, non-coding RNAs, promotor and enhancer regions and methylation sites [Salzberg, 2019]. In general, there are three sources of information: transcriptome studies, *ab initio* gene prediction and homology prediction based on known gene models [Keilwagen et al., 2018]. Transcriptome studies utilise RNA sequence data [Keilwagen et al., 2018] as experimental evidence of gene expression and have become more popular recently due to reduced cost and increased ease of such experimental approaches. *Ab initio* approaches to annotation include bioinformatic searches for features associated with genes such as splice sites, branch points, polypyrimidine tracts, start codons and stop codons [Wang et al., 2004]. Homology based alignments utilises conservation of gene structure in addition to the similarity of encoded amino acid sequences from well annotated species [Keilwagen et al., 2016] to identify intron positions and borders between exons and non-coding regions [Hoff et al., 2015].

A range of programs for gene prediction have been created incorporating available information. AUGUSTUS deploys *ab initio* approaches in addition to machine learning approaches, leveraging data from species that have already been well annotated [Stanke and Morgenstern, 2005]. MAKER2 is a pipeline that integrates *ab initio* predictors and RNA-seq data, which provides significant improvement of gene prediction [Holt and Yandell, 2011]. GeMoMa is a homology based gene prediction software that utilises amino acid and intron position conservation. Additionally, it can accept RNA-seq data to make predictions on protein coding regions [Keilwagen et al., 2016, Keilwagen et al., 2018]. However, when a draft genome is split into many small contiguous regions (contigs) whose order and scaffolding is unknown, genes may be broken among multiple contigs, making gene prediction unreliable [Salzberg, 2019].

3.1.4 Multiplexed Polymerase Chain Reaction

The development of Polymerase Chain Reaction (PCR) allowed for rapid *in vitro* amplification of almost any DNA sequence. A basic PCR reaction requires the sample, free nucleotide phosphates, primers for the region of interest and a polymerase in an aqueous solution [Saiki et al., 1988]. The exponential nature of the reaction allows for even a small amount of DNA to be amplified to appreciable quantities for analysis. A single cycle takes place in three steps, which are typically repeated 20-40 times, as needed, for a given final concentration of DNA:



PCR1: Tailed multiplex PCR adds Illumina sequencing primer sites to amplicons.

PCR2: Tailed PCR adds unique barcode sequences and Illumina capture sites to targets.

SequalPrep™ Plate normalization: Normalize and pool sample amplicons.

Illumina Sequencing: Single end 100 base reads with dual 6 base index sequencing.

Split Sequences into individual files: i7 sequence identifies plate and i5 sequence identifies well position.

```
@HISEQ:73:D29V8ACXX:4:1101:3805:2160 1:N:0:GGCTACTACGGG
AGGTAAGGCCTGTCTGACTATTTTGATTGTTTGGCTATTGAAATTCTACATTAAAACATGTAGCTAGATGTATTTTTTTTAAAACTGGCATCCATTGACCT
+
B@CDFFFFHHHHHIIJHJJIIIJJICHIIJJJIJIJHHIHIJIJJHIGIHHIIIIJJJIJJJJJJHIEIJHHIIJGIHFDDCCCDDDDBDDDDDDEDCDDD
@HISEQ:73:D29V8ACXX:4:1101:3892:2182 1:N:0:TAGTGTGGGCGC
AAAGTGGAGGCTGACCTGTTGTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAACA
+
@CCFDDFDHHHHHJJJJIGJIIIJJJJJJHIEFIFIFHIIGIIJJIJJJJJJIJJJIIFHHGFFFFFFDCEEEEDDBCBAC3:>C((50>8<B########
```

Genotype samples: Count occurrences of each allele at each locus and use allele ratios to generate genotypes.

```
  Allele 1 search sequence: TTTGGGGGGGATTGTTT
  Allele 2 search sequence: TTTGGGTGGGATTGTTT
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGGGGGATTGTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGGGGGATTGTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGTGGGATTGTTTTAATTATCTGTGTTTTTATTTAGTGATTCATCTTATGCAACTAGATCGAAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGTGGGATTGTTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGTGGGATTGTTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGGGGGATTGTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGTGGGATTGTTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAAATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGGGGGATTGTTTAATTATGTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGGGGGATTGTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGGGGGATTGTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGTGGGATTGTTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC
ACCCAACATGCTCTATTGCAGCTAGATTTTTGGGGGGGATTGTTTAATTATCTGTGTTTTTAATTTATTGATTCATCTTATGCAACTAGATCGGAAGAGC

Locus: Omy_rapd-167
G = 7
T = 5
A1:A2 ratio = 1.4
Genotype = GT
```

Figure 3.1: Overview of GT-seq [Campbell et al., 2014].

1. denaturation of the DNA sample (high temperature)
2. annealing of the polymerase (low temperature)
3. synthesis of the complementary strand (moderate temperature)

Multiplex PCR is when multiple primer pairs are deployed in a single PCR reaction, generating multiple amplicons. Multiplex marker PCR allows for inclusion of internal controls to indicate template quality and quantity, and efficiency of time in preparation and of reagents compared to uniplex PCR [Edwards and Gibbs, 1994]. In more recent times, multiplex PCR has been deployed to resolve the genotype of markers.

Multiplex primer design requires a range of primer characteristics to be considered including similar melting temperatures, avoidance of non-specific binding, generation of self-dimers, cross-dimers and hairpins [Wang et al., 2019]. MPprimer is a multiplex primer design software that utilises Primer3 and the primer specificity evaluation program MFEprimer to design and evaluate the candidate primers. Experimentally, it has been used to generate 79plexes in human [Shen et al., 2010] and 172plexes in human [Garaycoechea et al., 2018]. MPD is an alternative program that is capable of avoidance of placing primers over sites of known variation. Experimentally, it has multiplexed 313 samples simultaneously, identifying 224 variants [Wingo et al., 2017].

There were two main aims in this chapter. Firstly, a scalable, open genotyping system that cost $\sim$ \$ 5 per sample was to be generated for strawberry with a panel of markers that could be deployed for GP. Secondly, a rational selection approach would be implemented to select markers sets, by integrating factors that are likely to be informative for GP. It was envisaged that GT-seq would be deployed for cost-effective genotyping, so the design process was optimised for this.

3.2 Methods

3.2.1 Genome Annotation for Rational Marker Design

Sequence information surrounding markers was needed for amplicon design, but at the time this research was performed, no published assembly or genome annotation of the dessert strawberry genome was available. Work was underway to assemble the cultivar 'Redgauntlet' (Clavajo et. al. unpublished), and a draft genome was available (2017 version). The assembly was incomplete, fragmented into 130347 contiguous regions (contigs). Two approaches were used to identify genes in the draft assembly: an *ab initio* approach to identify genes across the genome and a targeted approach to identify genes believed to be involved in the everbearing trait. AUGUSTUS (V3.2.3) [Stanke and Morgenstern, 2005] was deployed for *ab initio* gene identification with default heuristic settings and the 'arabidopsis' gene model for training.

To assist rational design of marker sets for GP, an attempt was made to identify genes believed to be responsible for control of the everbearing trait. Seven genomic regions (Table 3.1) were identified as being associated with the everbearing trait in *F. vesca* on chromosome 4 (Hytonen, unpublished), based on the Fvb4_v4.0.a1 assembly [Li et al., 2019]. Identification of these regions in the dessert strawberry genome assembly would yield sequence information near these regions for amplicon development. The best

Gene name	Physical position
AP1	29660000 –29666200
CDF2	30453920 –30466065
FKF1	31702474 –31707462
FT2	30205976 –30211048
SEP4	29666150 –29674381
SPL6	29677951 –29686591
TCP14	30610251 –30619416

Table 3.1: Seven regions associated with the everbearing trait and their genome positions based on Fvb4_v4.0.a1 on chromosome 4 [Li et al., 2019].

available assembly of the dessert strawberry at the time, ‘Redgauntlet’ (2019 version; fragmented into 4122 contigs, Clavajo et. al. unpublished), was used as a reference genome for alignment.

Basic local alignment search tool (BLAST 2.6.0) [Donkor et al., 2014, Altschul et al., 1990] in nucleotide mode (BLASTN) was used to align the seven genomic regions and the ‘Redgauntlet’ genome, using default heuristic settings. A threshold of > 95 % match between the genomic regions was considered a true alignment. Additionally, Satsuma2 [Grabherr et al., 2010] was deployed for synteny identification between 29 - 33Mbp in the *F. vesca* genome and the ’Redgauntlet’ genome using default heuristic settings. After publication of the reference dessert strawberry genome [Edger et al., 2019], BLAST and Satsuma2 were deployed aligning the seven everbearing genomic regions with chromosome 4-3 of the reference genome (the ’vesca-like’ subgenome) to validate the quality of the ‘Redgauntlet’ assembly.

3.2.2 Rational Amplicon Design for Genotyping-in-Thousands Sequencing

The 90K SNP array is an experimental microarray for genotyping 95062 SNPs believed to be present in strawberry [Bassil et al., 2015]. Amplicons for rational design were targeted towards these markers as they represent a set of markers present in strawberry with high confidence. Moreover, the SNP array allows for experimental validation of the marker set and included known flanking regions, which could be useful for primer design. As markers were to be resolved by genotyping using the Illumina Miseq Reagent Kit V2 (2 x 250) (Illumina, Cambridge, UK), amplicons were limited in length to 500 bp. However, it is known that the paired end reads suffers from poorer accuracy at the tail [Schirmer et al., 2015], so a small overlap region sequenced would allow for increased depth to compensate.

Genetic maps generated from genotyping progeny from three biparental crosses (‘Redgauntlet’ × ‘Hapil’, ‘Emily’ × ‘Fenella’ and ‘Flamenco’ × ‘Chandler’) allowed experimental evidence of markers that segregated (Vickerstaff, unpublished). Additionally, data from genome-wide association studies (GWAS) identified 42 markers associated with resistance and susceptibility to verticilium wilt [Cockerton et al., 2018] and 12 markers associated with crown rot resistance [Nellist et al., 2019].

Given the poor quality of the *F. ananassa* ‘Redgauntlet’ genome assembly, it was decided that the published genome of *F. vesca* would be used as the sequence information for primer design. It was assumed that the genome of the woodland strawberry [Shulaev et al., 2011] represented each of the homeologous chromosomes in the dessert strawberry. Physical positions of probe targets in the array had been related to physical positions in the *F. vesca* (V1.1) genome assembly (Vickerstaff, unpublished).

A custom script was developed using Python (V.2.7.13; random seed = 1000 for all subsequent analyses using the 'random' module) to count the number of markers in all 450bp windows containing at least one marker in *F. vesca.* To evaluate potential amplicons and integrate information from the biparental cross and disease resistance markers, a heuristic scoring algorithm was applied to each amplicon. Each window was defined as a potential amplicon and was given a score equal to the number of markers that it contained. Every amplicon that contained markers that experimentally segregated in any of the biparental crosses was given a score of 1 multiplied by the number of subgenomes markers were identified in. Every marker identified in either disease resistance GWAS was assigned a score of $ln(0.001/p)$ where p is the p-value of the marker being associated with the disease and / is the floor divisor operator. If an amplicon overlapped with a coding region or exon based on the genome annotation of the woodland strawberry [Darwish et al., 2013] the amplicon was assigned a score of 1. The heuristic score of an amplicon was the sum of all these features.

To generate informative primer sets for GP that could be multiplexed, MPprimer [Shen et al., 2010] was deployed for multiplex design. Illumina i5 and i7 adaptors were added to the sequences in accordance to the manufacturer's recommendations. A Gillespie algorithm [Gillespie, 1977] was implemented in Python to select amplicon pairs with probability mass functions for each amplicon proportional to its heuristic score. These amplicons were input to MPprimer, using defined options (Table 3.2). Upon identifying a successful pair of amplicons, additional amplicons, as selected by the Gillespie algorithm, were added to the multiplex in an 'evolutionary' algorithm to build up multiplexes. Upon failure of multiplexing, the multiplex was discarded and a new twoplex was generated. In this way, a preliminary sixplex was generated targeting loci across multiple chromosomes to experimentally test the suitability of this rational design process. Design of the multiplexes were relatively insensitive to changes in most of the conditions, except primer product size.

3.2.3 Repeat Amplification Sequencing Primer Design

To explore the viability of rAmpSeq as an alternative to GT-seq for cost effective marker genotyping, a custom pipeline was implemented in Python to identify potential primers for rAmpSeq. Jellyfish [Marçais and Kingsford, 2011] was deployed to count every 20mer in the unmasked *F. ananassa* reference genome [Edger et al., 2019]. Each 20mer was filtered to be present in 100 - 10000 copies, contain 35% - 65% GC content, and include a G or a C within three bases of the end as a 'GC-clamp' to increase specificity of primer binding. For each 20mer in each contig, distances were calculated between the next instance of the same 20mer and where distances were between 125 - 200 base pairs, an amplicon was defined. The total number of expected amplicons for each 20mer was summed and 20mers were ranked by number of expected amplicons.

MPprimer Design Parameter	**Value**
PRIMER OPT SIZE	22
PRIMER MAX SIZE	35
PRIMER MIN SIZE	16
PRIMER OPT TM	60 0
PRIMER MAX TM	70 0
PRIMER MIN TM	50 0
PRIMER OPT GC PERCENT	50 0
PRIMER MAX GC	65 0
PRIMER MIN GC	35 0
PRIMER MAX END STABILITY	20 0
PRIMER SELF ANY	10 0
PRIMER SELF END	3 0
PRIMER NUM NS ACCEPTED	100
PRIMER PRODUCT SIZE RANGE	$[150, 600]$
PRIMER SALT CONC	50 0
PRIMER DIVALENT CONC	0
PRIMER DNTP CONC	0
PRIMER DNA CONC	50 0
PRIMER TM SANTALUCIA	1
PRIMER SALT CORRECTIONS	1
PRIMER FILE FLAG	0
PRIMER PICK INTERNAL OLIGO	0
PRIMER EXPLAIN FLAG	1

Table 3.2: Variable values inputted into MPprimer [Shen et al., 2010] to generate amplicon sets for GT-seq. Primer parameters tend to be more relaxed relative to uniplexes to increase probability of finding higher plex sets.

3.2.4 Read Depth Simulation

In order to estimate the total number of reads necessary to have sufficient power to resolve genotypes, an stochastic *in silico* simulation of the sequencing stage of GT-seq was implemented to model the number of reads required per genotype. A mixed population of genotypes dependant on the number of individuals to be genotyped and differences in concentrations of amplicons was simulated and amplicons were selected proportional to its concentration until a desired coverage of markers was genotyped. In this way, a minimum number of reads could be estimated.

3.2.5 Genetic Material and Primers

Young leaves from *F. ananassa* cultivars 'Redgauntlet', 'Hapil' and 'Emily' were harvested and stored in darkness overnight at 4 °C prior to DNA extraction. Extraction was performed with the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The final concentrations of DNA obtained was 507ngμL^{-1}, 621ngμL^{-1} and 262ngμL^{-1} respectively.

Primers with standard desalting purification were designed (Table 3.3) and synthesised by a commercial company (Integrated DNA Technologies, IDT, IA, USA). The final mix for thermocyling (total volume 50 µL) included forward and reverse primers at 1μM each, genomic template at 1ngμL^{-1} and Kapa Hifi HotStart ReadyMix (2×) Mix (Kapa Biosystems, Wilmington, MA, USA) according to the manufacturer's recommendations. Dilution in all cases was with DNAse free water.

3.2.6 Polymerase Chain Reaction

For uniplex and multiplex reactions, the enzyme was heat activated at 95°C for 3 minutes. 30 theromocyles were them implemented: 95°C for 30 seconds; annealing temperature (varied, subsection 3.3.4) for 30 seconds; 72°C for 30 seconds. The reaction was held at 72°C for 5 minutes to ensure extension completed after the thermocycles and then chilled to 4°C until analysis. For Touchdown PCR, the same conditions as the previously described PCR was conducted, except the annealing temperature was decreased by 0.2°C per cycle Figure 3.6.

3.2.7 Electrophoresis and Imaging

A 2% agarose gel was prepared using standard methods, spiked with Gel Red according to the manufacturer's recommendations. Each sample was mixed with DNA loading buffer according to the manufacturer's instructions and loaded into the wells with a DNA ladder (GeneRuler 1 kb Plus DNA Ladder, ThermoFisher, Waltham, MA, USA) at the ends of the gel. Electrophoresis was conducted by immersion of the loaded gel in Tris (40 mM), acetic acid (20 mM) and EDTA (1 mM) (TAE) buffer, maintaining potential difference across the gel at 200mV until the loading buffer reached the end of the gel. Pictures of the gel was captured using an imager (Gel Doc XR+, BioRad, Hercules, CA, USA) with darkness adjusted to maximise contrast.

ID	**LG**	**Start Position**	**Target Length**	**Temperature/°C**	**Sequence**
JH02_{F}	2	22622454	329	59.8	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGGGCATGTTGGAGCAGTGGC
JH02_{R}	2	22622454	329	59.9	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCGTGCAGCAGTTAAGCCAGCA
JH05_{F}	3	9523186	163	60.0	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGAGCCCCAGCCTGAGAAGAG
JH05_{R}	3	9523186	163	60.2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGGGCCAAAAGGGTCTGAGGGAA
JH01_{F}	4	7734927	260	58.3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCGGAACCGGTGGTAGCGAAAT
JH01_{R}	4	7734927	260	57.5	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGCAGACCTGTGTTGCAGCGA
JH04_{F}	5	20595999	426	59.0	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGCCCCTTCAACAAAGGCTCC
JH04_{R}	5	20595999	426	60.2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAGGCTCTCCGCTCCAGCAAGT
JH03_{F}	7	20416880	207	59.7	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTTGAAGACCGTAGCCCTCGT
JH03_{R}	7	20416880	207	59.7	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTTTCGCCCAAGCCCTCTTAGC
JH06_{F}	7	21334717	514	60.1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGAGCGCAGCAGCAGGAATGA
JH06_{R}	7	21334717	514	60.0	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCCAAGCCGAAGGCATCAAGGT

Table 3.3: Rationally designed primers for a 6plex PCR reaction. Start Position indicates expected start of amplicon with respect to the *F. vesca* genome [Shulaev et al., 2011] in given linkage groups (LG column) with expected amplicon sizes. Temperature indicates predicted annealing temperature based on MPprimer [Shen et al., 2010]

3.3 Results and Discussion

3.3.1 Genome Annotation for Rational Marker Selection

221967 genes were identified in the 'Redgauntlet' assembly through *ab initio* gene prediction using AUGUSTUS. The number of genes in diploid *F. vesca* is around 30000 (32831 in V1.1 [Darwish et al., 2013]; 33496 in TowU_Fve [Darwish et al., 2015], 34007 in v4.0.a2 [Li et al., 2019]) and expected to be around a quarter of the number in the octoploid *F. ananassa*. The significantly greater number of genes identified through *ab initio* prediction indicated there were potentially replicated contigs in the assembly. Alternatively, the heuristic or gene model settings for the *ab initio* prediction were unsuitable for *F. ananassa*. The most recent annotation of *F. ananassa*, V1.0.a2, indicates 108,447 genes with 97.5 % complete Benchmarking Universal Single-Copy Orthologs (BUSCOs) [Liu et al., 2021].

Alignment of the seven genomic regions to 'Redgauntlet' identified 24 contigs containing one of these regions (Figure 3.2). CDF2 was identified in 14 contigs, FKF1 was identified in 6 contigs, while SPL4, SEP4 and AP1 were not identified in 'Redgauntlet'. This could be due to duplication of contigs in the 'Redgauntlet' assembly or significant rearrangements in the dessert strawberry or the heuristics for BLAST were not optimised. Despite some of the everbearing regions being physically close to each other, no regions were identified on the same contigs in the 'Redgauntlet' assembly suggesting poor assembly or significant genomic rearrangements.

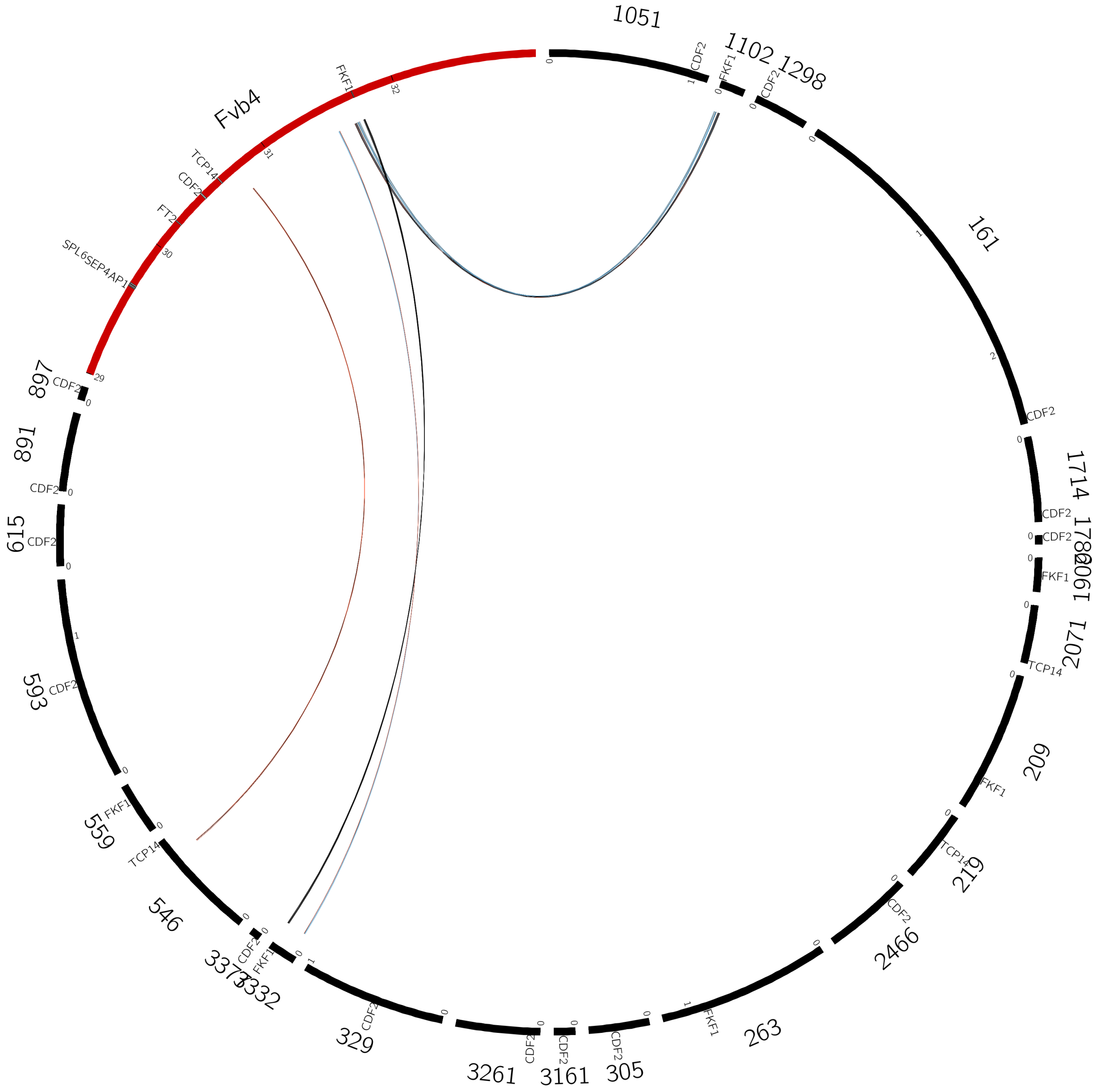
Synteny identification was poor between the 'Redgauntlet' contigs and the *F. vesca* genomic region (Figure 3.2). This may be due to poor assembly or significant sequence divergence since the last common ancestor to *F. vesca* and *F. ananassa*. Where there was synteny, however, the syntenic regions were approximately equidistant from identified genomic regions in both *F. vesca* and F. *ananassa*.

Comparison of the synteny plots between *F. vesca* to 'Redgauntlet' and *F. vesca* to 'Camarosa', the dessert strawberry reference genome [Edger et al., 2019], shows better alignment between *F. vesca* and 'Camarosa'. The seven genomic regions appear in the reference genome in approximately the same order and distances apart, consistent with the hypothesis that chromosome 4-3 descended recently from a common ancestor to *F. vesca*. Multiple instances of the same genomic region at approximately the same locus identified in the reference genome is likely an artefact of the BLAST algorithm. Synteny between the two chromosomes was strong with evidence of a small translocation from 32 –32.2 Mbp with respect to the *F. vesca* genome.

The two approaches to genome annotation, taken together, indicated the 'Redgauntlet' genome assembly was unreliable. It was decided that the 'Regauntlet' assembly would not be used in rational design of the marker set; instead, the annotated assembly of *F. vesca* (V.1.1) was used, with the assumption that all subgenomes of *F. ananassa* were identical to it.

3.3.2 Amplicon Design and Read Depth Simulation

94980 450mers containing at least one marker were identified through the amplicon design approach. The read depth simulation algorithm was implemented, assuming a 95%


Fvb4
FKF1
TCP14
CDF2
FT2
SPL6SEP4AP1
1051
CDF2
FKF1
1102
1298
CDF2
161
CDF2
1714
CDF2
1780
CDF2
FKF1
2071
TCP14
209
FKF1
219
TCP14
CDF2
2466
263
FKF1
305
CDF2
3161
CDF2
3261
CDF2
329
CDF2
3332
3373
FKF1
CDF2
546
TCP14
559
FKF1
593
CDF2
615
CDF2
891
CDF2
897
CDF2

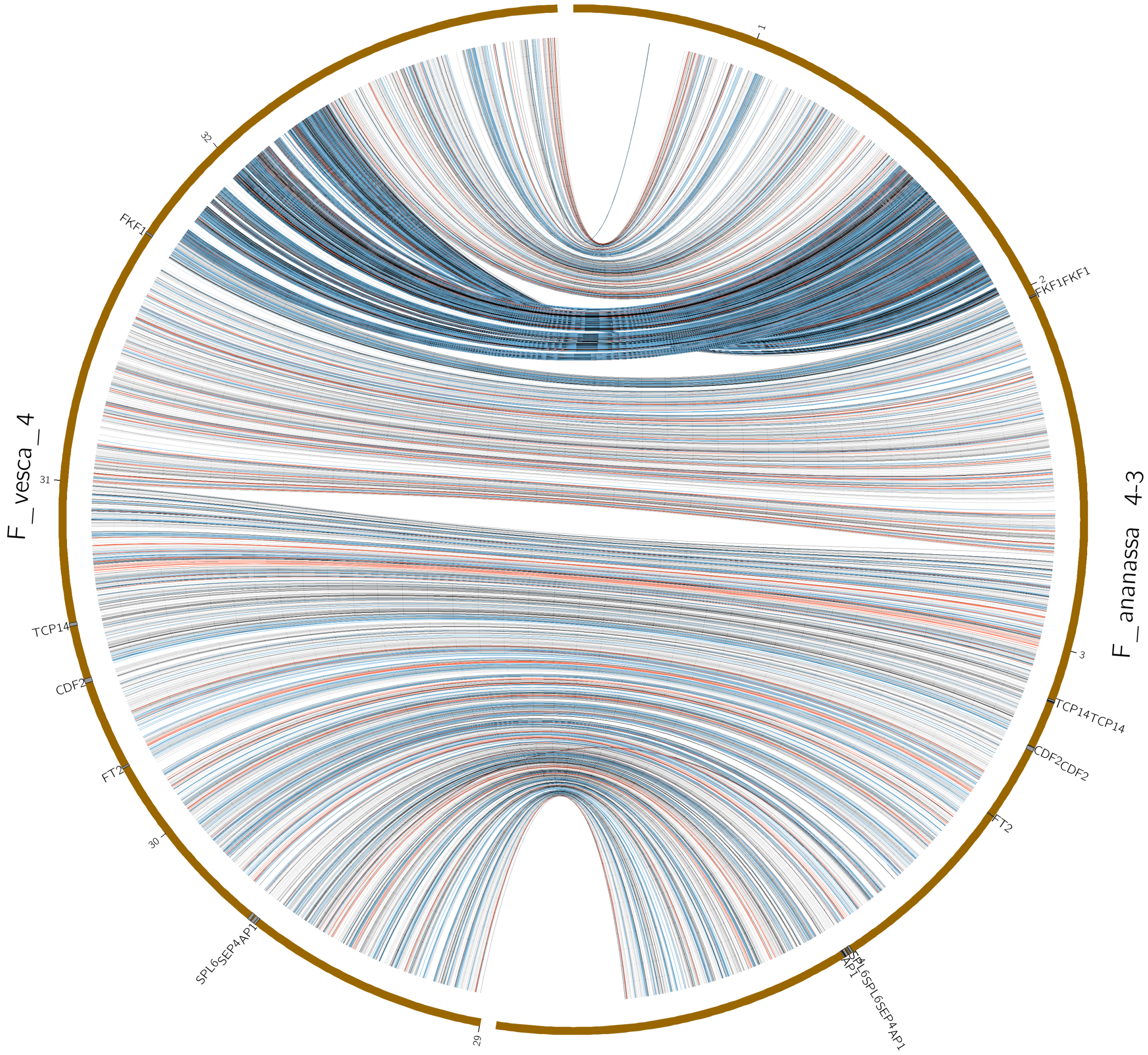


Figure 3.2: Upper image: synteny between *F. vesca* chromosome 4 (red) and assembly of *F. ananassa* cv. 'Redgauntlet' (black). Lower image: synteny between *F. vesca* and *F. ananassa* cv. 'Camarosa' chromosome 4-3 [Edger et al., 2019]. Ticks indicates base position in Mbp. External labels indicates contig name. Linking lines between contigs indicate synentic regions with colour indicating presence of a feature in the genome annotation of *F. vesca*. Blue line: coding sequence; red line: exon; black line: neither coding region nor exon. Plots simplified to include only contigs that include alignment of the seven genes associated with everbearing.

chance of achieving 10 fold coverage in 90% of samples in 192 amplicons in 2068 octoploids, 57123468 reads would be needed. A Illumina Miseq Reagent Kit V2 (2 x 250) generates ∼ 15 million reads per cell, indicating ∼ 4 cells would be needed to achieve this level of coverage (`https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/miseq-reagent-kit-v2.html`). The price of a cell is £945 (as of March 2021), indicating a sequencing cost of ∼ £2 per individual genotyped. Simulation took approximately 10 hours (MacBook Pro, Intel core i7 × 4 @ 3.1GHz, macOS High Sierra, Apple Inc, CA, USA).

3.3.3 Repeat Amplification Sequencing Primer Design

Primers were successfully designed using the described algorithm. Due to material reasons, experimental validation of rAmpSeq was not implemented.

3.3.4 Polymerase Chain Reaction

Uniplexes of each of the primer pairs generates the expected sized fragments with higher temperatures having greater specificity (Figure 3.3). As 64°C annealing temperature had the highest specificity, that temperature was used in subsequent experiments.

The 6plexes do not amplify evenly, with one band (JH06) dominating the reaction. This occurred regardless of the template DNA (figure 3.4). In order to further investigate the interactions of the multiplexes, 5plexes were generated, with each primer pair being sequentially removed (figure 3.5). JH06 appears to be the dominant band, but the appearance of additional bands when some primer pairs are removed suggests that the primers are interacting in unexpected and complex ways.

To investigate if the temperature used influenced the formation of side products in the multiplex reaction, a series of touchdown PCR reactions was conducted with a range of temperatures (Figure 3.6). Under all attempted conditions, side products appeared to form with a JH06 as a dominant band.

Taken together, it appears that the described strategy would be unsuitable to generate a high enough multiplex for GT-seq in strawberry. GT-seq for strawberry was not implemented as the multiplex designed demonstrated complex interactions that interfered with amplification of the amplicon set. As the designed set represented a small proportion of the number of markers that was to be multiplexed, it was expected that GT-seq with the full set of markers would generate many off-target interactions.

3.4 Conclusions

3.4.1 Multiplex Design for Genotyping-in-Thousands Amplicons

Experimental demonstration of GT-seq has been applied to steelhead trout [Campbell et al., 2014] and western rattlesnake [Schmidt et al., 2019], with no bioinformatics approaches used to predict multiplex primer interactions. In the rattlesnake experiment, three of 16 replicated GT-seq attempts failed at the PCR amplification stage, indication some potential unreliability in GT-seq. In the trout population 192 markers were selected

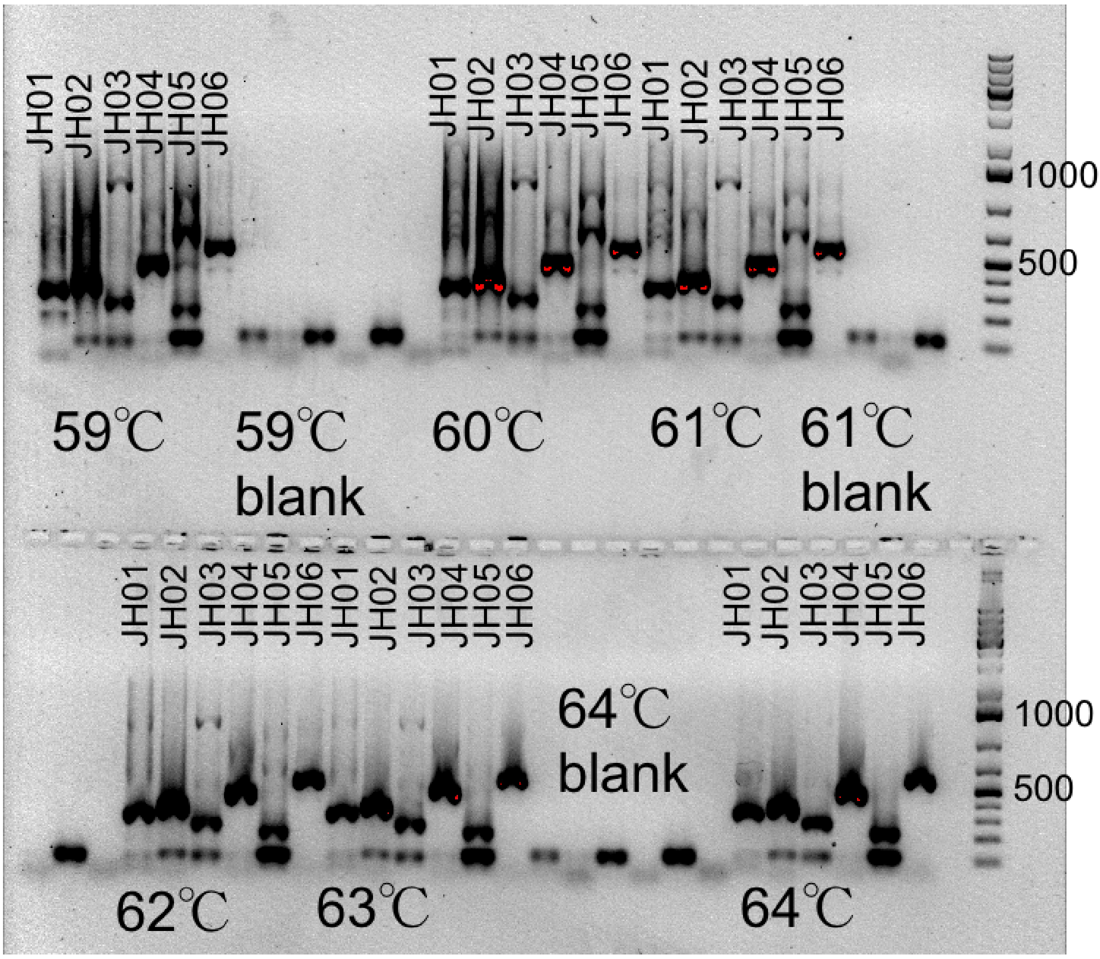


Figure 3.3: Gel electrophoresis of the products of pairs of primers designed through a rational process for GT-seq. Different annealing temperatures were used in blocks (indicated) with negative controls (no template DNA). Uniplexes amplify expected targets with highest specificity at 64 °C.

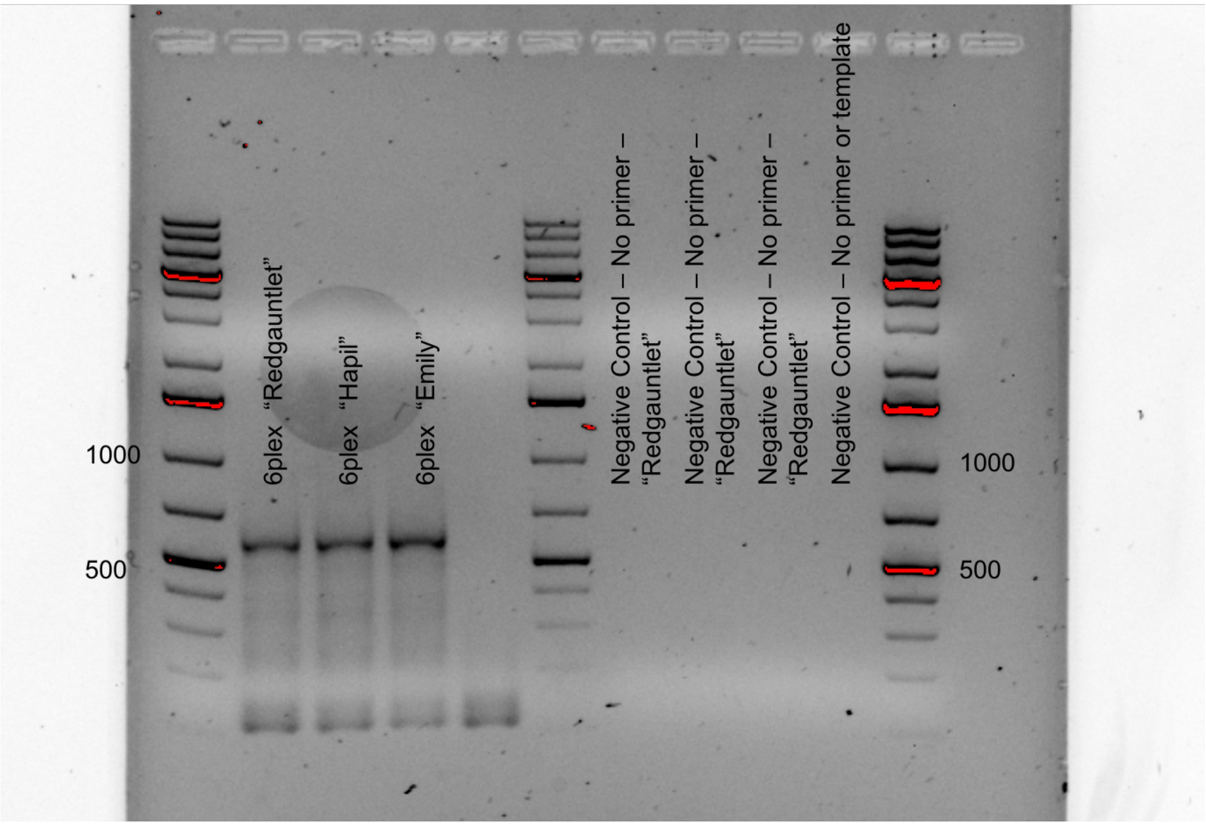

1000
500
6plex "Redgauntlet"
6plex "Hapil"
6plex "Emily"
Negative Control – No primer – "Redgauntlet"
Negative Control – No primer – "Redgauntlet"
Negative Control – No primer – "Redgauntlet"
Negative Control – No primer or template
1000
500

Figure 3.4: A 6plex PCR reaction was conducted with three cultivars of strawberry (annealing temperature at 64°C). Uneven amplification with fusion products appear independent of the strawberry cultivar amplified. Negative controls are to the right of the image. 10× dilution was performed prior to gel electrophoresis.

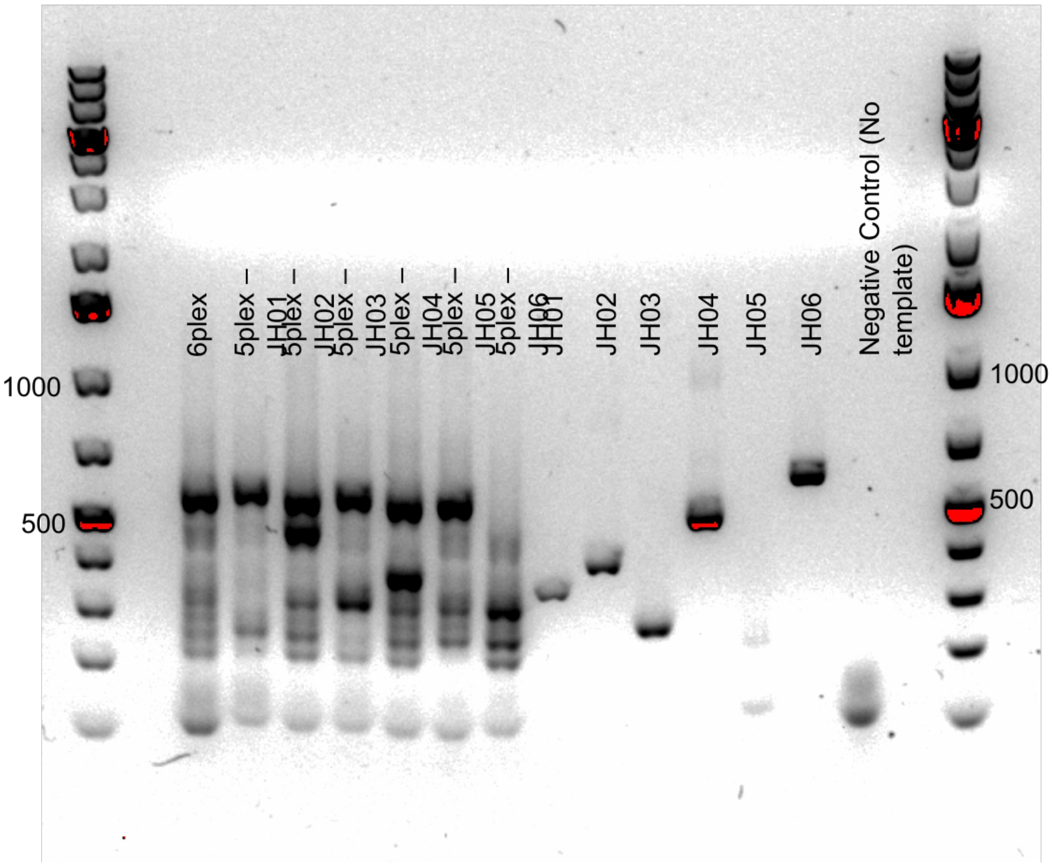


Figure 3.5: 5plexes were generated (labelled with single primer pair missing). Uniplex controls suggest that the multiplex banding pattern is a result of multiplex interactions. .

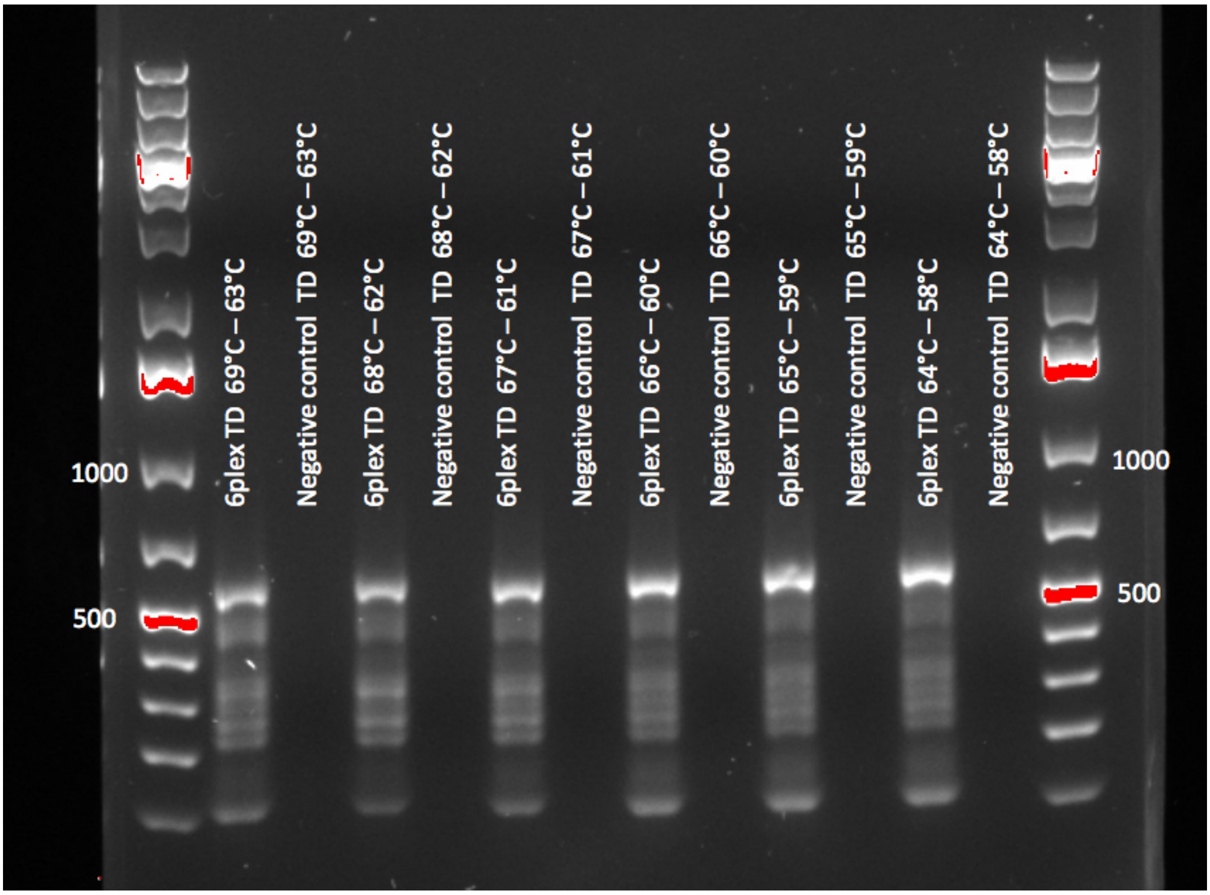

1000
500
6plex TD 69°C – 63°C
Negative control TD 69°C – 63°C
6plex TD 68°C – 62°C
Negative control TD 68°C – 62°C
6plex TD 67°C – 61°C
Negative control TD 67°C – 61°C
6plex TD 66°C – 60°C
Negative control TD 66°C – 60°C
6plex TD 65°C – 59°C
Negative control TD 65°C – 59°C
6plex TD 64°C – 58°C
Negative control TD 64°C – 58°C
1000
500

Figure 3.6: To investigate if complex primer interactions were due to some reactions dominating at certain temperatures, touchdown PCR was performed. under the conditions studied, uneven amplification occurred and primer interactions were observed.

based on association with divergent selection. Of these, 8 primer pairs were identified affecting primer amplification reads experimentally, generating artefact reads and were redesigned. This suggests that empirical design of primer sets for GT-seq may be more reliable than the bioinformatic design approach taken here. However, when single primer pairs from the strawberry multiplex were removed, artefact amplicons were still generated, suggesting that no single primer pair in the designed marker set was responsible for the artefact amplicons.

Multiplexing may also have failed due to inappropriate inputs for MPprimer. It was noted that the optimal annealing temperatures identified for all primers experimentally were higher than the temperatures predicted by MPprimer, indicating potential unreliability of MPprimer for multiplex design. Differences in the genomes of the organisms may also contribute to failure of multiplexing. As the diploid genome was used to generate primers and make predictions in the octoploid genome, interactions between different sequences across homeologous chromosomes may allow off target amplification. The version of the diploid genome deployed has undergone significant changes with the addition of more RNA-seq data [Li et al., 2019], but the quality of the diploid assembly did not appear to affect the accuracy of the uniplexes generated.

Alternative approaches to design of multiplexing primers can be explored. Primer-Pooler automatically designs multiple primersets when primer interactions are expected, allowing for flexibility in moving sets of interacting primers for multiplex in another tube rather than discarding them from the multiplex [Brown et al., 2017]. Additionally, experimental validation of the rAmpSeq primers may be an alternative method of cost effectively genotyping strawberry, though it would not be possible to target the rationally designed markers.

3.4.2 Refinement of the Rational Design Process

The rational design process described attempts to maximise the usefulness of the genetic markers genotyped for GP using GT-seq by targeting markers known to segregate, is associated with known traits of interest to breeders, overlap with predicted genes and amplifies across homeologous chromosome. Given the availability of a reference strawberry genome [Edger et al., 2019], several of these approaches can be refined and other factors that may affect GP can be incorporated into the design process.

Gene models have been predicted based on the reference genome [Edger et al., 2019], which could be used to bias the rational design algorithm towards regions known to encode for genes. As each subgenome is resolved in the reference genome, bioinformatic approaches could be deployed to identify if markers are expected to be amplified across subgenomes. However, transcriptional control is known to be important in controlling traits of interest to the breeder, for example in flavonoids and phenlypropanoids biosynthesis [Li et al., 2020]. Further research could estimate the magnitude of relative effects of transcriptional and genetic control of traits.

Additional factors can be included in the rational design process to increase likely usefulness for GP. It is known that the *F. vesca*-like subgenome is dominant in the dessert strawberry, retaining 20% more protein coding genes than other subgenomes, followed by the *innumae*-like, and the *F. virdis*-like and *F. nipponica*-like subgenomes [Edger et al., 2019]. This dominance suggests the stochastic rational design algorithm could be biased in the observed proportions to increase coverage in the dominant subgenome.

It seems possible to utilise a Practical Haplotype Graph approach to rationally select from haploblocks that show variation within the strawberry population. A population of strawberries representative of diversity of germplasm material in breeding programs would serve as the discovery population. This population would also allow estimation of allele frequency, and a bias for rare alleles may be incorporated into the design process. Selection for rare alleles increases the average genetic gain in a population as bringing it to fixation involves more individuals gaining beneficial alleles, but incorporation of rare alleles increase the proportion of populations for which that allele is not present [Shi and Lai, 2015].

Although the implemented method described is outdated, given the publication of the strawberry reference genome, the rational design approach remains a valid approach to design marker sets to maximise usefulness in GP. It is envisaged that an iterative approach using rationally selected markers in GP to maximise useful information combined with experimentally effective markers can be used to improve heuristic scores for the design process.

Chapter 4

Marker Assisted Prediction Methods in a Biparental Mapping Population

4.1 Introduction

4.1.1 Selection

One key task of plant breeders is to predict the performance of individuals in the future given current data, and select optimal genotype(s) for the next step of the breeding cycle. Traditionally, selection has been done based on assessment of the phenotype of individuals from the crop population, but in modern times, breeding programmes have been established to generate and select from expanded variation by controlled mating [Breseghello and Coelho, 2013]. With the advent of modern gene theory and understanding that DNA is the heritable material in life, it is possible to monitor the inheritance of variation by analysis of the genome, and select for specific recombinants.

When genetic marker data are available for a population, they can be used for a range of purposes, including verification of plant identity and planning crosses by better understanding of lineages [Whitaker et al., 2020]. Marker data can also be used to predict performance and select seedlings for progression onto subsequent stages of the breeding cycle, termed marker assisted prediction (MAP) and selection (MAS). MAP assumes that there exist molecular markers in close LD with QTLs controlling agronomically important traits and predicts the effects of the markers on the trait. MAS selects indirectly on traits of individuals from a breeding population by selection upon the markers [Xu and Crouch, 2008].

MAS typically operates in two steps, the training and breeding phases. In the training phase, significant relationships between phenotypes and genotypes are found using statistical approaches. In the breeding phase, genotype data are obtained in a breeding population, before favourable individuals are selected based on their genotypes [Nakaya and Isobe, 2012]. MAS can be divided into two variations; traditional MAS (tMAS) which treats each marker or small region of the genome separately and computes their effect on the trait, and genomic prediction (GP) which simultaneously considers all markers in an individual to generate genetic estimated breeding values (GEBV) [Heffner et al., 2009, Meuwissen et al., 2001, Crossa et al., 2017].

MAP is expected to benefit plant breeding by a range of mechanisms. MAP may allow for (i) more effective identification and quantification of genetic variation in available germplasm resources, (ii) introgression of QTLs associated with agronomically important traits, (iii) manipulating genetic variation in plant breeding population (iv) pyramiding multiple traits into single individuals [Xu and Crouch, 2008]. Additionally, GP may allow (v) control of traits that are difficult to manage through conventional phenotypic selection - because they are expensive or time-consuming to measure, (vi) traits whose selection depends on specific environments or developmental stages that influence the expression of the target phenotype and (vii) elimination of some field experiments and better planning of crosses by providing information on relatedness [Gezan et al., 2017].

tMAS has been most successfully deployed in strawberries to associate markers with disease resistance, including resistance to red stele rot [Haymes et al., 1997, Mangandi et al., 2017, Noh et al., 2018], anthracnose fruit rot [Denoyes-Rothan, 1997], angular leaf spot [Oh et al., 2020] and colletotrichum crown rot [Anciro et al., 2018]. Markers associated with QTLs controlling the everbearing trait [Negi et al., 2020] and fruit quality parameters including mesifurane content [Zorrilla-Fontanesi et al., 2012], γ-decalactone content [Chambers et al., 2014], a range of flavonoids [Labadie et al., 2020] and phenylpropanoids [Pott et al., 2020] have also been identified [Whitaker, 2011, Whitaker et al., 2020].

GP uses markers covering the whole genome so that all genetic variance can be explained by the markers. It is assumed that the markers are dense enough that they are in LD with QTLs controlling traits, and therefore the markers can be used to select for the favourable individuals [Goddard, 2008]. GP is expected to predict traits with greater accuracy than tMAS approaches as firstly, the bi-parental mapping populations used in most QTL studies do not readily translate into breeding applications; and secondly the identification of major effect loci in tMAS is ineffective for traits controlled by multiple traits of small effect [Meuwissen et al., 2001, Heffner et al., 2009]. Additionally, tMAS introduces bias in effect size estimation by assigning arbitrary thresholds for significant markers, thus shrinking markers with smaller effect sizes to zero [Nakaya and Isobe, 2012] and overestimating the effect size of the remaining markers [Xu, 2003]. Simulation studies indicate the estimates of phenotypic variances associated with correctly identified QTLs were particularly overestimated when the mapping population was small [Beavis, 1994].

Originally developed for cattle breeding, GP has now been deployed in a wide range of plants ranging from the maritime pine [Bartholomé et al., 2016] to flax [He et al., 2019]. Success has been demonstrated in overcoming some shortcomings in tMAS, particularly in staple calorie crops [Voss-Fels et al., 2019]. For example, in wheat, GP allowed prediction of yield, thousand grain weight and days to heading, despite these being traits controlled by many small effect QTLs [Poland et al., 2012]. Success in GP is typically represented by the Pearson's correlation coefficient between prediction and validation datasets, termed the selection accuracy [González-Recio et al., 2014]. In soybean, GP had a selection accuracy of 0.80 - 0.85 compared to 0.64 - 0.74 using tMAS for seed weight [Zhang et al., 2016]. GP was deployed in rice to select hybrids, resulting in a 16 % increase in yield compared with the average of all potential hybrids [Xu et al., 2014].

GP has been deployed in strawberries for five complex fruit quality traits over two years. Six models of prediction were deployed including Genomic Best Linear Unbiased Prediction (GBLUP), and Bayesian methods. BayesB was found have the highest prediction accuracy, while GBLUP performed worst when predicting between years, but

GBLUP was found to be more accurate when cross validating. Prediction accuracies were found to be consistent with accuracies of GS deployed in other crops [Gezan et al., 2017]. Deployment of GP assessing verticillium wilt resistance found no large effect segregating markers, with the three models tested performing similarly. Utilisation of both years' data generated higher selection accuracies than use of either year's data alone. Selection accuracy decreased when predicting verticillium wilt resistance between years over within the same year [Pincot et al., 2020].

4.1.2 Prediction Models

Various models have been deployed to predict plant performance based on previous trait data, with linear models being commonly used. There are a large number of classes of equations to model the effects of genetics on phenotype, but often the solutions to such systems are mathematically challenging and requires knowledge of constants which may be difficult to estimate. There is no particular reason to assume that the effect of the variation in genetic data on phenotype is linear, but restricting the class of equations to linear additive models simplifies the modelling and allows utilisation of a range of tools already developed for linear analysis. Typically, the only constants that need to be known are the first and second moments of the variables to be estimated [Henderson, 1984, Robinson, 1991]. Moreover, empirically, linear models have shown success in prediction of plant and animal performance in breeding populations [Bates et al., 2015].

The simple linear model assumes a linear correlation of the genotype on the trait, and least squares is a popular method to estimate the effect of the genotype. However, the assumption of the simple linear model is that each observation is independent of the others, which is not true for many experiments. Typically, experiments have a randomised block design, where replicates are spatially co-located to control for spatial variation and when experiments take place over multiple years, it is likely that there are systemic effects associated with the year of measurement.

In cases where genotypic data is unavailable, a linear mixed model (LMM) is often deployed to predict plant traits based on data gathered from previous years or other populations. The LMM assumes the observed phenotype is a linear combination of a set of fixed and random effects.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{4.1}$$

where $\mathbf{y}$ is a vector of observed phenotypes, $\mathbf{X}$ is a design matrix for the fixed effects, $\mathbf{b}$ is a vector of fixed effects, $\mathbf{Z}$ is a design matrix for the random effects and $\mathbf{e}$ is the error. The random effects are assumed to be drawn from a normal distribution with mean $\mathbf{0}$ and a known variance covariance matrix $\mathbf{u} \sim \text{MVN}~(\mathbf{0},~\mathbf{G})$; $\mathbf{e} \sim \text{MVN}~(\mathbf{0},~\mathbf{R})$. In the context of plant breeding, $\mathbf{G}$ represents the genetic covariance between the individuals and $\mathbf{R}$ represents their environmental covariance. When marker panels are dense with many markers of small effect, the infinitesimal model is appropriate and the probability of a random marker being identical by state to a random locus, allowing the following estimation:

$$\hat{\boldsymbol{G}} = \Phi \mathbf{X}\mathbf{X}' \tag{4.2}$$

where $\mathbf{X}$ is the genotype matrix and Φ is a proportionality constant [Endelman and Jannink, 2012].

In cases where the number of effects to be estimated differs from the number of unique measurements, then no unique solution can be obtained (more precisely, in the system $\mathbf{Ax} = \mathbf{b}$ with augmented matrix $[\mathbf{A}|\mathbf{b}]$, a unique solution exists if and only if $rank[\mathbf{A}] = rank[\mathbf{A}|\mathbf{b}]$).

Under this model, it is common to treat plant performance as a random variable, making predictions with the best linear unbiased prediction (BLUP) [Molenaar et al., 2018]. Originally developed for animal breeding, BLUP models the genotype effect on the observed phenotype as a random variable, with other effects including location, block and experimental year classed as fixed variables. BLUPs have properties that are desirable in prediction models: of all the linear models where predictions are unbiased, it has the minimum variance. Additionally, it incorporates shrinkage towards the mean, which is a desirable statistical property of an estimator, as it increases accuracy, leading to a smaller mean squared error [Piepho et al., 2008, Endelman and Jannink, 2012]. BLUP and best unbiased linear estimation (BLUE) solutions to the linear mixed model can be computed using Henderson's mixed model equations [Henderson, 1984]. These formulations are preferable when implementing computational algorithms for the linear mixed model for computational efficiency [Robinson, 1991, Piepho et al., 2008].

$$\begin{bmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} \\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{X'R^{-1}y} \end{bmatrix} \tag{4.3}$$

Mean and variances of the plant traits are to be estimated using the available data, with maximum likelihood (ML) being the most popular method. Assuming that the phenotypic measurements are drawn from a normal distribution, we can calculate the likelihood of μ and σ^2 given the data. Maximising the likelihood (or log-likelihood in practice as this function is monotonic and easier to perform calculus on) gives us the maximum likelihood (ML) estimate of σ^2. It is known that the ML estimation of variance is biased downwards in cases where a finite population is sampled. The bias in ML variance estimation arises due to the loss of one degree of freedom required in estimation of each of the fixed effect parameters (i.e. the mean) [Foulley, 1993]. The restricted maximum likelihood (REML) maximises a modified likelihood that has no mean component and thus avoids bias [Piepho et al., 2008].

In organisms that cannot be clonally propagated and exhibit sexual dimorphism, such as dairy cattle, best linear unbiased prediction (BLUP) has been used for decades to select sires with the highest estimated breeding values (based on measurements of the offspring from previous matings with each sire) to breed superior families [Henderson, 1984]. For plants, BLUP has been deployed to breed for various traits in ornamental carnation and geraniums, finding selection based on family indices worked at least as well as individual selection [Molenaar et al., 2018]. In potato, BLUP was deployed to breed for resistance to late blight resistance [Sood et al., 2020]. BLUP was also deployed to predict for expansion volume and yield and select families in maize [Viana et al., 2011].

When marker information is available, MAP can be implemented. MAP estimates the magnitude of effect of some marker(s) on the trait and estimates the plant trait (in the case of a linear models) as the sum of the effect of each marker that an individual has. The Kruskal Wallis (KW) test, a non-parametric implementation of ANOVA, tests if samples originate from the same distribution and can be applied to each marker to determine if it is associated with differences in traits [Broman, 2003]. The generated

H value is to be compared with the KW distribution (with an appropriate number of degrees of freedom) to determine a p-value, but when the number of groups or individuals are large, computation of the KW distribution becomes computationally infeasible. It remains unknown the best method to approximate the KW distribution when groups are large (as in the case of field experiments where hundreds of genotypes are assessed), but the χ^2 distribution is often used as an approximation [Meyer and Seaman, 2008].

When marker data are not densely or evenly spread throughout the genome, power to detect QTLs in sparse regions falls [Akond et al., 2019]. Consideration of QTLs in small intervals, using nearby markers increases the power to detect such QTLs and is termed interval mapping (IM) [Lande and Thompson, 1990]. A logarithm of odds (LOD) score measures the likelihood that a particular interval is associated with a QTL. A LOD score of 3 is often considered evidence of a true marker [Akond et al., 2019], but the probability of a false positive is dependent on the number of markers, size of the intervals considered and the heritability of the trait. Frequently, an empirical test for significance is performed by bootstrapping with a permutation test [Visscher et al., 1996, Churchill and Doerge, 1994].

A range of GP models have been deployed on both simulated and experimental datasets. The first model for GS utilises BLUPs to predict effects of the markers on traits [Meuwissen et al., 2001]. Genomic BLUP (GBLUP) utilises the linear model [Endelman, 2011]:

$$\mathbf{y} = \boldsymbol{\mu} + \sum_{i=1}^{n} \mathbf{W}\boldsymbol{q_i} + \mathbf{e} \tag{4.4}$$

where $\mathbf{y}$ is the observed phenotype, $\boldsymbol{\mu}$ is the mean, $\mathbf{W}$ is the genotypic design matrix, $\boldsymbol{q_i}$ is the effect of each SNP and n is the total number of markers. The variances of each SNP is assumed to be equal. As the number of markers (effects) is typically greater than the number of phenotypic records (measurements), the system is typically underdetermined and some form of regularisation is deployed to solve the system. Although stepwise algorithms exist to select a subset of markers, [Meuwissen et al., 2001, Habier et al., 2007], this method remains highly biased when strongly correlated markers are present. Ridge regression adds λ to the least squares estimator as a penalty, shrinking the effect size of each marker equally towards 0 to overcome the underdetermined system, whilst still using all markers [Piepho and Möhring, 2007]:

$$\hat{\boldsymbol{b}} \text{ (least squares)} = \arg\min_{b} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i b})^2 \tag{4.5}$$

$$\hat{\boldsymbol{b}} \text{ (ridge regression)} = \arg\min_{b} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i b})^2 + \lambda \sum_{j=1}^{p} b_j^2 \qquad \lambda \geq 0 \tag{4.6}$$

where j is the jth phenotypic record of p total phenotypic records. The ridge regression λ parameter controls the magnitude of the penalty and parameterises the relative importance of the data-dependent empirical error [Ogutu et al., 2012]. If variances of markers are assumed to be equal, $\lambda = \sigma_e^2/\sigma_u^2$, which is the ratio of the residual and marker variances, usually estimated through maximum likelihood methods [Endelman, 2011]. Under this model, when residual errors are large, more shrinkage allows for control of bias, and when marker effects are large, shrinkage is reduced to allow for estimation of true positives.

To better estimate variances of marker effects, Bayesian approaches were proposed. Bayesian models assume a more realistic distribution of marker effects as a prior and utilise data to update the prior to estimate marker effects on the phenotype. Two models of prior marker effects having an inverted χ^2 distribution and an inverted χ^2 distribution with an additional density distributions around zero, termed BayesA and BayesB respectively, were developed [Meuwissen et al., 2001]. Additional Bayesian models have been proposed utilising different combinations of prior distributions, density distributions around zero, and hyperpriors [Kärkkäinen and Sillanpää, 2012]. For example, BayesCπ and BayesDπ were developed to better estimate the constant part of the prior in BayesB [Habier et al., 2011]. Bayesian models have also been extended to utilise LASSO and elastic nets as methods of shrinkage.

Non-linear models of GP include reproducing kernel Hilbert spaces regression (RKHS) [Gianola and Van Kaam, 2008] and machine learning approaches utilising random forests support vector machines and neural networks [González-Recio et al., 2014, Heslot et al., 2012]. Although most experimental implementations of GS deploy multiple models and assess the prediction accuracy of the different models, in general, different models of GS perform similarly. Comparison of 11 GP models found that most models generated similar accuracies, with slightly better performance when deploying RKHS [Heslot et al., 2012]. Comparison of GBLUP, RKHS and BayesCπ for wheat yield showed little difference in selection accuracy [He et al., 2016]. Comparing three different Bayesian models, no significant improvement in selection accuracy was observed [Habier et al., 2011]. Comparison of three statistical approaches found limited effect of model selection, but accuracy increased significantly with higher marker counts in maize, cattle and pig [Zhang et al., 2019]. In strawberry, selection accuracies were slightly higher using RKHS than BayesB or GBLUP, but other factors had greater effects [Gezan et al., 2017].

Accuracy of selection positively correlates with the heritability, density of markers and training population size. Lower heritability traits are expected to have lower prediction accuracies as MAP cannot make predictions of the effect of environment on traits. In wheat significantly lower accuracy was observed when the size of the training population was reduced, regardless of the GP model deployed [He et al., 2016]. For maize, increasing training population size from 16% to 50% of the total population size increased selection accuracy from 0.24 to 0.33 for grain yield under well watered conditions [Zhang et al., 2017]. In an experiment in wheat, increases in prediction accuracy were observed as the training population size was increased plateauing at approximately 70% of the total population [Cericola et al., 2017]. Consistent with these observations, in pea, when the training population is at 70% of the total population size with significant drops in accuracy as training population size fell to less than 35% of the population size [Tayeh et al., 2015]. Genomic prediction in another experiment in wheat suggests plateauing of selection accuracy at around 70% of the total population for four plant quality traits [Norman et al., 2018]. Treating haploblocks as polyallelic markers for GP increased selection accuracy in maize, but use of individual SNP markers was more accurate in rice [Matias et al., 2017].

When dealing with polyploids, such as strawberry, markers may be in LD with QTLs on only one subgenome. When the resolution to a homeologous subgenome is unclear, the linkage of a detected marker to the QTL may be unclear. Although effort has been deployed to generate phasing techniques for polyploids [He et al., 2018], GP has also been deployed in polyploids with little modification of models. 15 traits were assessed using GP in triploid banana with selection accuracies that were not improved by inclusion of dosage information [Nyine et al., 2018]. Starch content and chopping quality predictions in autotetraploid potato had prediction accuracies 0.56 and 0.73 respectively, utilising the standard GBLUP and Bayesian models, though the authors note that large training populations were needed for high accuracies [Sverrisdóttir et al., 2017]. However, simulation of strawberry using pSBVB for sugar content indicated that the underlying genetic architecture and knowledge of markers from homeologous chromosomes was important in selection accuracy [Zingaretti et al., 2019]. The 90K SNP array and the 35K SNP array include 'haploSNPs', which utilise probes for known nearby subgenome variation to establish the homeologous subgenome a marker is associated with [Bassil et al., 2015].

To evaluate the utility of models, comparisons of selection accuracy between models can be made, as well as comparisons to estimations of heritabilities of traits. The broad sense heritability is defined as the proportion of total genetic variance to total phenotypic variance [Schmidt et al., 2019]:

$$H^2 = \sigma_G^2/\sigma_P^2 \tag{4.7}$$

where σ_G^2 is the genotypic variance and σ_P^2 is the phenotypic variance [Piepho and Möhring, 2007]. In plant trials, where genetically identical individuals are replicated, variance of the phenotypic observations within a genotype represents phenotypic variance while marker data allows estimation of genotypic variance due to identity by state. Under a randomised block design, there is assumed to be no genotype by environment correlations [Kruijer et al., 2014].

There were two main aims in this chapter. Firstly, based on a biparental strawberry mapping population, three between years prediction approaches (phenotype only, tMAS and GP) of 15 strawberry fruit quality traits relevant to breeders were to be assessed. Secondly, biological correlations and efficacy of selection was to be computed. Together, these datasets offer models for strawberry breeders on the methods and traits suitable for selection.

4.2 Methods

4.2.1 Plant Material

The biparental mapping population was used previously for genetic mapping [Antanaviciute, 2016]. Briefly, 188 seedlings were raised from a cross between two *F. ananassa* cutivars 'Redgauntlet' and 'Hapil', of which 120 were randomly selected. These individuals were clonally propagated with six replicates in the Autumn of 2015. Additionally, the parental genotypes and two check varieties, 'Sonata' and 'Elsanta', were included in the experiment, making a total of 744 individuals. The experiment took place at East Malling Research, at 51°17'15"N 0°27'12"E.

Seedlings were distributed in a randomised block design within three tunnels, with three beds per tunnel and two rows per bed. Each block was one-and-a-half rows.

Seedlings were planted in a double row zig-zag 35cm high, 50cm wide with 40cm between plants. Plants were allowed to establish over winter and dead material were removed in early 2016 and again in May 2016. Irrigation and fertigation was installed and performed according to conventional practice. Plants were also sprayed against common pests and diseases according to common practice. Harvesting took place three times a week (Monday, Wednesday and Friday) from when the first fruits developed until all fruits were harvested (17/06/2016 - 21/07/2016). In each harvest, all ripe fruits were collected from all plants for phenotypic analysis, except during the peak season, where only one or two tunnels were harvested for logistical reasons; in any week, all plants were harvested at least once. Harvesting was initiated in early morning at approximately 05:00 and classed *in situ* as marketable or unmarketable before delivery to a centralised location, where phenotyping of other traits took place.

4.2.2 Phenotypic and genotypic data

Phenotypic assessment of the plants took place on the same day as harvest, except during peak periods, where assessment took place over the day of harvest and the day after. Where assessment took place the next day, fruits were stored at 4°C overnight. Assessment was conducted using a modified RosBREED protocol for strawberry, with their standards defining the extremes and midpoints of the fruit quality traits where applicable [Mathey et al., 2013, Antanaviciute, 2016]. Assessment was primarily conducted by J. He and A. Karlstrom with occasional assistance from others. For all individuals, examples of phenotypes corresponding with measurements on the appropriate scales (Table 4.1) were demonstrated and agreed before phenotyping took place.

A total of 15 traits were assessed. Marketable and unmarketable fruits were collected and weighed separately for each plant, and summed for all harvests throughout the season. For each other phenotype, a single value was generated from assessment of ten randomly selected fruits (where available) from the marketable portion of each plant at each harvest, except pH, soluble solids, and firmness, where three, twenty and ten fruits were randomly selected for analysis over the season respectively.

pH was measured by releasing a drop of strawberry juice onto a pH meter; firmness was assessed by gentle depression of the fruit by a robotic arm and measurement of the deformation (Firmtech Umweltanalytische Produkte GmbH). Soluble solids content was measured by releasing a drop of juice from a randomly selected fruit onto an interferometer; cap size was a visual assessment of the width the cap relative to the neck of the fruit; appearance was a visual assessment of the fruit ranging from very malformed to symmetrical and attractive; external colour was a visual assessment of the fruit colour; glossiness was a visual assessment of the shine of the fruit; achene position was a visual assessment of how protruding the achenes were; seediness was a relative measure of the density of visible achenes; fruit shape was a visual assessment of the ratio of fruit height to width; neck line was visual assessment of the shape of the neck; skin strength was the number of fruits with broken skin after ten fruits were rubbed gently with a thumb; and internal colour was the relative colours of the inside of the fruit after bisection [Mathey et al., 2013] Table 4.1. In addition to data collected in 2016, phenotypic data from a previous study on the population from 2013 to 2015 were included in the analysis [Antanaviciute, 2016]. The 2013 - 2015 dataset was defined as the prediction dataset and the

Trait	Assessment	Scale/units
Marketable Yield	Weigh on scales	g
Unmarketable Yield	Weigh on scales	g
Firmness	Depression by robotic arm	gmm^{-1}
Soluble Solids Content	Meniscus of juice in refractometer	Brix°
pH	Juice in pH meter	
Achene Position	Protrusion of achenes from skin	$1-3$
Seediness	Relative number of achenes per area	$1-3$
Cap size	Diameter of cap relative to sholder	$1-3$
Appearance	Relative attractiveness of fruit	$1-5$
Shape	Oblate - long conic shape	$1-5$
Redness	Redness of external colour	$1-5$
Glossiness	Gloss of external skin	$1-5$
Neck Line	Depression of neck relative to shoulder	$1-5$
Skin strength	Relative skins broken after gentle stroking	$1-5$
Internal Colour	Redness of internal after bisection	$1-5$

Table 4.1: 15 fruit quality traits and their scales assessed in 2016 in a biparental 'Redgauntlet' × 'Hapil' mapping population

2016 dataset was defined as the validation dataset. DNA extraction for genotyping was performed on single young leaves using the Qiagen DNeasy kit according to the manufacturer's instructions. Genotyping was performed using the 90K array with genotypes being calculated in accordance to the manufacturer's instructions [Verma et al., 2017] (Cockerton, unpublished). Filtering was performed on the dataset to remove non-segregating markers and remove redundant information by maintaining only one instance of markers that segregated identically. After filtration, 3436 segregating markers were identified and included for genotypic analysis. Genetic rogues, defined as individuals with non-parental genotypes or individuals that were genetically identical to apparently other genotypes were excluded from analysis. After data filtration, 66297 phenotypic records were used in the training dataset and 24908 phenotypic in the validation set, amounting to a 77% and 23% data split respectively.

4.2.3 Phenotypic Predictions and Correlations of Fruit Quality Traits

116 individuals including the parental and check varieties were included for phenotypic analysis. To explore the data, the phenotypic values were plotted against the genotypes, with the parental and check cultivars highlighted. For all traits, the mean across all blocks in the years was computed, except marketable yield and unmarketable yield, where the sum of all records were computed, and pH, where the mean of the concentration of hydronium ions was calculated, and the result converted to the logarithmic pH scale. The differences in rank of the parental strains were also computed.

In the absence of genotypic data, given the unbalanced data, it is conventional to deploy BLUP to predict plant performance. Calculation of the BLUP was performed using the 'lmer' command from the 'lme4' package in R [Bates et al., 2015]. The prediction model included all data from 2013 - 2015, treating the year and blocks as fixed effects and the genotypes as random effects. It was assumed that there was no differential interaction effect between genotype and block or year. In order to compute a comparable figure for the validation dataset, a similar linear model was fitted for the 2016 data, treating blocks as fixed effects and the genotypes as random effects. The variance estimation method for both models was 'REML' [Bates et al., 2015]. Random effects were extracted from the model using the 'ranef' command and their Pearson's correlation coefficients were computed using the 'cor' function as a measure of prediction accuracy. The concordance correlation coefficient (CCC) of the predictions were also computed using the 'CCC' function from the 'DescTools' package [Signorell et al., 2021].

To estimate correlations between traits, the Pearson's correlation coefficients of BLUPs for every pair of traits from the prediction, validation and total datasets were also computed. p-values under the null hypothesis that the correlations were not different from 0 were calculated and a Bonferroni correction was performed using the number of pairwise tests performed ($p < 0.05$, $n = 315$).

4.2.4 Traditional Marker Assisted Prediction

103 progeny were included in genotypic analyses. In order to maximise power to detect markers associated with QTL, all phenotypic data from all years were included in the marker discovery phase. Two methods of marker discovery were implemented: the Kruskal Wallis (KW) test and interval mapping (IM). Marker discovery for both methods were conducted using MapQTL5 in accordance to the user manual [van Ooijen, 2009]. The mean of all traits were calculated as input for the 'qua' file, except marketable yield and unmarketable yield, where the sum of all records were computed, and pH, where the concentration of aqueous hydronium ions were calculated.

For the KW analysis, the resulting H statistic (and their associated degrees of freedom) was extracted from MapQTL5 for p value estimation and multiple testing correction. As an approximation to the KW distribution, the H statistic was compared to the χ^2 distribution with the appropriate number of degrees of freedom to yield a p value for each marker being associated with a QTL. Computation of the χ^2 distribution was performed using the 'pchisq' function from the 'stats' package in R. The Benjamani-Hochburg (BH) correction [Benjamini and Hochberg, 1995] was applied to adjust for multiple testing of markers to control false discovery rate. The critical value for false discovery rate was set to an exploratory rate of 0.2. Computation of the BH correction was performed using the 'p.adjust' function, also from the 'stats' package in R.

For the IM test, significance thresholds were first generated. A permutation test was conducted using the 'permutation test' function of MapQTL5. 100 permutations were simulated for all traits and the 95th percentile of the genome-wide significance levels were taken as the threshold for a statistically significant marker. As markers physically close together are likely to be in LD with each other as well as QTLs, a simple clustering algorithm was implemented to determine if a set of markers with significant LOD values described the same QTL. When a LOD peak was identified at a locus, all other markers and intervals were scanned from both directions until a marker was identified with a LOD score 2 units less than peak. All markers scanned were clustered as describing the same

QTL, with the marker with the highest LOD score selected as representative of that peak.

Markers from KW and IM were pooled and used for prediction of plant performance. Prediction was calculated as the mean of the trait values estimated by MapQTL5 for a given allele for each marker. Prediction accuracy was defined as the Pearson's correlation coefficient between the tMAP estimations of the prediction data and the BLUPs of the validation data as described in 4.2.3. Additionally, the CCC between the values was calculated.

4.2.5 Genomic Prediction

GP was implemented through a two step process. In the first step, BLUPs were calculated for the prediction and validation data as described in 4.2.3. GBLUP was conducted using the 'mixed.solve' function from the 'rrBLUP' package [Endelman, 2011]. The Y and Z matrices were defined as the phenotypic BLUPs subsection 4.2.3 and the design matrix of the genotypic markers respectively. The predictor of an individual is the GEBV and was defined as the sum of all corresponding marker effects of the individual [Gezan et al., 2017]. To evaluate accuracy, two methods of prediction accuracies were calculated. The BV BLUP cor method correlates the GEBV of the training population with the BLUPs of the validation population and the GEBV cor method correlates the GEBVs of the training population and the GEBVs of the validation population.

4.2.6 Heritability

Heritability for the validation dataset was calculated according the following model [Piepho and Möhring, 2007]:

$$H^2 = \sigma_G^2/(\sigma_G^2 + \sigma_{Ge}^2/m + \sigma^2/rm) \tag{4.8}$$

where σ_{Ge}^2 is the genotype-environment variance and σ^2 is the residual error variance. It was assumed that $\sigma_{Ge}^2 = 0$ due to the randomised block design of the experiment.

4.3 Results and Discussion

4.3.1 Phenotypic Data

To explore the range of data, the mean phenotypic traits of each genotype along with the parental and check varieties were plotted (Figure 4.1). Comparison with traits described in literature may be unreliable because the scores of individuals are dependent on their environment. For example, the average colour for 'Sonata' and 'Elsanta' were found to be statistically different, but almost indistinguishable to the human eye [Giné Bordonaba and Terry, 2010]. Total sugar content in 'Sonata' is higher than 'Elsanta' when fully irrigated, but not statistically distinguishable when under water stress. Similarly, 'Sonata' contains more total acid than 'Elsanta' when fully irrigated, but statistically indistinguishable when under water stress [Giné Bordonaba and Terry, 2010].

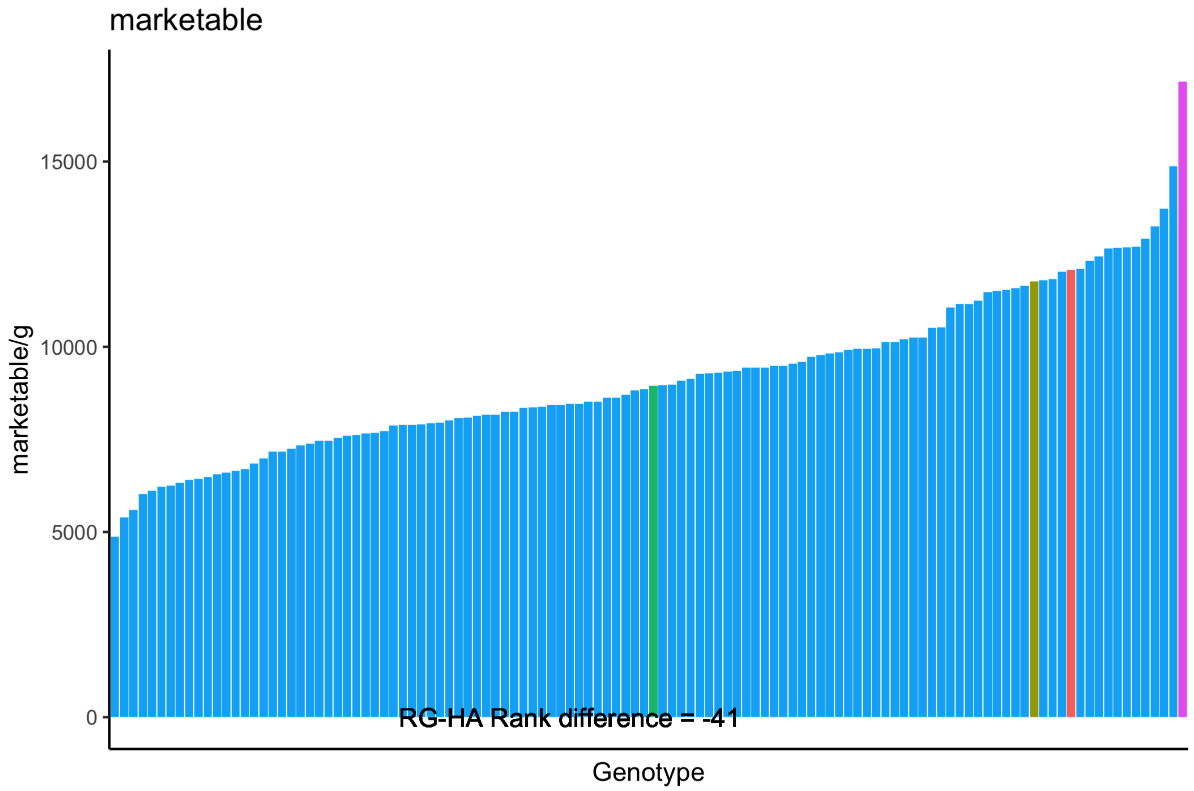
However, consistent with expectations of a modern cultivar on the market, the marketable yield of 'Elsanta' and 'Sonata' were both high, with 'Sonata' being markedly

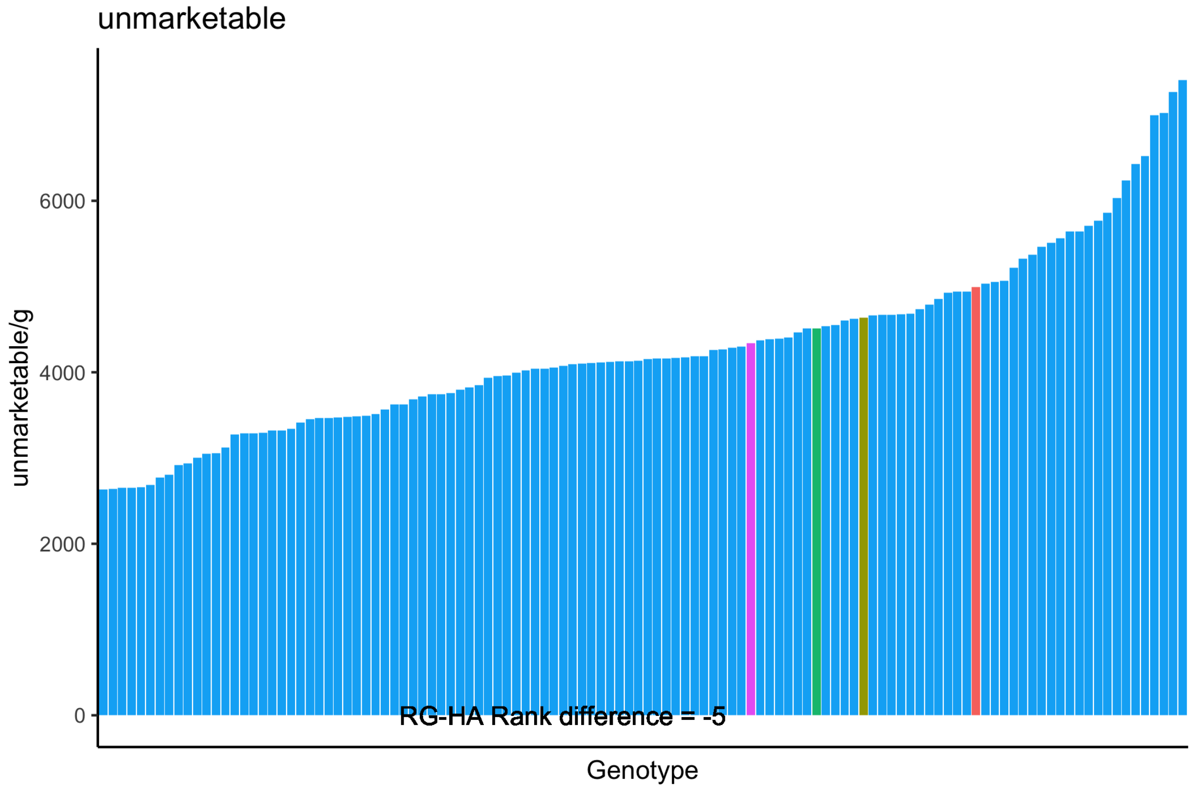
Trait	Heritability
Marketable Yield	0.33
Unmarketable Yield	0.33
Firmness	0.15
Soluble Solids Content	0.13
pH	0.39
Achene Position	0.32
Seediness	0.14
Cap size	0.24
Appearance	0.15
Shape	0.25
Redness	0.31
Glossiness	0.08
Neck Line	0.30
Skin strength	0.17
Internal Colour	0.34

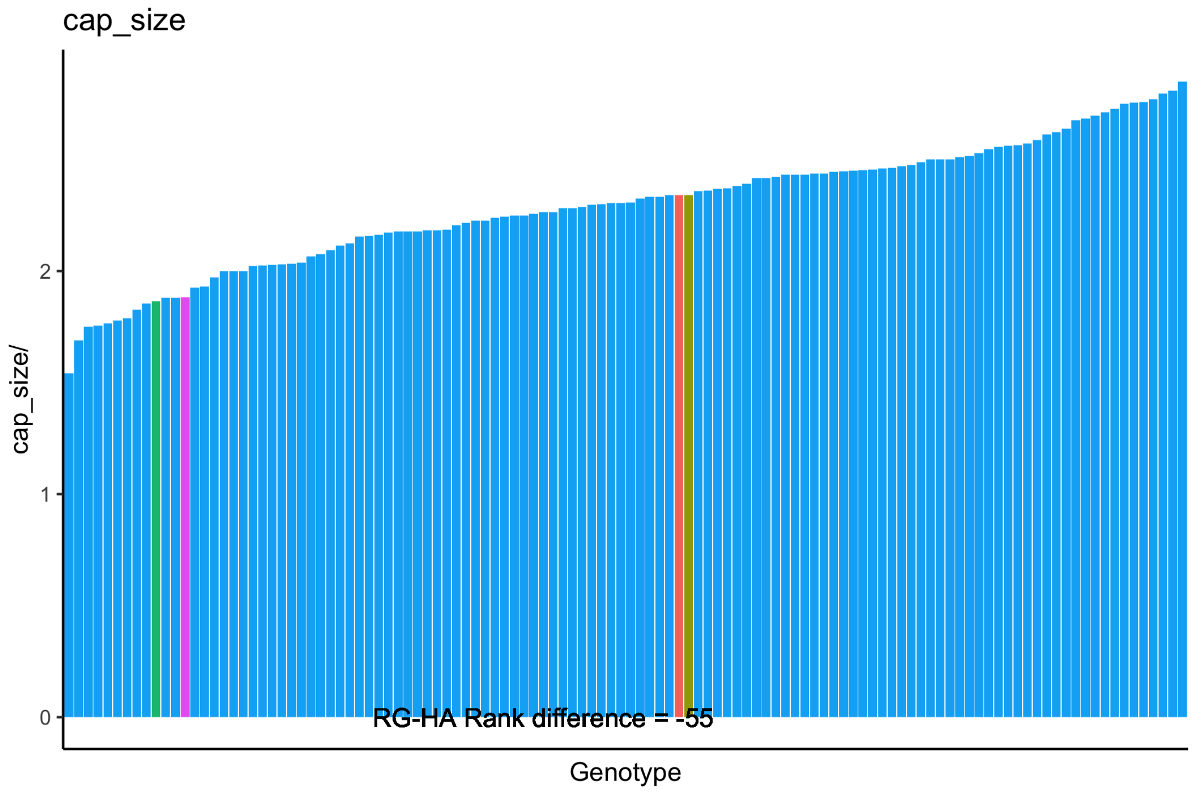
Table 4.2: heritability of 15 fruit quality traits in a biparental 'Redgauntlet' × 'Hapil' mapping population

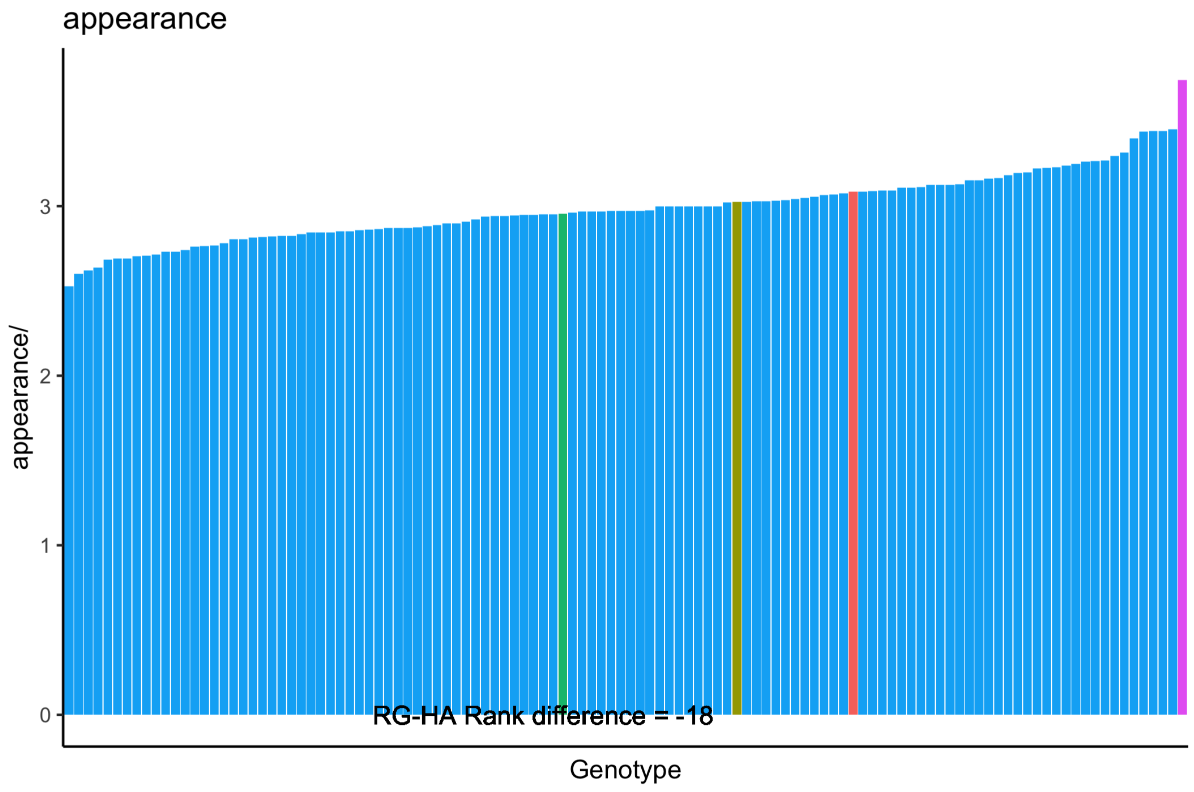
higher than any other genotype. 'Sonata' scored significantly higher than any other genotype for appearance with 'Elsanta' scoring moderately. Both check varieties scored highly in glossiness, and skin strength, both of which are desirable traits for the market. Moreover, both check varieties were significantly firmer than other genotypes, a trait that breeders select for as firmness correlates with post harvest shelf life [Salentijn et al., 2003]. 'Sonata' is described as producing oblate fruits, and this trait is apparent in its low shape and neck line score. Interestingly, both check species have relatively low brix values, despite consumers indicating sweetness as an important trait [Colquhoun et al., 2012] and neither have low unmarketable yields. The latter observation can be explained as marketable yield and unmarketable yield are correlated and commercial cultivars are expected to have high yields. Taken together, the data distributions are consistent with known and expected traits of the check varieties, suggesting the measured phenotypes are representative of the phenotypes of strawberries.

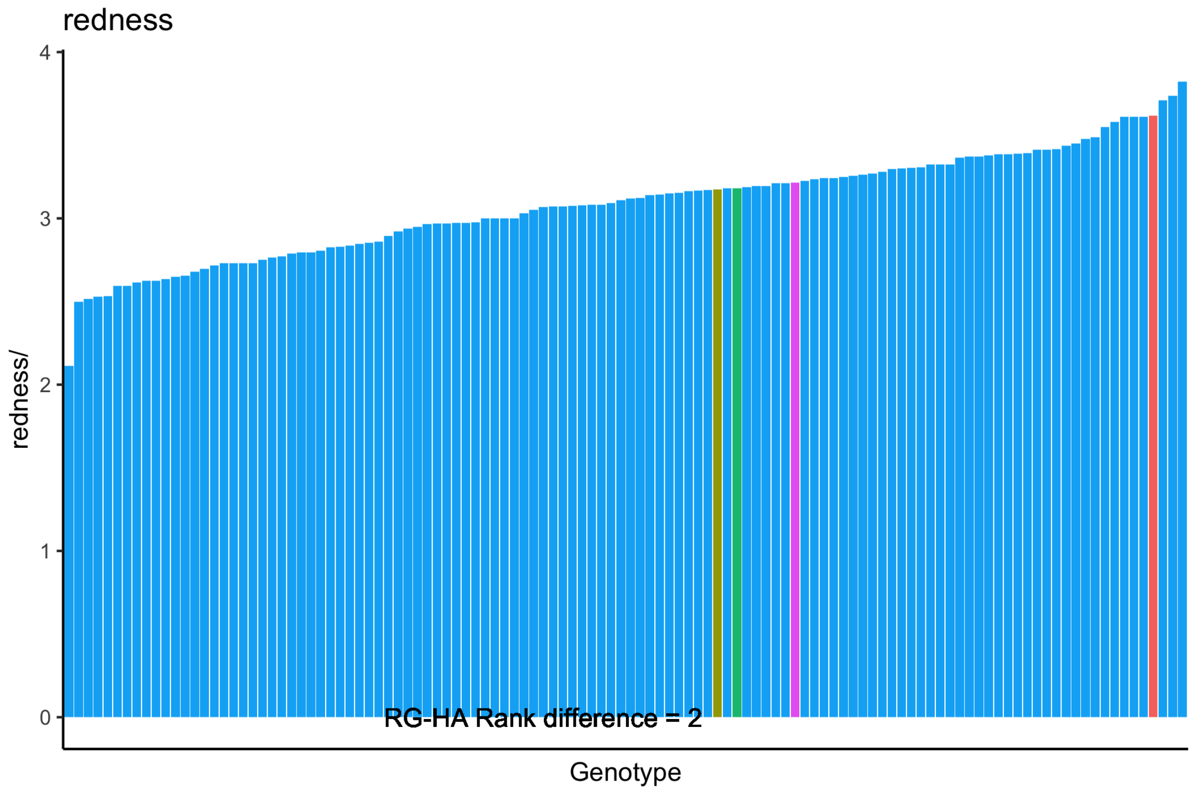
In the case of glossiness, achene position and pH, parental phenotypes have a large rank difference (> 70), suggesting that the parents have differing alleles controlling the traits, with offspring inheriting heterozygously, displaying intermediate phenotypes. In the case of unmarketable yield, appearance, redness, seediness, shape and brix, the difference in rank of the parents is small (< 25). The distribution of achene position, shape and internal colour shows progeny with marked extremes. This is typical of traits under the control of a few large effect QTLs as segregation pyramids those QTLs by chance in a few individuals. Thus, it may be expected to identify large effect markers using tMAP for these traits. Heritabilities were also calculated Table 4.2.

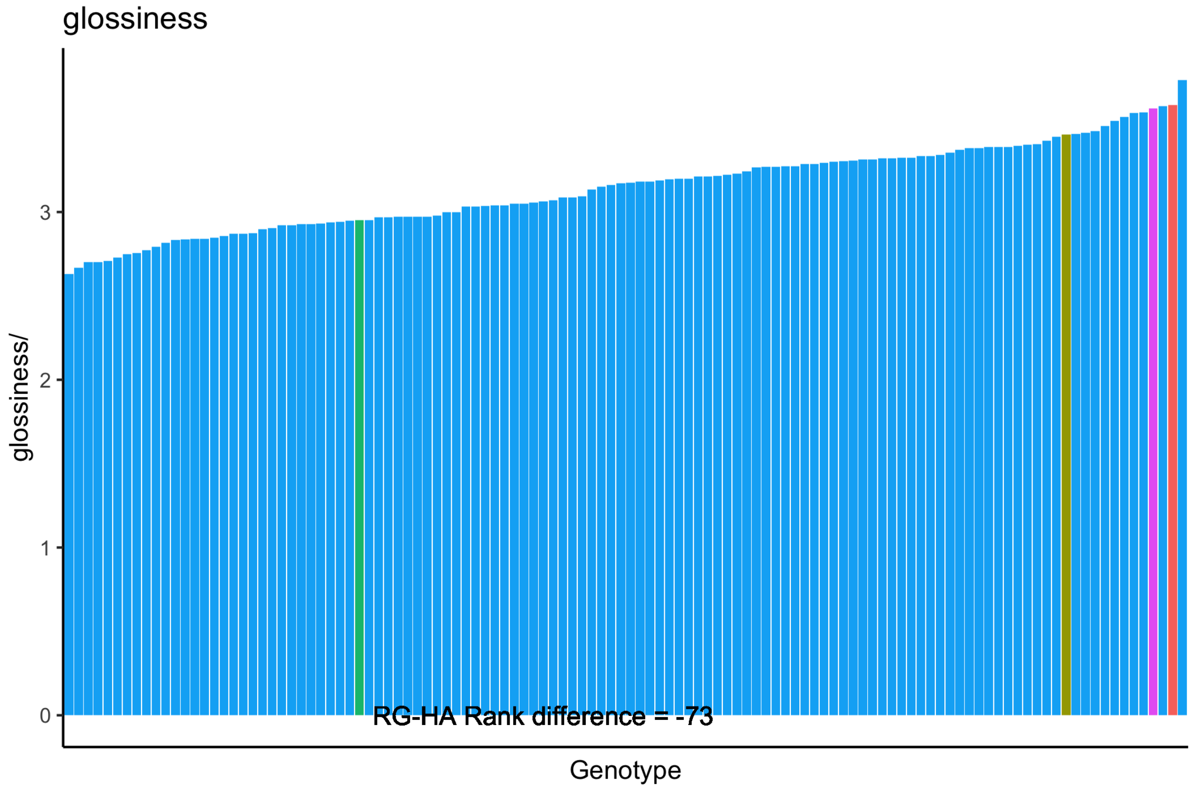

marketable
marketable/g
15000
10000
5000
0
RG-HA Rank difference = -41
Genotype

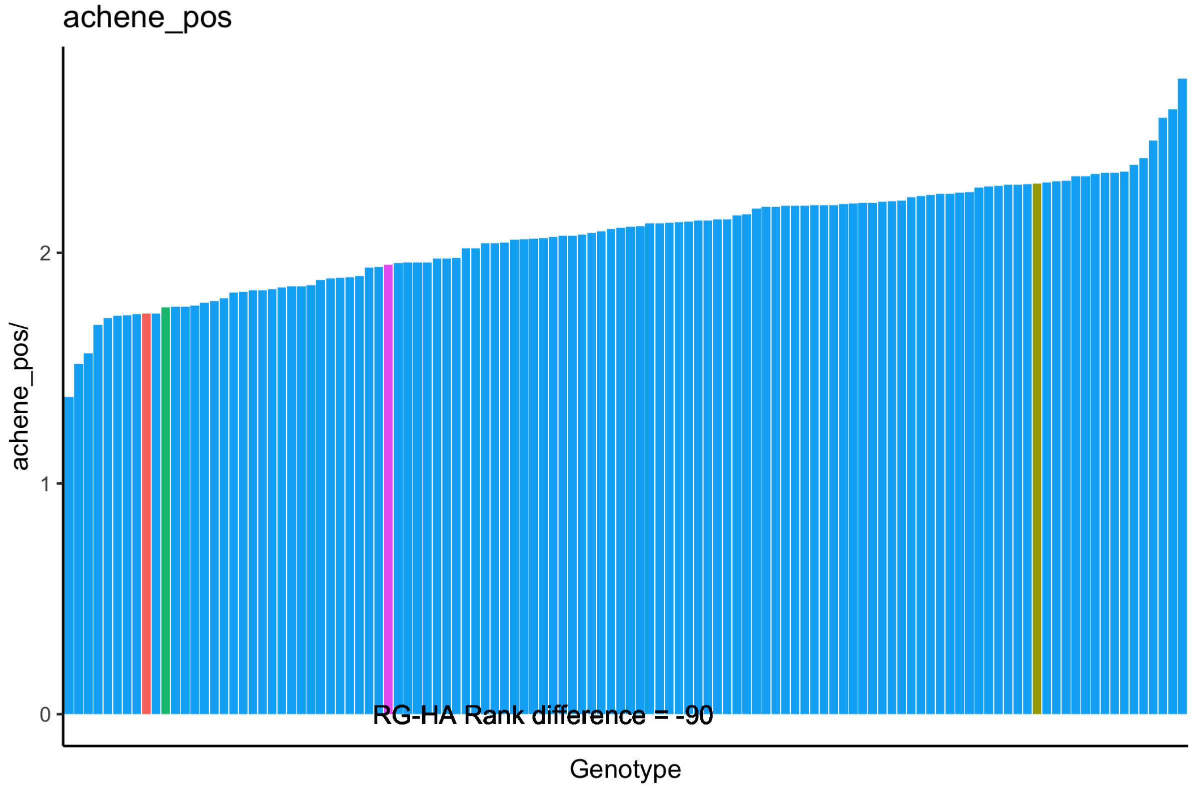

unmarketable
unmarketable/g
6000
4000
2000
0
RG-HA Rank difference = -5
Genotype

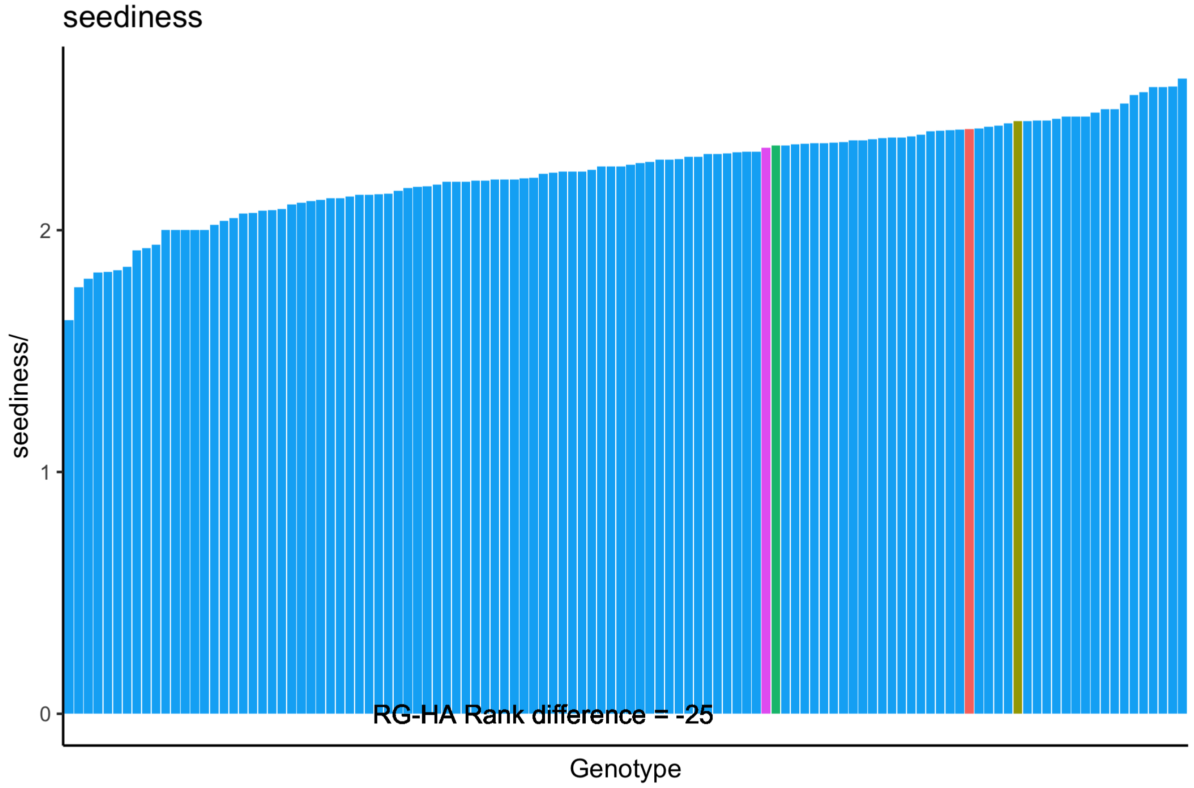

cap_size
cap_size/
0
1
2
RG-HA Rank difference = -55
Genotype

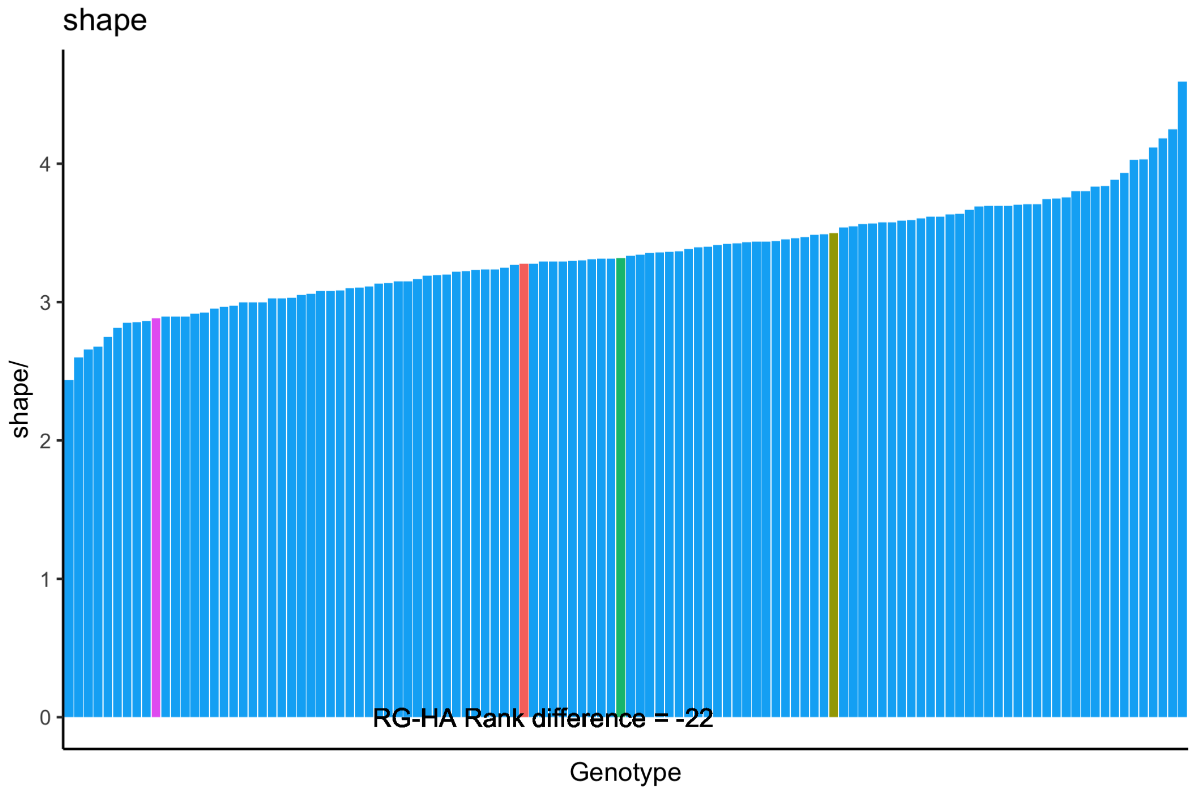

appearance
appearance/
0
1
2
3
RG-HA Rank difference = -18
Genotype

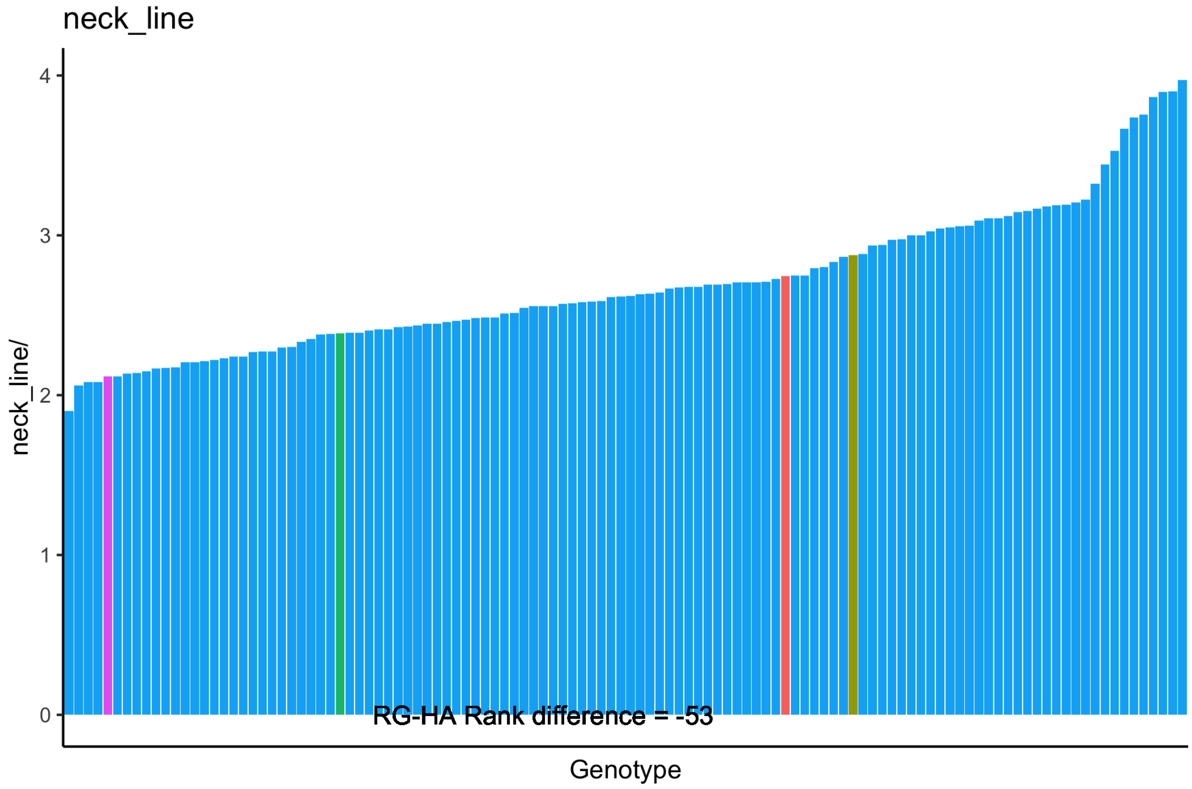

redness
4
3
2
1
0
redness/
RG-HA Rank difference = 2
Genotype

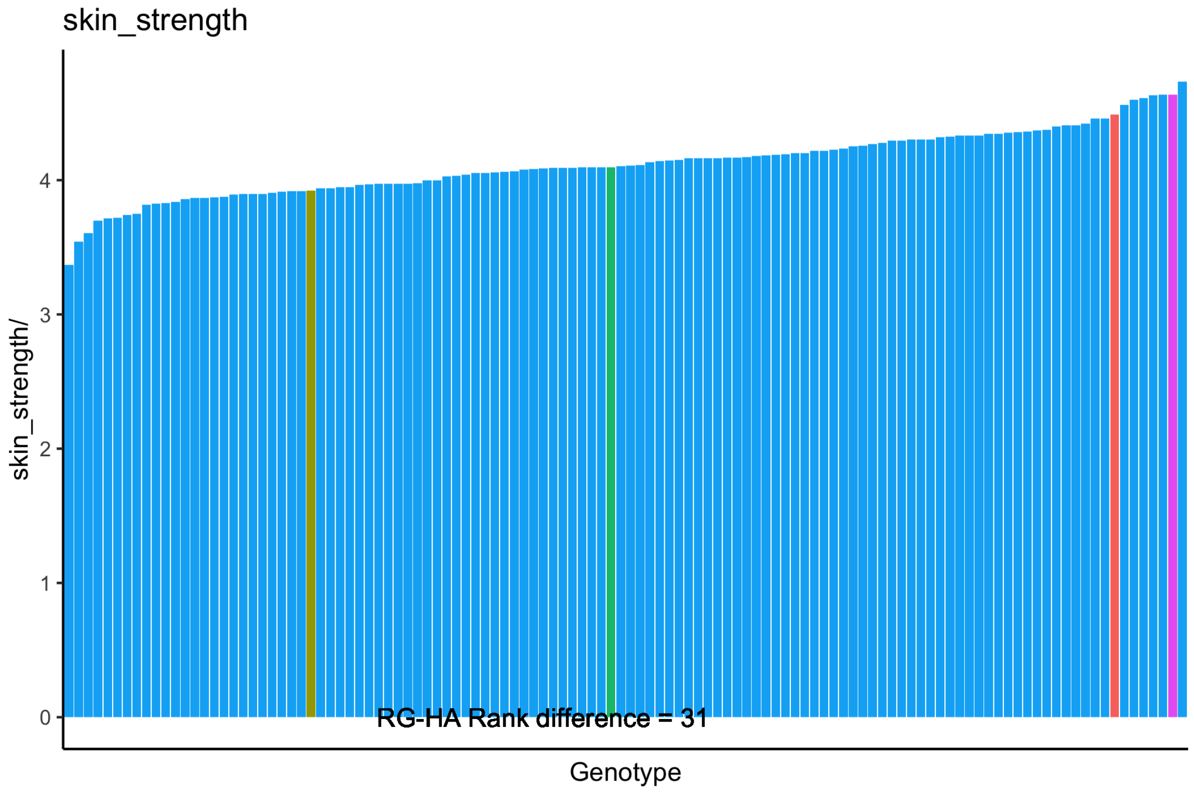

glossiness
3
2
1
0
glossiness/
RG-HA Rank difference = -73
Genotype

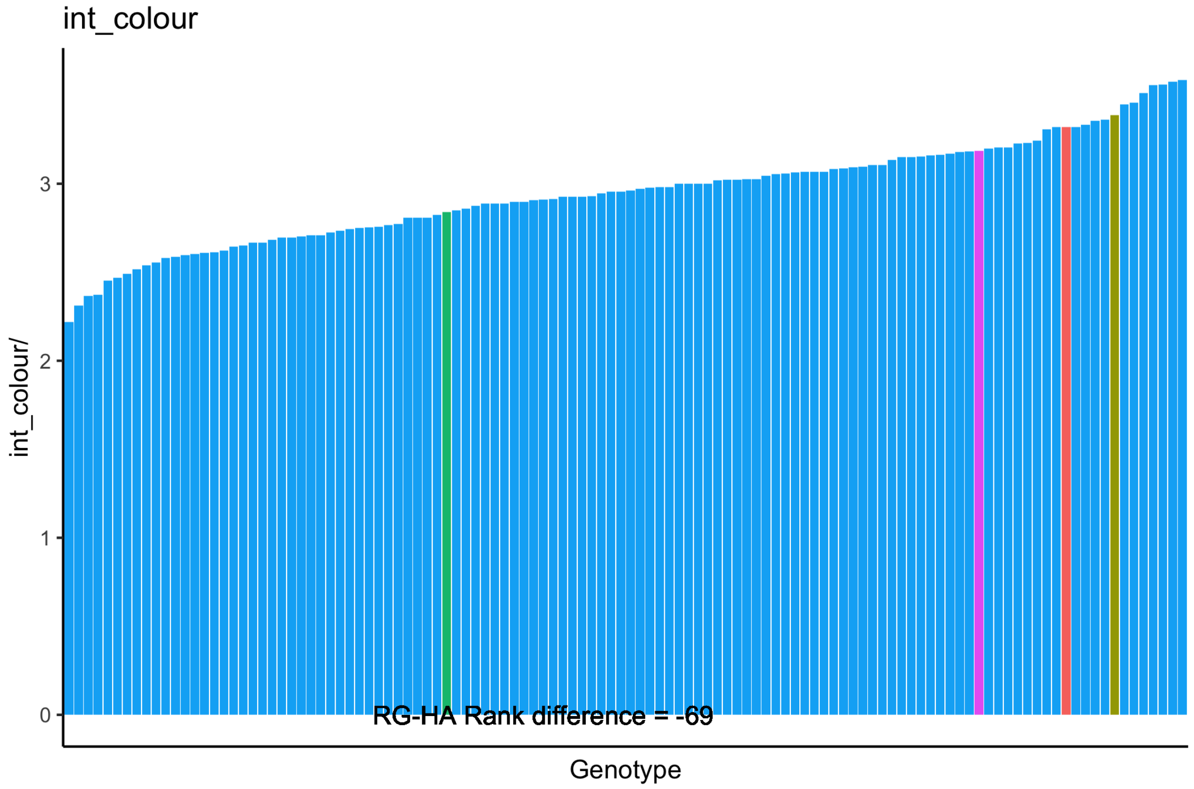

achene_pos
achene_pos/
2
1
0
RG-HA Rank difference = -90
Genotype

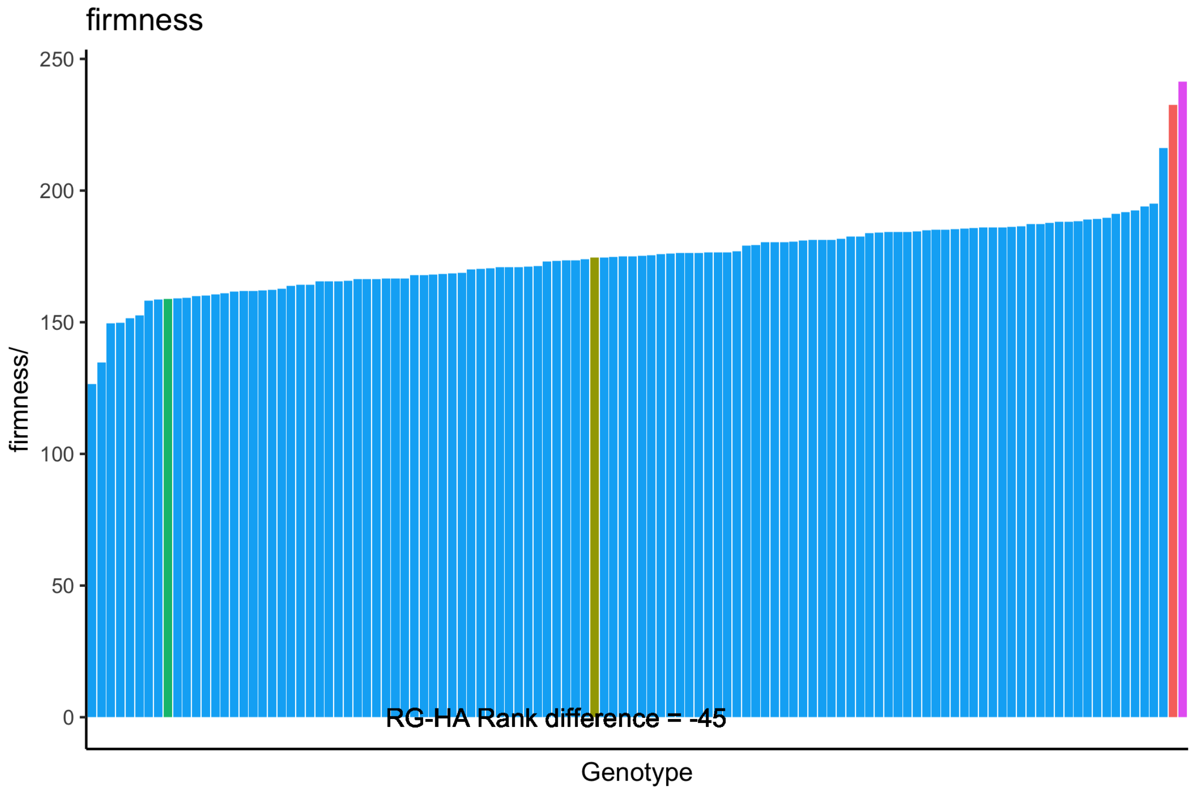

seediness
seediness/
2
1
0
RG-HA Rank difference = -25
Genotype

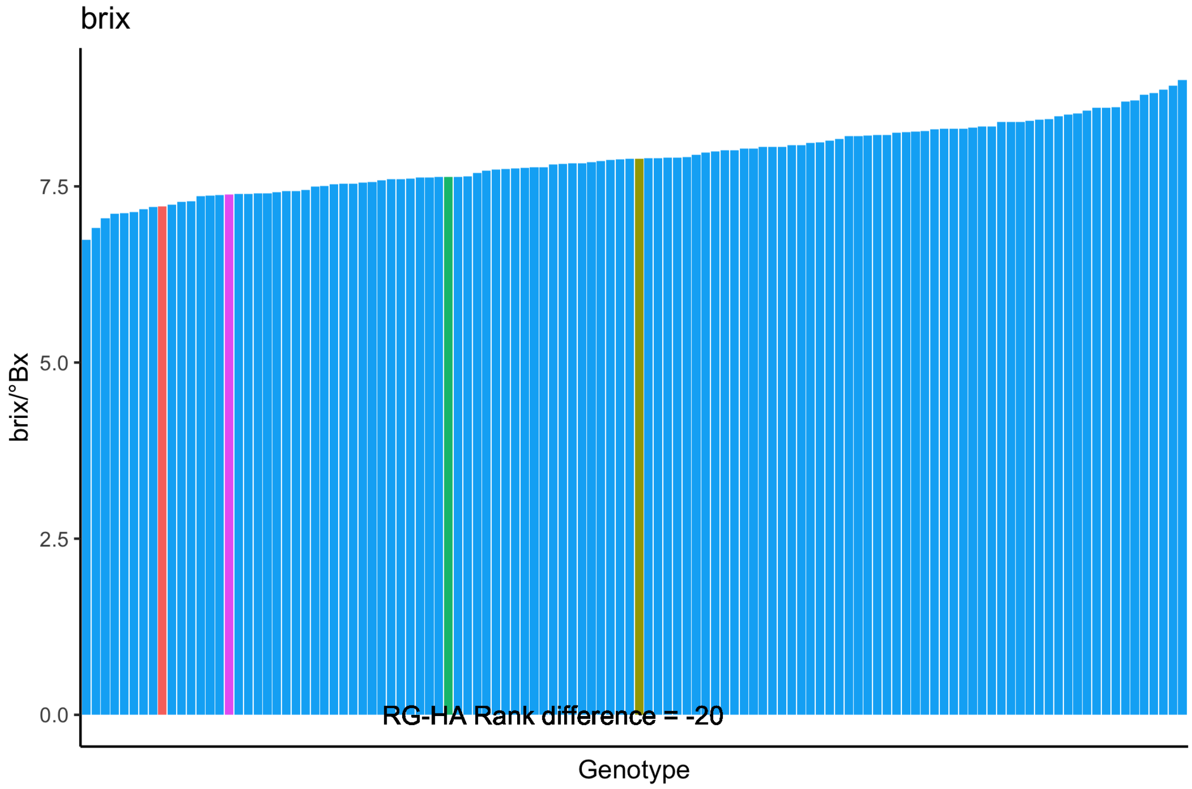

shape
shape/
0
1
2
3
4
RG-HA Rank difference = -22
Genotype


neck_line
neck_line/
0
1
2
3
4
RG-HA Rank difference = -53
Genotype


skin_strength
skin_strength/
0
1
2
3
4
RG-HA Rank difference = 31
Genotype


int_colour
int_colour/
0
1
2
3
RG-HA Rank difference = -69
Genotype


firmness
250
200
150
100
50
0
firmness/
RG-HA Rank difference = -45
Genotype


brix
7.5
5.0
2.5
0.0
brix/°Bx
RG-HA Rank difference = -20
Genotype

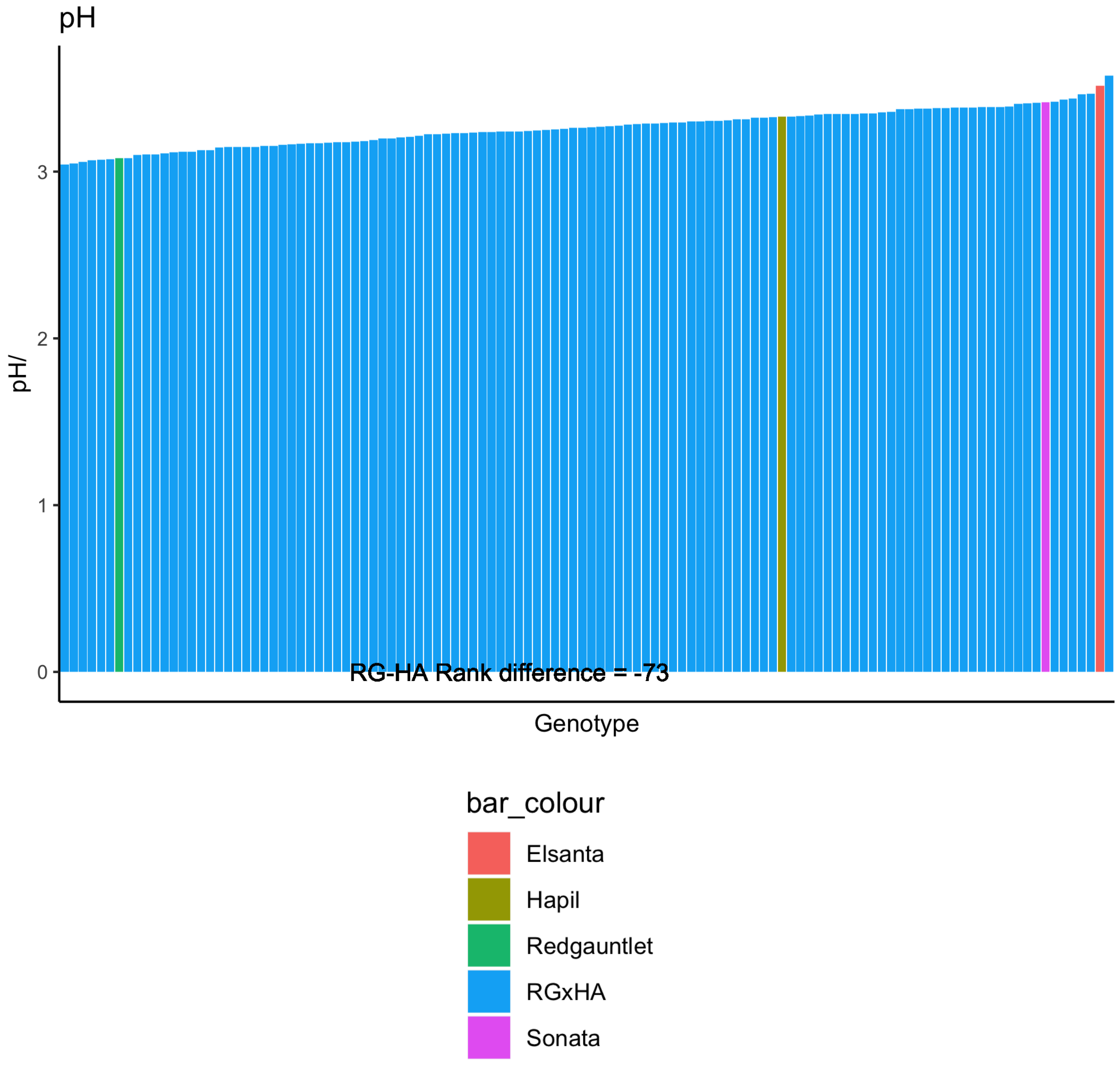


Figure 4.1: Distribution of means of individuals from a biparental mapping population for 15 fruit quality traits across all years.

4.3.2 Correlations between Phenotypes

Fifteen statistically significant correlations between traits were identified based on analysis using all data. Of these traits, the strongest was between redness and internal colour (0.82), followed by neck line and shape (r = 0.51), marketable yield and unmarketable yield (r = 0.50) and skin strength and firmness (r = 0.50). 14 of the 15 identified correlations were positive, with a single negative correlation identified between firmness and neck line. In order to investigate the reliability of correlations between the validation and prediction datasets, pairwise correlations of traits were also computed for the prediction and validation datasets. The three pairs of traits that had the strongest correlations were consistently correlated across years; internal colour and redness, neck line and shape and marketable yield and unmarketable yield.

In an experiment in Bangladesh, strong pairwise correlations were found between total fruit weight, fruit length, fruit diameter and brix values [Mehraj and Jamal Uddin, 2014]. However, in this study, no such correlations were observed. Consistent with the correlations observed in previous years [Antanaviciute, 2016], a strong correlation was observed between redness and internal colour between redness and glossiness and but not between cap size and shape. However, no other consistent correlation was observed and many fewer correlations were observed in this experiment. This may be due to differences in the method for calculating correlations and it is unclear if multiple testing corrections were applied for the result of Antanaviciute.

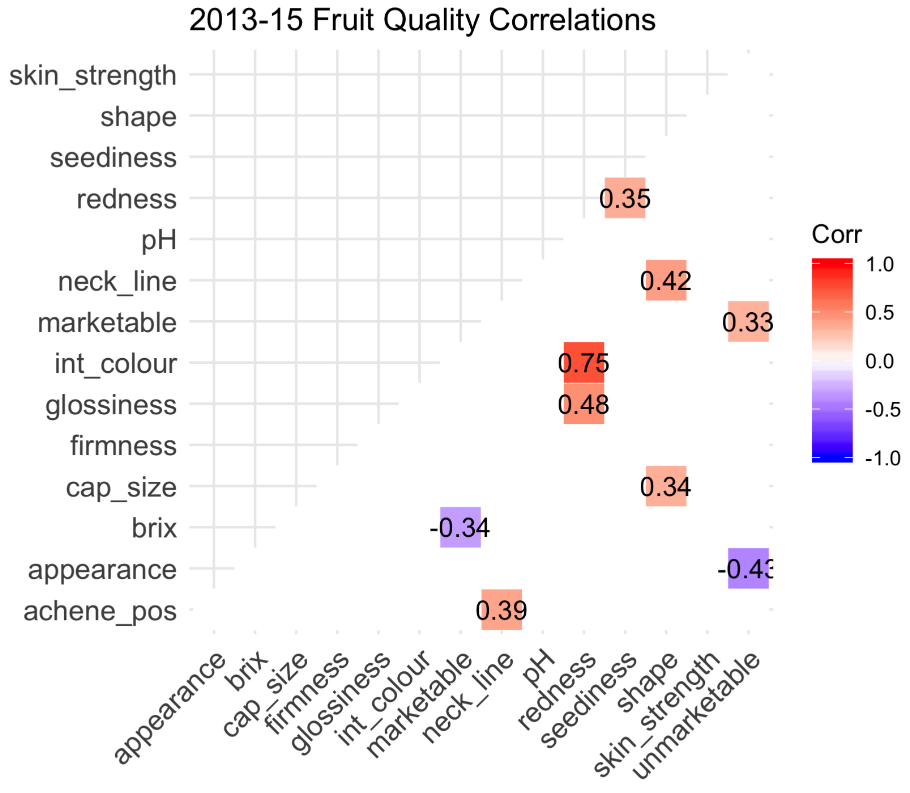
4.3.3 Between Years Phenotypic Prediction

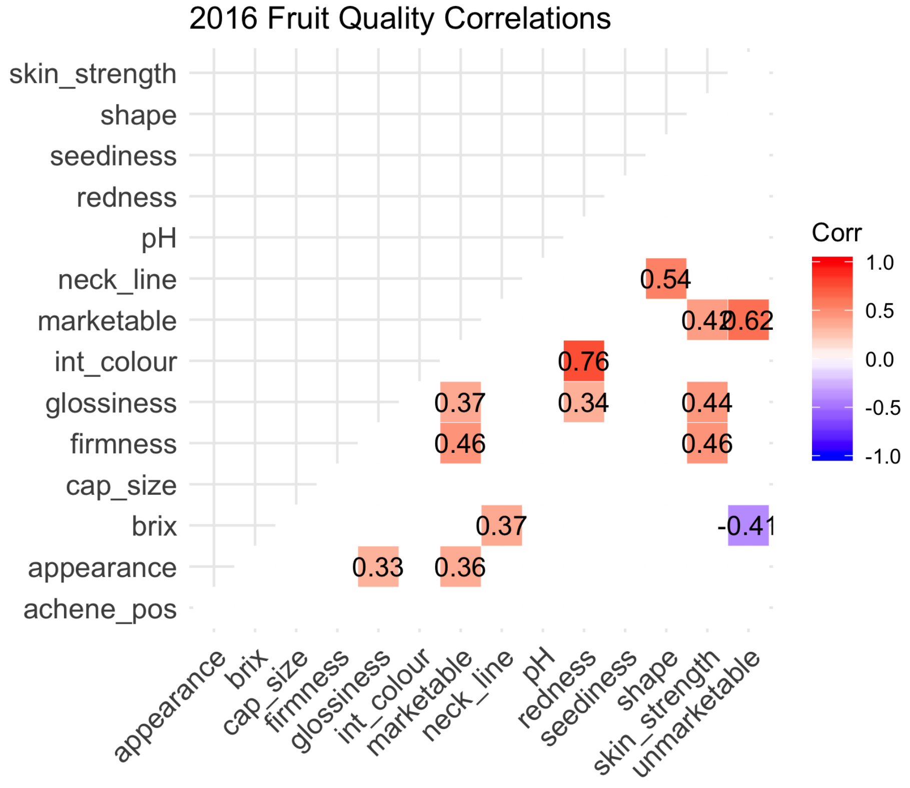
Correlations and concordances were computed between BLUPs of the prediction and validation dataset (Figure 4.3). High correlations (> 0.7) were found for firmness, neck line and redness. Low correlations (< 0.4) were found in glossiness, marketable yield, skin strength and unmarketable yield. High concordances (> 0.5) were found for redness and shape while low concordances (< 0.2) were found for appearance, firmness, glossiness, internal colour, marketable yield, pH, skin strength, and unmarketable yield.

In the case of traits such as brix and seediness, concordance was similar to correlation, indicating the absolute values were similar where there was correlation. In cases such as marketable yield, internal colour, pH and skin strength, the concordance was much lower than correlation, indicating that even when there was correlation, the absolute values of these traits did not match. In the case of marketable fruit, this is likely a reflection of the significant effect of environment on the trait; in the case of internal colour, this may be a rater effect.

4.3.4 Traditional Marker Assisted Prediction

Using IM, one marker for achene position, redness and unmarketable yield each were identified, on LG4B at 53.0cM, LG7A 49.9cM and LG1C at 9.9 cM respectively. The LOD values for each marker were 3.91, 4.14 and 5.14 respectively, exceeding the permutation


2013-15 Fruit Quality Correlations
skin_strength
shape
seediness
redness
pH
neck_line
marketable
int_colour
glossiness
firmness
cap_size
brix
appearance
achene_pos
0.35
0.42
0.33
0.75
0.48
0.34
-0.34
-0.43
0.39
appearance
brix
cap_size
firmness
glossiness
int_colour
marketable
neck_line
pH
redness
seediness
shape
skin_strength
unmarketable
Corr
1.0
0.5
0.0
-0.5
-1.0


2016 Fruit Quality Correlations
skin_strength
shape
seediness
redness
pH
neck_line
marketable
int_colour
glossiness
firmness
cap_size
brix
appearance
achene_pos
0.54
0.42
0.62
0.76
0.37
0.34
0.44
0.46
0.46
0.37
-0.41
0.33
0.36
appearance
brix
cap_size
firmness
glossiness
int_colour
marketable
neck_line
pH
redness
seediness
shape
skin_strength
unmarketable
Corr
1.0
0.5
0.0
-0.5
-1.0

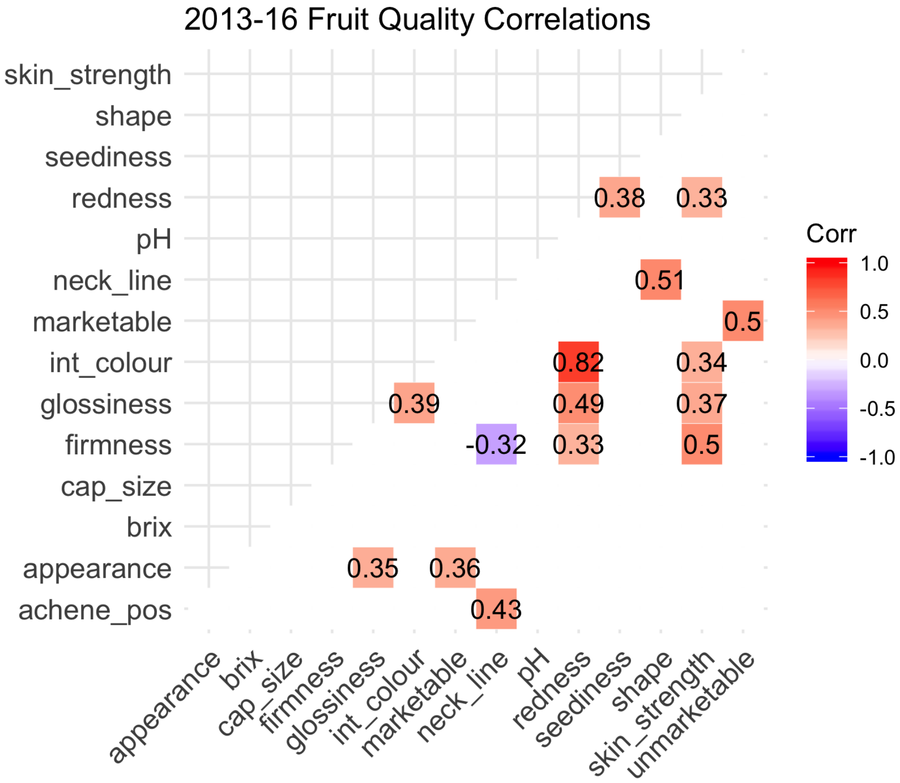


Figure 4.2: Correlation matrix between BLUPs for fruit quality traits across different years. Colours and numbers indicate the magnitude and direction of the correlation. For clarity, only statistically significant correlations are included in the plot ($p < 0.05$, Bonferroni correction $n = 315$).

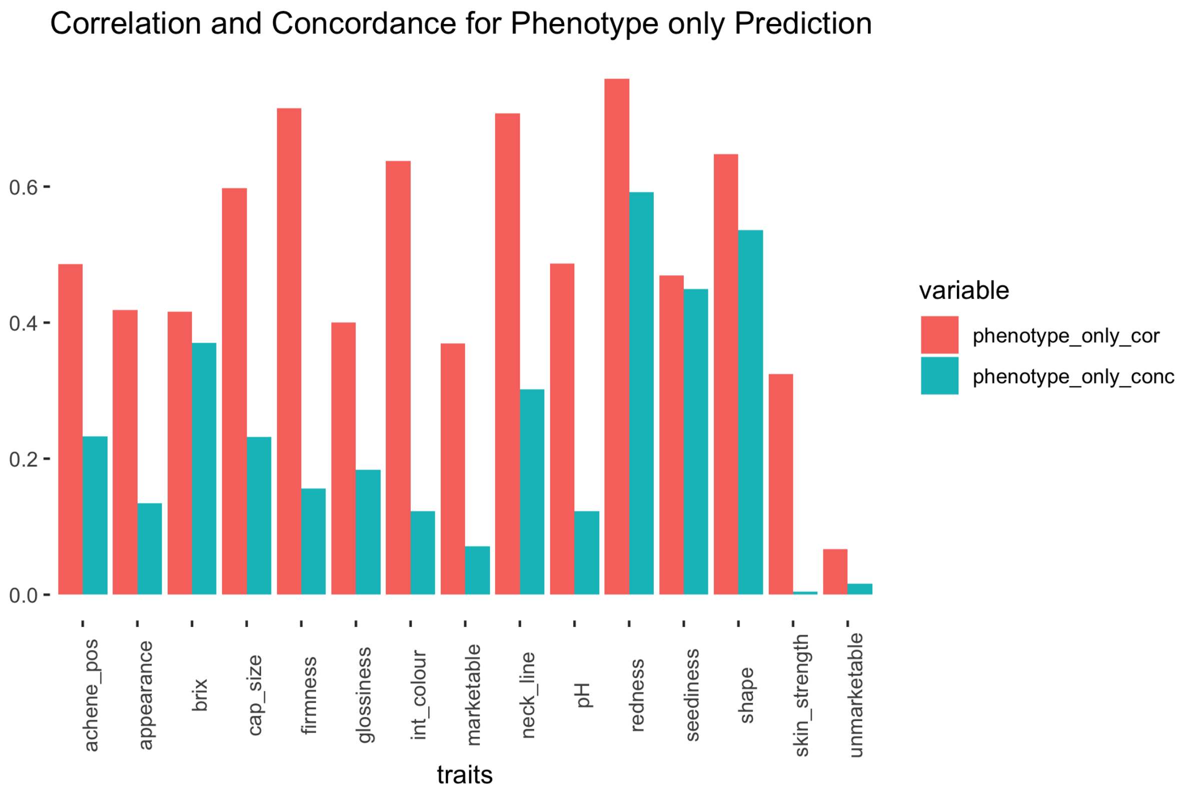


Figure 4.3: Correlations and concordances between phenotype only prediction in the prediction and validation datasets.

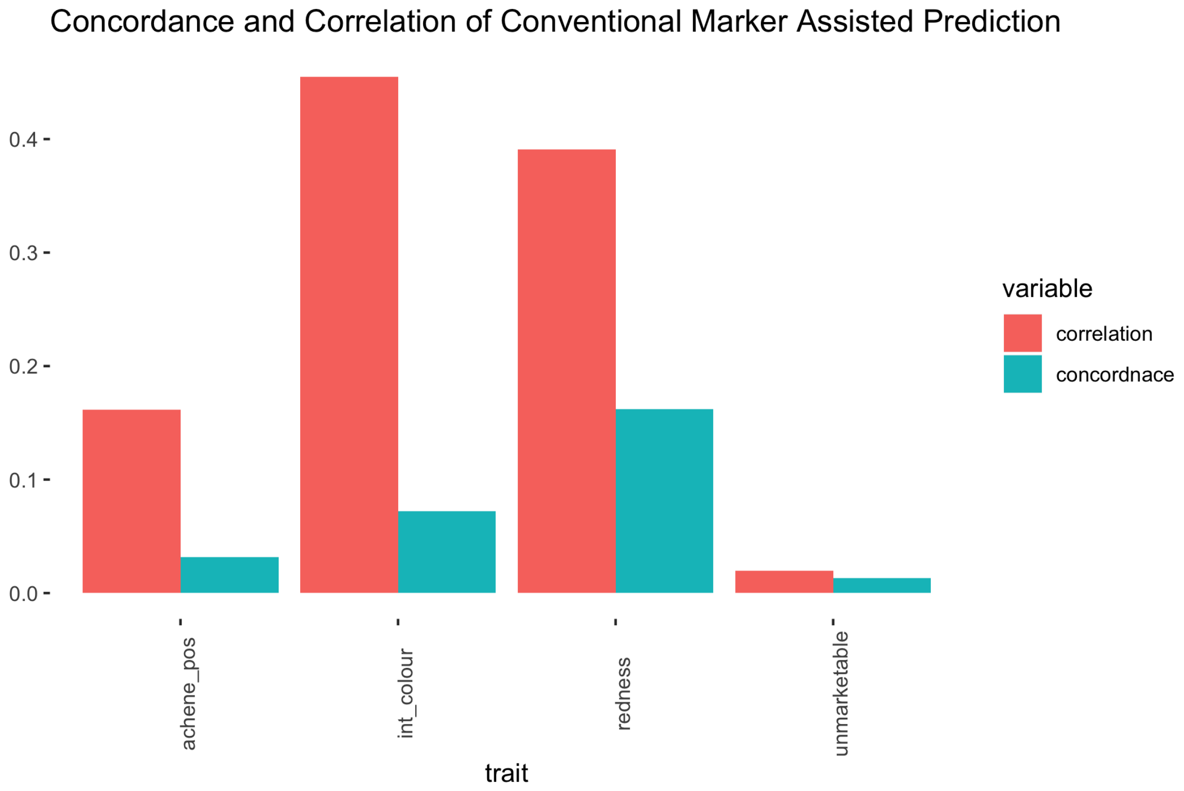


Figure 4.4: Correlations and concordances between tMAP model and validation data

test thresholds of 3.4, 3.5 and 5.1 respectively. The mean genotypic information content for each marker were 0.99, 1.0 and 0.99 respectively, with each marker explaining 17.0%, 15.5% and 21.0% of the variance respectively.

Using the KW test, 2 markers on LG3C at 90cM, 4 markers on LG4C at 64cM and 6 markers on LG7A at 52cM were identified, all for internal colour. The significance of the markers, after controlling for false discovery were between 0.15 and 0.18. Deployment of tMAP generated predictions for plant performance. The highest prediction accuracy was for internal colour (0.45) with the lowest being for unmarketable yield (0.02). Interestingly, internal colour was a trait for which markers were found, consistent with the skewed phenotypic distribution described.

Prediction and concordances of the tMAP model were generally poor (Figure 4.4). Of the studied traits, markers were identified for only four traits. Comparison with a similar genome wide association study from a previous study on the same population [Antanaviciute, 2016] found no markers for the same traits on the same chromosome as identified in this study. More markers were identified in the previous study as multiple testing correction did not appear to have been applied. Inconsistencies in identified markers across different experiments is common in genome wide association studies as markers are only reported when they exceed a critical threshold, resulting in overestimation of the effect size [Xu, 2003]. In order for markers identified in this experiment to be deployed in MAS, validation of their predictions should be performed in other populations and in different

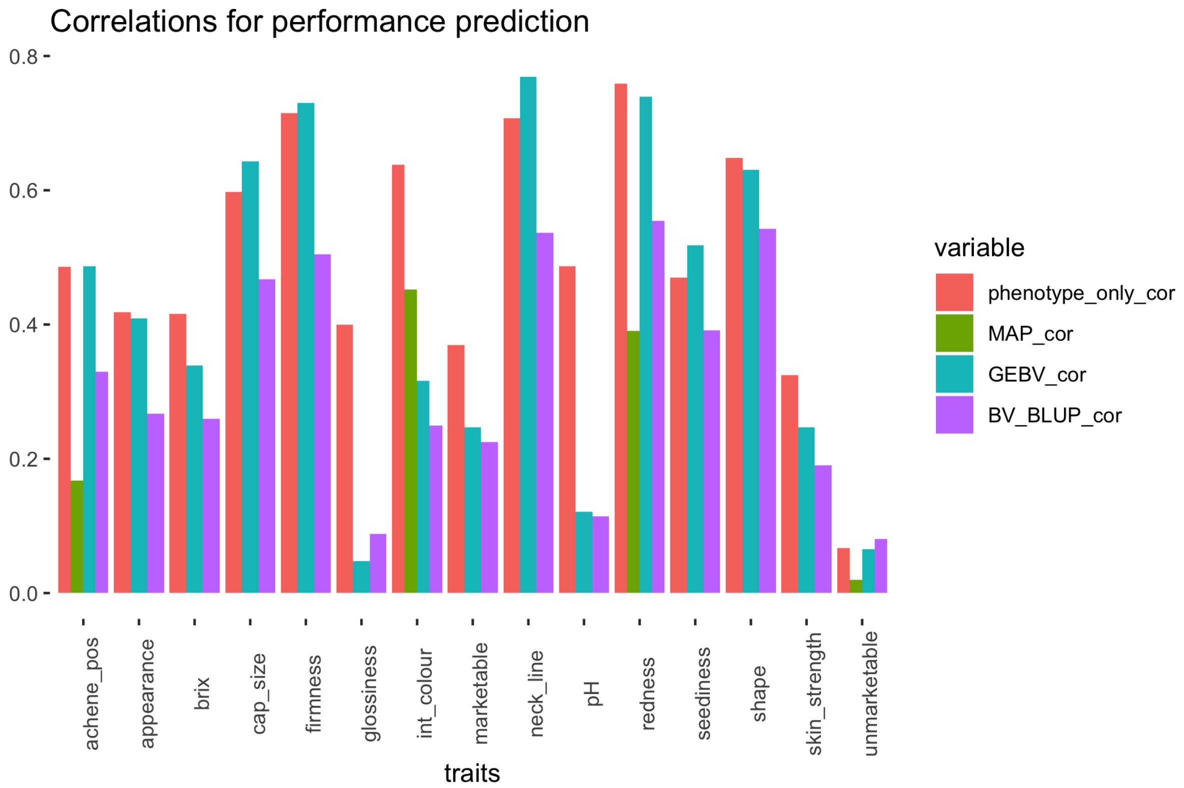


Figure 4.5: Correlations between prediction and validation datasets for phenotype only prediction, tMAP and GBLUP. Two methods of assessing GBLUP was presented, correlations between prediction GEBVs and validation BLUPs (purple) and correlations between prediction GEBVs and validation GEBVs (blue). tMAP correlations (green) were calculated only for achene position, internal colour, redness and unmarketable yield as markers were not identified for other traits.

environments.

4.3.5 Genomic Prediction

Deployment of GBLUP generated GEBVs for all traits (Figure 4.5). High (> 0.6) prediction accuracies could be achieved for cap size, firmness, internal colour, neck line, redness and shape. Glossiness, skin strength, marketable yield and unmarketable yield had low (< 0.4) prediction accuracies. The highest prediction accuracy was for redness (0.74) with the lowest being for unmarketable yield (0.08). The concordances between GEBVs estimated using the prediction and validation datasets were close to zero (data not shown). This is likely due to significant shrinkage, so slight differences in means between validations and prediction datasets would result in lack of concordance. Interestingly, no predictive model performed well for unmarketable yield, perhaps due to its low heritability.

In theory, MAP in this study must be at least as good as phenotype only prediction when available information is utilised optimally. This is because the data included in the phenotype only prediction models include phenotype data only, whereas the MAP models include identical phenotype as well as genotype information. If there is no relationship between marker data and plant performance, then optimal use of MAP ought to be identical to phenotype only prediction; any predictive power that the genotypic information has improves the model.

One possible reason that the tMAP models explored here perform poorer than phenotype only prediction is the introduction of bias. tMAP performs the poorest of all the models in the traits measured. One source of bias in tMAP is the accept/reject nature of marker identification, which assigns an effect to markers that fall above a critical threshold, whilst rejecting effects that fall below. In this approach, markers with small effects, which together may account for a significant proportion of the variation, may not be identified, thus biasing the effect of the markers. One expected effect of this, which was observed in this study, is that markers for some polygenic traits cannot be identified, making tMAP incapable of predicting performance.

In the cases of most traits, prediction accuracies between phenotype only predictions and GEBVs between the training and validation populations are similar. This indicated that there is little additional information that genotype data add to make predictions. The results presented may underestimate the performance of GP in a strawberry breeding population. In a real breeding population, there is potential to leverage a much larger training population including individuals that were genotyped/phenotyped in previous years and other locations because their relationship with the breeding population is known through shared markers.

4.4 Conclusions

4.4.1 Implications for Strawberry Breeding

The biparental population used in this study shows relatively large variances in yield, cap size, achene position, shape, neck line, internal colour and firmness, suggesting there is diversity even in a biparental population. It appears that tMAS is ineffective for selection for most of the traits studied here, with no markers identified for 11 of the 15 traits measured in this study. These traits are likely to be polygenic with multiple QTLs of small effect. However, with the same marker data, GP generates higher selection accuracies and can be performed on all traits. Selection accuracies are high (> 0.6) for cap size, firmness, neck line, redness, and shape, indicating that these traits are amenable to GP. As no additional data are required for GP compared with tMAS for the described experimental conditions, GP is a more efficient use of available genomic data and is preferable for MAS. For most of the traits studied, GP performs approximately as well as phenotype only prediction, indicating that genomic information for these traits offers little improvement in selection accuracy. This observation is consistent with other studies. Comparison of GP to direct selection for plant height and time to anthesis in wheat showed GP had approximately two-thirds times the response [Watson et al., 2019].

GP allows for prediction and thus selection on traits that are difficult to assess phenotypically by leveraging available data from a related training population. For example, yield cannot be reliably assessed in the first phase of breeding as there are typically too few replicates. With GP, selection on yield allows for greater genetic gain in this phase.

The MAP models presented here can be deployed in strawberry breeding, though the applicability of the models to other populations has not been experimentally tested. The GP model has higher accuracy than the tMAP model.

In cases where multiple traits are desired to be simultaneously selected, a selection index may be used. A selection index is a single value calculated from a function of measurements of multiple traits of agronomic importance [Geraldi, 2005]. Correlations between traits suggests that the QTLs controlling these traits are linked or pleiotropic. Due to correlation between traits, it is more efficient to select based on an index rather than individual traits [Céron-Rojas and Crossa, 2018]. Breeding efforts attempting to establish counter relations between these traits are likely to be less successful. The correlations identified between traits here can guide the choices of breeders for the combinations of traits to select for. Most correlations identified here are relatively weak (< 0.5) and do not significantly limit the possibility of selection for or against pairs of traits. The exception is the strong correlation between redness and internal colour, indicating that an externally red but, internally light coloured (or vice versa) strawberry would be a challenging breeding goal.

For evaluation of the efficacy of GP, utilising the prediction dataset to predict the validation dataset may not be optimal due to the effect of year as a systemic error. It may be more appropriate to randomly select training and validation datasets with cross validation. Five-fold cross validation has been performed in a studies of GP in strawberries [Gezan et al., 2017, Osorio et al., 2021] and energy cane [Olatoye et al., 2019], to reduce over fitting. This approach is appropriate for assessing the long term efficacy of GP as a technique for strawberry as the cross-validation removes between individual environmental variation, but this value may not be most useful for breeders. This is because a breeder, in practical deployment of GP, would utilise a training population that (likely) would have been phenotyped in previous years, making environmental variation a significant factor. Moreover, as the typical strawberry variety has a finite lifespan on the market, the long-term prediction accuracies that a cross validation generates may overestimate the breeding value of a given potential novel variety.

For assessment of selection accuracy, it is conventional to use the Pearson's correlation coefficient [González-Recio et al., 2014, Gezan et al., 2017, Zhang et al., 2019], but this value may not be the most informative value for strawberry breeders. Pearson's correlation assumes each element of the prediction and validation datasets is drawn from the same distribution, each measurement of prediction and validation has equal reliability and it does not address the existence of bias. A consistent bias may mislead the breeder in incorrectly rejecting an individual that has a higher GEBV than a check variety, or vice versa. Accuracy can alternatively be measured with regression coefficients, predictive mean squared error and area under the receiver operating characteristic curve [González-Recio et al., 2014].

Strawberry breeders are interested in the absolute value of traits in addition to selecting the best individual from their population. This is because they must ensure that their selections are superior to existing check varieties as growers have the option to select varieties from clonally produced stock. In this study, the CCC is used as a metric of selection accuracy [Lin et al., 2002]. This metric is the product of the Pearson's correlation coefficient and another metric ($0 \leq C_b \leq 1$) that assesses the slope and intercept of the correlation curve [Lin et al., 2014]. This allows the assessment of the absolute value of the prediction method and has the advantage of being easy to compute, with a high value having a simple interpretation. A high CCC implies the prediction and validation data are concordant, lying approximately on a 45 °line with intercept at (0, 0).

4.4.2 Limitations

There are two types of limitations to this study, at the level of the experimental approaches and at the level of the capacity to make useful inferences for strawberry breeders. Phenotyping approaches were conducted by eye on 10 of the traits thus errors may be associated with the rater. Additionally, the prediction and validation data were conducted by different people, making it difficult to estimate the between years rater effect. For example,the concordances for skin strength is low, but the correlation is high. The assessment metric relies on the assessor rubbing the skin of the fruit and checking for breaking of the skin; as assessors were different between the training and validation dataset, a systemic difference is likely the cause.

A bi-parental mapping population was chosen for this experiment as is standard for marker discovery. Additionally, the population chosen has allowed leverage of data from 2013 - 2015 to increase the power of the analysis and one parent, 'Redgauntlet' is in process of sequencing and assembly to serve as a reference genome for strawberry. It is recognised that a real breeding population is unlikely to comprise of a bi-parental cross, but for comparison of techniques' accuracies, the same dataset was analysed. The reduced variation from a bi-parental cross underestimates variation likely to be present in a real breeding population, but high number of replications of offspring increases the number of recombinations, allowing for more precise estimates of marker effects.

A real breeding commercial strawberry program would utilise training data from a similar population to the breeding population. As the training population in this study is genetically identical to the validation population and thus the selection accuracies for all models studied may be higher than a real breeding population. However, when the training population is different from the breeding population, the linear model described in this study for phenotype only prediction cannot be deployed.

Measured correlations in this study were assumed to be due to biological effects, but may be due to an artefact of the measurement process. For instance, the correlation between neck line and shape may be visual measurement of the same aspect of the fruit rather than a biological (pleiotropic) effect between two distinct traits. Controlling for this kind of effect in a breeding program may be done through appropriate weightings of a selection index. Epistasy prevents assessment of some traits. For example, when the marketable yield is low or zero for a plant, measurements of other fruit characteristics are not possible. This reduction of reliability of data is non-random, which may lead to unknown biases.

For some of the traits measured, the scale ranged from 1 - 3 or 1 - 5, which may not be sufficiently wide for analysis using parametric approaches. Treatment of scales as ordinal is possible for GP and BLUP approaches. While there is no rule regarding the scale that qualify a measurement as ordinal or continuous, there was no observed difference in accuracy when a 9 point scale for disease resistance in strawberry was treated as ordinal or continuous data when deploying GP [Pincot et al., 2020].

The performance of GP is comparable to phenotype only selection in many traits, indicating that the addition of genotypic information may not generate additional information. In the case of glossiness, internal colour and pH, however, it appears that GP performs significantly worse than phenotype only prediction. There may be different reasons for the difference in predictive power between marker assisted prediction and phenotype only prediction. At the level of the data, slightly different datasets were included in each model. Due to DNA quality issues, some individuals in the phenotype only modelling population were excluded from genotypic analysis. This may be due to non-additive effects that GBLUP does not capture. Epigenetic effects may be a mechanism for non-additive (and undetectable by genotyping) effects in strawberry. For example, ripening-induced hypomethylation (and thus colour) in strawberries has been observed to be caused by down regulation of RNA-directed DNA methylation genes [Cheng et al., 2018].

4.4.3 Future Work

In principle, GP can replace all stages of the breeding cycle after crosses are made for strawberry. However, significant effects of GxE interactions have not been measured, which hinders GP for traits, resulting in low heritabilities. It may not be ever possible to deploy GP to replace latter stages of breeding trials (where large numbers of replicates of a potential new cultivar are assessed in different environments) as the training population sizes to generate equivalent power are likely to be infeasible. Moreover, epistatic interactions may not be measurable due to the exponentially increasing population size needed to measure higher numbers of interacting markers [Le Rouzic, 2014]. Finally, epigenetic effects of traits are not measured by genotyping platforms, and it is unknown the relative contribution of genetic to epigenetic effects in inheritable traits. Thus it is envisaged the first implementation of GP in commercial breeding programmes replaces the first stage of selection to reduce the burden of latter trials with larger populations.

There are currently no known experimental validations of a speed breeding programme for strawberry utilising GP to make selections before plant maturation. However, simulations of GS and speed breeding for wheat suggests several-fold increases in genetic gain per unit time are possible [Liu et al., 2019, Bhatta et al., 2021]. An experimental study in wheat generated an F_5 generation in 15 months using GS, with the authors noting that more optimal utilisation of resources could shorten the time required to 12 months. Comparison of GP to direct selection for plant height and time to anthesis showed GP had approximately two-thirds times the response, but perhaps 4-5 times the speed, suggesting significant expected gain [Watson et al., 2019]. The potential to decouple the phenotyping and selection phases of the breeding process allows for significant increase in the number of breeding cycles per unit time. The observation that the predictive power of most GP is comparable to phenotype only prediction for most traits in strawberry indicates that accuracy when basing selection on genotypic information is unlikely to limit the efficacy of speed breeding.

When multiple traits (potentially of different import) are to be selected for, it is more efficient to utilise a selection index [Geraldi, 2005]. The weighting of the index is dependent on the goals of the breeding programme. The response term of the Breeder's equation can be replaced by a selection index without change to the equation [Cobb et al., 2019]. Multi-trait GP deploying an index is more effective in increasing gain compared to single trait selection in both simulated and real data [Ceron-Rojas et al., 2015]. Where multiple traits are measured and to be selected on in the same population, as is the case for real breeding populations and the traits measured in this study, it would be more efficient to predict based on a selection index. In order to generate a selection index, different traits need to be differentially weighted, which is dependent on the goals of a particular breeding programme. Studies have been conducted quantifying financial value of strawberry traits, which could be a basis for weighting of the selection index [Gallardo et al., 2014]. In this study, 15 fruit quality traits encompassing the majority of traits of interest to strawberry breeders [Mathey et al., 2013] were measured and analysed within the same population, making it potentially a useful basis for development of selection indices. Moreover, the calculated correlations between phenotypes will be useful in informing breeders regarding selection upon correlated traits.

To further optimise selection in the context of a breeding programme, efficient allocation of resources to different stages of selection should be considered. For instance, decision support tools simultaneously considering the major costs associated with breeding programmes have been created [Edge-Garza et al., 2015]. Effective simulation tools for GP have been developed specifically for polyploids, aiding design of breeding programmes [Zingaretti et al., 2019]. However, there are no known models that calculate cost efficiency for strawberry breeding programmes with savings and increases in genetic gain per unit time associated with speed breeding deploying GP.

Chapter 5

Conclusion

5.1 Perspectives

With the global population projected to reach nearly 10 billion in 2050, changes in the composition of human calorie consumption is needed to ensure sufficient food for all, including significant transition to plant based calorie intake [Berners-Lee et al., 2018]. GS has the potential to increase the rate of genetic gain in breeding efforts and thus contribute towards feeding the projected population growth. Strawberries are a valuable commodity that are a source of nutrients, notably vitamin C and manganese as well as associated with protective effects for cancer and cardiovascular disease. Production and consumption in the UK, and the rest of the world are on an upwards trend.

Commercial strawberry breeding must generate money from its breeding efforts, and while the amounts that are generated depend on the nature of the breeding program, it is also dependant on the quality of their output, novel strawberry varieties. In this thesis, research is presented in three areas to improve deployment of GP in strawberry breeding and thus potentially improve the quality of novel strawberry varieties. The automated high-throughput 3D phenotyping platform increases the precision and reliability of measurement of seven strawberry fruit quality traits, which should allow for more accurate GP and reduction of labour costs associated phenotyping. Use of objective measurement scales also allows for easier interpretation of results. The rational design of amplicon sets for GP attempts to integrate parameters likely to be informative for GP in an open, scalable, genotyping system. This reduces cost by measuring only the most informative markers and allows scaling of genotyping effort dependant on the resources of the breeding program. The deployment of predictive models in a biparental strawberry population serves as experimental validation of the efficacy of GP in strawberry breeding compared to MAP and phenotype only prediction approaches, offering models for breeders to utilise.

The most impactful of the three areas for deployment of GP in commercial breeding programmes is likely to be the reduction in cost of genotyping. Cost of strawberry breeding in one program was $\sim$ \$40000 - \$50000 dollars per year for seedling trials for $\sim$ 3500 seedling crosses over three years [Wannemuehler et al., 2020]. Even assuming that these costs could be eliminated completely through GP, the deployment of the \$50 35K SNP array [Gezan et al., 2017] for genotyping would still be more costly. Greater costs can be justified by increases in selection accuracy leading to greater genetic gains, but increases in accuracy of GP over conventional prediction of plant performance based on phenotypic data is minimal. The second most important area is the experimental validation of GP models in strawberry. Currently, published GP models in strawberries exist for early marketable yield, total marketable yield, unmarketable proportion of fruit, soluble solid content [Osorio et al., 2021, Gezan et al., 2017] and verticillium resistance [Pincot et al., 2020], which encompass only a small proportion of traits of interest to strawberry breeders. A high throughput automated phenotyping platform serves to increase the precision and accuracy of the data gathered, leading to increased selection accuracy. It also would likely decrease the cost of labour in phenotyping, making both conventional breeding and GP cheaper.

5.2 Towards Genomic Selection in Commercial Strawberry Breeding

It appears from simulation studies that genotyping may not be cost effective in saving costs associated with maintaining plants in the field, even assuming perfect Mendelian inheritance and accuracy, for MAS. This is due to the relatively low cost of maintaining strawberries and the high labour and reagent costs associated with performing genotyping [Wannemuehler et al., 2020, Edge-Garza et al., 2015]. In the case of GS, a large marker panel is required, of which most markers have small or zero effect; indeed, in the case of some models such as BayesB, a prior with some proportion of markers with zero effect is assumed [Meuwissen et al., 2001]. The cost of genotyping an individual for GS is thus higher than the cost of genotyping for tMAS. Attempts to rationally design a marker set multiplexing the large number of individuals in a breeding population to reduce costs is challenging due to poor predictive programs for multiplex primer design. Current published experimental validations of GP in strawberry utilise one of the available SNP arrays [Gezan et al., 2017, Pincot et al., 2020, Osorio et al., 2021].

The main benefit of GP in strawberries, is expected in deployment of speed breeding [Bhatta et al., 2021]. For conventional selection, assessment of fruit quality traits requires plant maturation, which requires approximately a year of growth, whereas genotyping and thus GP can be performed at the seedling stage from a single leaf. The increase in genetic gain per unit time is inversely proportional to the duration of a selection cycle as given by the breeder's equation [Voss-Fels et al., 2019]:

$$R = \frac{H^2 S}{t} \tag{5.1}$$

where R is the response to selection, H^2 is the heritability, S is the selection intensity and t is the time for a selection cycle. Reducing the duration of a breeding cycle is the most desirable factor to increase genetic gain; heritability is a function of the population and varies only between $[0, 1]$, limiting improvements possible and increasing selection intensity reduces genetic diversity, which reduces response to selection in the future. Halving the duration of the selection cycle doubles the genetic gain per unit time and selection at the seedling stage may shorten the duration of a selection cycle, potentially increasing genetic gain per unit time (Figure 5.1). Under speed breeding, phenotype and genotyping would be decoupled, with a training population undergoing successive rounds of GP to increase accuracy of prediction for the breeding population. The breeding

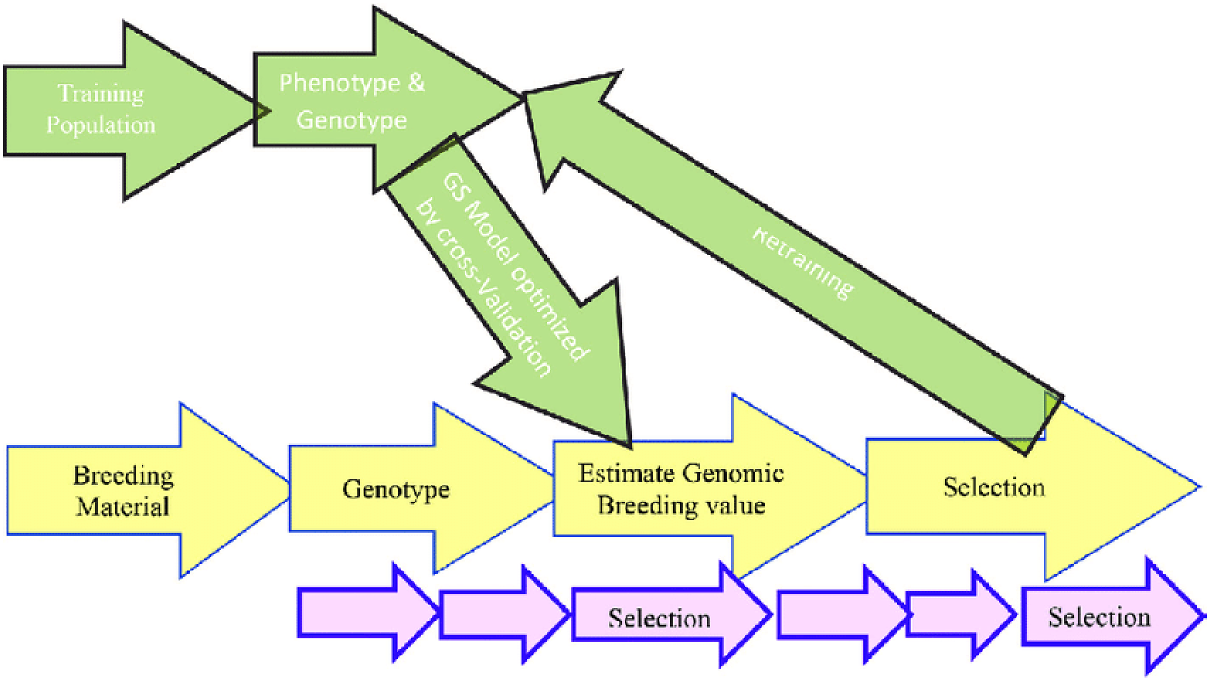


Figure 5.1: Schematic of GS where training and breeding is decoupled. Selection in the breeding population is made using the GP model, which is then evaluated and used in training future models of GP [Shamshad and Sharma, 2018].

population would be selected upon through genotyping and GP, with results being used to improve the model [Shamshad and Sharma, 2018].

In order for a commercial breeding program to benefit from GS, selections at the seedling stage are required. This requires almost all traits of agronomic importance to be amenable to GP as if some traits cannot be predicted on genomic data, maturation of the plant for assessment of those traits is needed. Depending on the breeding programme, over 40 traits may be of interest to breeders. Fortunately, there appears to be no experimental evidence of a trait that cannot be predicted using GP, though some traits have low accuracies of prediction. Note that MAS would not be suitable for a similar speed breeding scheme by genotyping and selection at the seedling stage as many traits do not have large effect markers that can be identified. Traits of interest in strawberry breeding can be split into fruit quality and plant habit traits [Antanaviciute, 2016]. The GP model presented in this thesis encompasses almost all the traits of interest to strawberry breeders [Mathey et al., 2013], potentially allowing for selections to be made prior to fruit development. This would reduce the selection cycle by a few months if plant habit traits can be adequately assessed before fruit maturation.

5.3 Future Research

If GP allows for increased genetic gain per unit time through reduction of the duration of the breeding cycle, then adoption of GP is a binary choice for strawberry breeders. Gradual introduction of GP for some traits cannot be performed (as can be envisaged for MAS) as shortening of breeding cycle cannot be achieved in these cases. Thus, experimental evidence for efficacy of GP in strawberry, as well as availability of datasets must be convincing before commercial adoption. Research focus should be targeted towards demonstration of speed breeding by ensuring that almost all traits of interest to breeders are experimentally modelled with GP.

Selection of models typically has a smaller effect on selection accuracy than number of markers or trait to be predicted, and is not expected to be the primary mechanism by which GP improves selection in strawberries. Moreover, if the duration of selection cycles are reduced to months (from planting of breeding population to selection), then the days or weeks of computation required for the more computationally intensive models of GP (such as BayesA) may become significant in speed breeding. Models for cost efficiency of MAP should be extended to account for the potential to remove the growth time for the first stage selections under speed breeding and GP.

Deployment of GP requires most traits of strawberry to be predictable by GP. However, due to correlation between traits, selection on individual traits may be inefficient; selection indices allow for appropriate weighing of traits that may be correlated. Correlation between traits is primarily due to LD between controlling genes and pleiotropy [Pedruzzi and Rouzine, 2019, Chebib and Guillaume, 2019]. Effort should be made to ensure multiple traits of agronomic importance are measured in the same strawberry population so these correlations can be quantified to inform construction of selection indices. Establishment of selection indices are dependent on the goals of the breeding program, but typically combination of traits of agronomic import are non-linear. Often there are thresholds that must be met for some characteristics to ensure that a potential novel cultivar exceeds performance of some check species, making unpredictable traits particularly disadvantageous for GP. If there is a trait that cannot be selected for using GP, or has a low accuracy compared to phenotypic assessment methods, loss of prediction for it must be compensated for by increases in genetic gain from speed breeding.

Chapter 6

Bibliography

[Afrin et al., 2016] Afrin, S., Gasparrini, M., Forbes-Hernandez, T. Y., Reboredo-Rodriguez, P., Mezzetti, B., Varela-López, A., Giampieri, F., and Battino, M. (2016). Promising Health Benefits of the Strawberry: A Focus on Clinical Studies. *Journal of Agricultural and Food Chemistry*, 64(22):4435–4449.

[Akond et al., 2019] Akond, Z., Alam, M. J., Hasan, M. N., Uddin, M. S., Alam, M., and Mollah, N. H. (2019). A Comparison on Some Interval Mapping Approaches for QTL Detection. *Bioinformation*, 15(2):90–94.

[Alenyà et al., 2011] Alenyà, G., Dellen, B., and Torras, C. (2011). 3D modelling of leaves from color and ToF data for robotized plant measuring. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 3408–3414.

[Altieri et al., 2010] Altieri, R., Esposito, A., and Baruzzi, G. (2010). Use of olive mill waste mix as peat surrogate in substrate for strawberry soilless cultivation. *International Biodeterioration and Biodegradation*, 64(7):670–675.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). BLAST_article.pdf.

[Alvarez-Suarez et al., 2014] Alvarez-Suarez, J. M., Giampieri, F., Tulipani, S., Casoli, T., Di Stefano, G., González-Paramás, A. M., Santos-Buelga, C., Busco, F., Quiles, J. L., Cordero, M. D., Bompadre, S., Mezzetti, B., and Battino, M. (2014). One-month strawberry-rich anthocyanin supplementation ameliorates cardiovascular risk, oxidative stress markers and platelet activation in humans. *Journal of Nutritional Biochemistry*, 25(3):289–294.

[Amani, 2014] Amani, R. (2014). Flavonoid-rich beverage effects on lipid profile and blood pressure in diabetic patients. *World Journal of Diabetes*, 5(6):962–968.

[Amatori et al., 2016] Amatori, S., Mazzoni, L., Alvarez-Suarez, J. M., Giampieri, F., Gasparrini, M., Forbes-Hernandez, T. Y., Afrin, S., Errico Provenzano, A., Persico, G., Mezzetti, B., Amici, A., Fanelli, M., and Battino, M. (2016). Polyphenol-rich strawberry extract (PRSE) shows in vitro and in vivo biological activity against invasive breast cancer cells. *Scientific Reports*, 6(1):1–13.

[Anciro et al., 2018] Anciro, A., Mangandi, J., Verma, S., Peres, N., Whitaker, V. M., and Lee, S. (2018). FaRCg1: a quantitative trait locus conferring resistance to Colletotrichum crown rot caused by *Colletotrichum gloeosporioides* in octoploid strawberry. *Theoretical and Applied Genetics*, 131(10):2167–2177.

[Antanaviciute, 2016] Antanaviciute, L. (2016). *Genetic mapping and phenotyping plant characteristics , fruit quality and disease resistance traits in octoploid strawberry (Fragaria x ananassa).* PhD thesis, University of Reading.

[Asalf et al., 2014] Asalf, B., Gadoury, D. M., Tronsmo, A. M., Seem, R. C., Dobson, A., Peres, N. A., and Stensvand, A. (2014). Ontogenic resistance of leaves and fruit, and how leaf folding influences the distribution of powdery mildew on strawberry plants colonized by *Podosphaera aphanis*. *Phytopathology*, 104(9):954–963.

[Ashley et al., 2003] Ashley, M. V., Wilk, J. A., Styan, S. M., Craft, K. J., Jones, K. L., Feldheim, K. A., Lewers, K. S., and Ashman, T. L. (2003). High variability and disomic segregation of microsatellites in the octoploid *Fragaria virginiana* Mill. (Rosaceae). *Theoretical and Applied Genetics*, 107(7):1201–1207.

[Ayesha et al., 2011] Ayesha, R., Fatima, N., Ruqayya, M., Faheem, H., Qureshi, K. M., Hafiz, I. A., Khan, K. S., Ali, U., and Kamal, A. (2011). Influence of different growth media on the fruit quality and reproductive growth parameters of strawberry (*Fragaria ananassa*). *Journal of Medicinal Plant Research*, 5(26):6224–6232.

[Badenes and Byrne, 2012] Badenes, M. L. and Byrne, D. H. (2012). Chapter 9: Strawberry. In *Fruit Breeding*, pages 305–325.

[Bae and Kim, 2008] Bae, Y. J. and Kim, M. H. (2008). Manganese supplementation improves mineral density of the spine and femur and serum osteocalcin in rats. *Biological Trace Element Research*, 124(1):28–34.

[Bartholomé et al., 2016] Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., and Bouffier, L. (2016). Performance of genomic prediction within and across generations in maritime pine. *BMC Genomics*, 17(1):1–14.

[Bassil et al., 2015] Bassil, N. V., Davis, T. M., Zhang, H., Ficklin, S., Mittmann, M., Webster, T., Mahoney, L., Wood, D., Alperin, E. S., Rosyara, U. R., Putten, H. K.-v., Monfort, A., Sargent, D. J., Amaya, I., Denoyes, B., Bianco, L., van Dijk, T., Pirani, A., Iezzoni, A., Main, D., Peace, C., Yang, Y., Whitaker, V., Verma, S., Bellon, L., Brew, F., Herrera, R., and van de Weg, E. (2015). Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria ananassa*. *BMC Genomics*, 16(1):155.

[Bates et al., 2015] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

[Beavis, 1994] Beavis, W. D. (1994). The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Proceedings of the Forty-ninth Annual Corn & Sorghum Industry Research Conference*, pages 250–266.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Bernard et al., 2014] Bernard, H., Faber, M., Wilking, H., Haller, S., Höhle, M., Schielke, A., Ducomble, T., Siffczyk, C., Merbecks, S., Fricke, G., Hamouda, O., Stark, K., Werber, D., Askar, M., Gunsenheimer-Bartmeyer, B., Behnke, S., Breidenbach, J., Fiebig, L., Gilsdorf, A., Greutélaers, B., Hecht, J., Hermes, J., Jentsch, F., Milde-Busch, A., Preußel, K., Prohl, M., Remschmidt, C., Ritter, S., Rosner, B., Santos-Hövener, C., Schaade, L., Schütze, M., Tolksdorf, K., Wetzstein, M., Bätzing-Feigenbaum, J., Oppermann, H., and Popp, A. (2014). Large multistate outbreak of norovirus gastroenteritis associated with frozen strawberries, Germany, 2012. *Eurosurveillance*, 19(8).

[Berners-Lee et al., 2018] Berners-Lee, M., Kennelly, C., Watson, R., and Hewitt, C. N. (2018). Current global food production is sufficient to meet human nutritional needs in 2050 provided there is radical societal adaptation. *Elementa Science of the Anthrropocene*, 6(52):1–14.

[Bhatta et al., 2021] Bhatta, M., Sandro, P., Smith, M. R., Delaney, O., Voss-Fels, K. P., Gutierrez, L., and Hickey, L. T. (2021). Need for speed: manipulating plant growth to accelerate breeding cycles. *Current Opinion in Plant Biology*, 60(January):101986.

[Black et al., 2002] Black, B. L., Enns, J. M., and Hokanson, S. C. (2002). A comparison of temperate-climate strawberry production systems using eastern genotypes. *HortTechnology*, 12(4):670–675.

[Blasco et al., 2003] Blasco, J., Aleixos, N., and Moltó, E. (2003). Machine vision system for automatic quality grading of fruit. *Biosystems Engineering*, 85(4):415–423.

[Blasco et al., 2007] Blasco, J., Aleixos, N., and Moltó, E. (2007). Computer vision detection of peel defects in citrus by means of a region oriented segmentation algorithm. *Journal of Food Engineering*, 81(3):535–543.

[Boyer et al., 2016] Boyer, L. R., Feng, W., Gulbis, N., Hajdu, K., Harrison, R. J., Jeffries, P., and Xu, X. (2016). The use of arbuscular mycorrhizal fungi to improve strawberry production in coir substrate. *Frontiers in Plant Science*, 7(AUG2016):1–9.

[Boyera et al., 1998] Boyera, N., Galey, I., and Bernard, B. A. (1998). Effect of vitamin C and its derivatives on collagen synthesis and cross-linking by normal human fibroblasts. *International Journal of Cosmetic Science*, 20(3):151–158.

[Breseghello and Coelho, 2013] Breseghello, F. and Coelho, A. S. G. (2013). Traditional and modern plant breeding methods with examples in rice (*Oryza sativa* L.). *Journal of Agricultural and Food Chemistry*, 61(35):8277–8286.

[Broman, 2003] Broman, K. W. (2003). Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics*, 163(3):1169–1175.

[Brown et al., 2017] Brown, S. S., Chen, Y. W., Wang, M., Clipson, A., Ochoa, E., and Du, M. Q. (2017). PrimerPooler: Automated primer pooling to prepare library for targeted sequencing. *Biology Methods and Protocols*, 2(1):1–10.

[Buckler et al., 2016] Buckler, E. S., Ilut, D. C., Wang, X., Kretzschmar, T., Gore, M. A., and Mitchell, S. E. (2016). rAmpSeq: Using repetitive sequences for robust genotyping. *bioRxiv, p.096628.*

[Campbell et al., 2014] Campbell, N. R., Harmon, S., and Narum, S. R. (2014). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, 15(4):855–867.

[Cao et al., 1998] Cao, G., Russell, R. M., Lischner, N., and Prior, R. L. (1998). Serum antioxidant capacity is increased by consumption of strawberries, spinach, red wine or vitamin C in elderly women. *The Journal of Nutrition*, 128(12):2383–2390.

[Cassidy et al., 2013] Cassidy, A., Mukamal, K. J., Liu, L., Franz, M., Eliassen, A. H., and Rimm, E. B. (2013). High Anthocyanin Intake Is Associated With a Reduced Risk of Myocardial Infarction in Young and Middle-Aged Women. *Circulation*, 127(2):188–196.

[Cericola et al., 2017] Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., and Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. a case of study in advanced wheat breeding lines. *PLoS ONE*, 12(1):1–20.

[Céron-Rojas and Crossa, 2018] Céron-Rojas, J. J. and Crossa, J. (2018). *Linear Selection Indices in Modern Plant Breeding Foreword by Daniel Gianola.*

[Ceron-Rojas et al., 2015] Ceron-Rojas, J. J., Crossa, J., Arief, V. N., Basford, K., Rutkoski, J., Jarquín, D., Alvarado, G., Beyene, Y., Semagn, K., and DeLacy, I. (2015). A genomic selection index applied to simulated and real data. *G3: Genes, Genomes, Genetics*, 5(10):2155–2164.

[Chalavi et al., 2003] Chalavi, V., Tabaeizadeh, Z., and Thibodeau, P. (2003). Enhanced resistance to *Verticillium dahliae* in transgenic strawberry plants expressing a *Lycopersicon chilense* chitinase gene. *Journal of the American Society for Horticultural Science*, 128(5):747–753.

[Chalidabhongse et al., 2006] Chalidabhongse, T., Yimyam, P., and Sirisomboon, P. (2006). 2D/3D vision-based mango's feature extraction and sorting. In *9th International Conference on Control, Automation, Robotics and Vision, 2006, ICARCV '06.*

[Chambers et al., 2014] Chambers, A. H., Pillet, J., Plotto, A., Bai, J., Whitaker, V. M., and Folta, K. M. (2014). Identification of a strawberry flavor gene candidate using an integrated genetic-genomic-analytical chemistry approach. *BMC Genomics*, 15(1):1–15.

[Chandler et al., 2012] Chandler, C. K., Folta, K., Dale, A., Whitaker, V. M., and Herrington, M. (2012). Strawberry. In *Fruit Breeding*, pages 305–325. Springer US, Boston, MA.

[Chang et al., 1988] Chang, C., Bowman, J. L., DeJohn, A. W., Lander, E. S., and Meyerowitz, E. M. (1988). Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, 85(18):6856–60.

[Chebib and Guillaume, 2019] Chebib, J. and Guillaume, F. (2019). Pleiotropy or linkage? Their relative contributions to the genetic correlation of quantitative traits and detection by multi-trait GWA studies. *bioRxiv*.

[Chen et al., 2018] Chen, P., Bornhorst, J., and Aschner, M. (2018). Manganese metabolism in humans. *Frontiers in Bioscience*, 23(9):1655–1679.

[Chéné et al., 2012] Chéné, Y., Rousseau, D., Lucidarme, P., Bertheloot, J., Caffier, V., Morel, P., Belin, É., and Chapeau-Blondeau, F. (2012). On the use of depth camera for 3D phenotyping of entire plants. *Computers and Electronics in Agriculture*, 82:122–127.

[Cheng et al., 2018] Cheng, J., Niu, Q., Zhang, B., Chen, K., Yang, R., Zhu, J. K., Zhang, Y., and Lang, Z. (2018). Downregulation of RdDM during strawberry fruit ripening. *Genome Biology*, 19(1):1–14.

[Choi, 2015] Choi, J. W. (2015). *Estimating Market Equilibrium Values of Fruit Attributes for Apple and Strawberry Using Choice Experiments with Consumers and Producers*. PhD thesis.

[Christenhusz and Byng, 2016] Christenhusz, M. J. M. and Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3):201–217.

[Churchill and Doerge, 1994] Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971.

[Cignoni et al., 2008] Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., and Ranzuglia, G. (2008). MeshLab: an Open-Source Mesh Processing Tool. *Sixth Eurographics Italian Chapter Conference*, pages 129–136.

[Cobb et al., 2019] Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., Hagen, T., Quinn, M., and Ng, E. H. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theoretical and Applied Genetics*, 132(3):627–645.

[Cockerton et al., 2019] Cockerton, H. M., Li, B., Vickerstaff, R. J., Eyre, C. A., Sargent, D. J., Armitage, A. D., Marina-Montes, C., Garcia-Cruz, A., Passey, A. J., Simpson, D. W., and Harrison, R. J. (2019). Identifying *Verticillium dahliae* resistance in strawberry through disease screening of multiple populations and image based phenotyping. *Frontiers in Plant Science*, 10(1):1–17.

[Cockerton et al., 2018] Cockerton, H. M., Vickerstaff, R. J., Karlström, A., Wilson, F., Sobczyk, M., He, J. Q., Sargent, D. J., Passey, A. J., McLeary, K. J., Pakozdi, K., Harrison, N., Lumbreras-Martinez, M., Antanaviciute, L., Simpson, D. W., and Harrison, R. J. (2018). Identification of powdery mildew resistance QTL in strawberry (*Fragaria ananassa*). *Theoretical and Applied Genetics*, 131(1):1–13.

[Collard et al., 2005] Collard, B. C., Jahufer, M. Z., Brouwer, J. B., and Pang, E. C. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142(1-2):169–196.

[Colquhoun et al., 2012] Colquhoun, T. A., Levin, L. A., Moskowitz, H. R., Whitaker, V. M., Clark, D. G., and Folta, K. M. (2012). Framing the perfect strawberry: An exercise in consumer-assisted selection of fruit crops. *Journal of Berry Research*, 2(1):45–61.

[Costa et al., 2009] Costa, C., Menesatti, P., Paglia, G., Pallottino, F., Aguzzi, J., Rimatori, V., Russo, G., Recupero, S., and Recupero, G. R. (2009). Quantitative evaluation of Tarocco sweet orange fruit shape using optoelectronic elliptic Fourier based analysis. *Postharvest Biology and Technology*, 54(1):38–47.

[Crooks and Soule, 1999] Crooks, J. A. and Soule, M. E. (1999). Lag times in population explosions of invasive species: Causes and implications. In *INVASIVE SPECIES AND BIODIVERSITY MANAGEMENT*, pages 103–125.

[Crossa et al., 2017] Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., and Varshney, R. K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, 22(11):961–975.

[Cupec et al., 2011] Cupec, R., Nyarko, E., and Filko, D. (2011). Fast 2.5D Mesh Segmentation to Approximately Convex Surfaces. *5th European Conference on Mobile Robots*, pages 3–8.

[Dadwal and Banga, 2012] Dadwal, M. and Banga, V. K. (2012). Color Image Segmentation for Fruit Ripeness Detection : A Review. *2do International Conference on Electrical, Electronics and Civil Engineering*, pages 190–193.

[Darrier et al., 2019] Darrier, B., Russell, J., Milner, S. G., Hedley, P. E., Shaw, P. D., Macaulay, M., Ramsay, L. D., Halpin, C., Mascher, M., Fleury, D. L., Langridge, P., Stein, N., and Waugh, R. (2019). A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. *Frontiers in Plant Science*, 10(April):1–14.

[Darrow, 1966] Darrow, G. M. (1966). *The Strawberry.* New England Institute for Medical Research.

[Darwish et al., 2015] Darwish, O., Shahan, R., Liu, Z., Slovin, J. P., and Alkharouf, N. W. (2015). Re-annotation of the woodland strawberry (*Fragaria vesca*) genome. *BMC Genomics*, 16(1):29.

[Darwish et al., 2013] Darwish, O., Slovin, J. P., Kang, C., Hollender, C. A., Geretz, A., Houston, S., Liu, Z., and Alkharouf, N. W. (2013). SGR: An online genomic resource for the woodland strawberry. *BMC Plant Biology*, 13(1).

[De Tender et al., 2016] De Tender, C. A., Debode, J., Vandecasteele, B., D'Hose, T., Cremelie, P., Haegeman, A., Ruttink, T., Dawyndt, P., and Maes, M. (2016). Biological, physicochemical and plant health responses in lettuce and strawberry in soil or peat amended with biochar. *Applied Soil Ecology*, 107:1–12.

[DEFRA, 2018] DEFRA (2018). Wholesale fruit and vegetable prices, weekly average.

[Denoyes-Rothan, 1997] Denoyes-Rothan, B. (1997). Inheritance of resistance to *Colletotrichum acutatum* in strawberry (*Fragaria x ananassa*). *Acta Horticulturae*, 439(4):809–814.

[Deschamps et al., 2012] Deschamps, S., Llaca, V., and May, G. D. (2012). Genotyping-by-sequencing in plants. *Biology*, 1(3):460–483.

[Devore et al., 2013] Devore, E. E., Kang, J. H., Breteler, M. M. B., and Grodstein, F. (2013). Dietary intake of berries and flavonoids in relation to cognitive decline. *Annals of Neurology*, 72(1):135–143.

[Dimeas et al., 2014] Dimeas, F., Sako, D. V., Moulianitis, V. C., and Aspragathos, N. A. (2014). Design and fuzzy control of a robotic gripper for efficient strawberry harvesting. *Robotica*, 33(5):1085–1098.

[DiMeglio et al., 2014] DiMeglio, L. M., Staudt, G., Yu, H., and Davis, T. M. (2014). A phylogenetic analysis of the genus *Fragaria* (strawberry) using intron-containing sequence from the *ADH-1* gene. *PLoS ONE*, 9(7):1–12.

[Ding et al., 2004] Ding, S., Mannan, M. A., and Poo, A. N. (2004). Oriented bounding box and octree based global interference detection in 5-axis machining of free-form surfaces. *Computer-Aided Design*, 36(13):1281–1294.

[Dixon and Brereton, 2009] Dixon, S. J. and Brereton, R. G. (2009). Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on. *Chemometrics and Intelligent Laboratory Systems*, 95(1):1–17.

[Donkor et al., 2014] Donkor, E., Dayie, N., and Adiku, T. (2014). Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). *Journal of Bioinformatics and Sequence Analysis*, 6(1):1–6.

[Duarte et al., 2009] Duarte, T. L., Cooke, M. S., and Jones, G. D. D. (2009). Gene expression profiling reveals new protective roles for vitamin C in human skin cells. *Free Radical Biology and Medicine*, 46(1):78–87.

[e Sousa et al., 2019] e Sousa, M. B., Galli, G., Lyra, D. H., Granato, Í. S. C., Matias, F. I., Alves, F. C., and Fritsche-Neto, R. (2019). Increasing accuracy and reducing costs of genomic prediction by marker selection. *Euphytica*, 215(2):1–14.

[Edge-Garza et al., 2015] Edge-Garza, D. A., Luby, J. J., and Peace, C. (2015). Decision support for cost-efficient and logistically feasible marker-assisted seedling selection in fruit breeding. *Molecular Breeding*, 35(12):1–15.

[Edger et al., 2019] Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., Smith, R. D., Teresi, S. J., Nelson, A. D. L., Wai, C. M., Alger, E. I., Bird, K. A., Yocca, A. E., Pumplin, N., Ou, S., Ben-Zvi, G., Brodt, A., Baruch, K., Swale, T., Shiue, L., Acharya, C. B., Cole, G. S., Mower, J. P., Childs, K. L., Jiang, N., Lyons, E., Freeling, M., Puzey, J. R., and Knapp, S. J. (2019). Origin and evolution of the octoploid strawberry genome. *Nature Genetics*, 51(1):541–547.

[Edwards and Gibbs, 1994] Edwards, M. C. and Gibbs, R. A. (1994). Multiplex PCR: Advantages, development, and applications. *Genome Research*, 3(4):S65–S75.

[Endelman, 2011] Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, 4(3):250–255.

[Endelman and Jannink, 2012] Endelman, J. B. and Jannink, J. L. (2012). Shrinkage estimation of the realized relationship matrix. *G3: Genes, Genomes, Genetics*, 2(11):1405–1413.

[Felgines et al., 2003] Felgines, C., Talavéra, S., Gonthier, M.-P., Texier, O., Scalbert, A., Lamaison, J.-L., and Rémésy, C. (2003). Strawberry Anthocyanins Are Recovered in Urine as Glucuro- and Sulfoconjugates in Humans. *The Journal of Nutrition*, 133(5):1296–1301.

[Fernandez et al., 2001] Fernandez, G. E., Butler, L. M., and Louws, F. J. (2001). Strawberry growth and development in an annual plasticulture system. *HortScience*, 36(7):1219–1223.

[Finn et al., 2013] Finn, C. E., Retamales, J. B., Lobos, G. A., and Hancock, J. F. (2013). The chilean strawberry (*Fragaria chiloensis*): Over 1000 years of domestication. *HortScience*, 48(4):418–421.

[Food and Nutrition Board, 2002] Food and Nutrition Board, I. o. M. (2002). *Dietary Reference Intakes.*

[Forbes et al., 2016] Forbes, T., Gasparrini, M., Afrin, S., Mazzoni, L., Reboredo-Rodríguez, P., and Giampieri, F. (2016). A comparative study on cytotoxic effects of strawberry extract on different cellular models. *Journal of Berry Research*, 6:263–275.

[Foulley, 1993] Foulley, J. L. (1993). Foulley1993.Pdf.

[Freedman et al., 2007] Freedman, N. D., Park, Y., Subar, A. F., Hollenbeck, A. R., Leitzmann, M. F., Schatzkin, A., and Abnet, C. C. (2007). Fruit and vegetable intake and esophageal cancer in a large prospective cohort study. *International Journal of Cancer*, 121(12):2753–2760.

[Freedman et al., 2008] Freedman, N. D., Park, Y., Subar, A. F., Hollenbeck, A. R., Leitzmann, M. F., Schatzkin, A., and Abnet, C. C. (2008). Fruit and vegetable intake and head and neck cancer risk in a large United States prospective cohort study. *International Journal of Cancer*, 122(10):2330–2336.

[Gabbay et al., 2010] Gabbay, K. H., Bohren, K. M., Morello, R., Bertin, T., Liu, J., and Vogel, P. (2010). Ascorbate synthesis pathway: Dual role of ascorbate in bone homeostasis. *Journal of Biological Chemistry*, 285(25):19510–19520.

[Gallardo et al., 2014] Gallardo, R. K., Li, H., Mccracken, V., Yue, C., Luby, J., and Mcferson, J. R. (2014). Market intermediaries' willingness to pay for apple, peach, cherry, and strawberry quality attributes. *Agribusiness*, 31(2):259–280.

[Garaycoechea et al., 2018] Garaycoechea, J. I., Crossan, G. P., Langevin, F., Mulderrig, L., Louzada, S., Yang, F., Guilbaud, G., Park, N., Roerink, S., Nik-Zainal, S., Stratton, M. R., and Patel, K. J. (2018). Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature*, 553(7687):171–177.

[Gasparrini et al., 2018] Gasparrini, M., Giampieri, F., Forbes-Hernandez, T. Y., Afrin, S., Cianciosi, D., Reboredo-Rodriguez, P., Varela-Lopez, A., Zhang, J. J., Quiles, J. L., Mezzetti, B., Bompadre, S., and Battino, M. (2018). Strawberry extracts efficiently counteract inflammatory stress induced by the endotoxin lipopolysaccharide in Human Dermal Fibroblast. *Food and Chemical Toxicology*, 114(February):128–140.

[Geraldi, 2005] Geraldi, I. O. (2005). *Selection Indices for Population Improvement Programmes.*

[Gezan et al., 2017] Gezan, S. A., Osorio, L. F., Verma, S., and Whitaker, V. M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Horticulture Research*, 4(October 2016):1 – 9.

[Giampieri et al., 2012] Giampieri, F., Tulipani, S., Alvarez-Suarez, J. M., Quiles, J. L., Mezzetti, B., and Battino, M. (2012). The strawberry: Composition, nutritional quality, and impact on human health. *Nutrition*, 28(1):9–19.

[Gianola and Van Kaam, 2008] Gianola, D. and Van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178(4):2289–2303.

[Gillespie, 1977] Gillespie, D. T. (1977). Exact Stochastic Simulation of couple chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

[Giné Bordonaba and Terry, 2010] Giné Bordonaba, J. and Terry, L. A. (2010). Manipulating the taste-related composition of strawberry fruits (*Fragaria ananassa*) from different cultivars using deficit irrigation. *Food Chemistry*, 122(4):1020–1026.

[Goddard, 2008] Goddard, M. (2008). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, 136(2):245–257.

[Goddard and Hayes, 2007] Goddard, M. E. and Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, 124(6):323–330.

[Golbach et al., 2015] Golbach, F., Kootstra, G., Damjanovic, S., Otten, G., and van de Zedde, R. (2015). Validation of plant part measurements using a 3D reconstruction method suitable for high-throughput seedling phenotyping. *Machine Vision and Applications*, 27(5):663–680.

[González-Recio et al., 2014] González-Recio, O., Rosa, G. J., and Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*, 166(1):217–231.

[Grabherr et al., 2010] Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., di Palma, F., and Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, 26(9):1145–1151.

[Graham et al., 1997] Graham, J., Gordon, S. C., and McNicol, R. J. (1997). The effect of the CpTi gene in strawberry against attack by vine weevil (*Otiorhynchus sulcatus* F. Coleoptera: Curculionidae). *Annals of Applied Biology*, 131(1):133–139.

[Gunady et al., 2012] Gunady, M. G., Biswas, W., Solah, V. A., and James, A. P. (2012). Evaluating the global warming potential of the fresh produce supply chain for strawberries, romaine/cos lettuces (*Lactuca sativa*), and button mushrooms (*Agaricus bisporus*) in Western Australia using life cycle assessment (LCA). *Journal of Cleaner Production*, 28:81–87.

[Guo et al., 2020] Guo, R., Dhliwayo, T., Mageto, E. K., Palacios-Rojas, N., Lee, M., Yu, D., Ruan, Y., Zhang, A., San Vicente, F., Olsen, M., Crossa, J., Prasanna, B. M., Zhang, L., and Zhang, X. (2020). Genomic Prediction of Kernel Zinc Concentration in Multiple Maize Populations Using Genotyping-by-Sequencing and Repeat Amplification Sequencing Markers. *Frontiers in Plant Science*, 11(May):1–15.

[Habier et al., 2007] Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397.

[Habier et al., 2011] Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12.

[Hamner, 2012] Hamner, B. (2012). Metrics: Evaluation metrics for machine learning. R package version 0.1.1.

[Hancock et al., 1999] Hancock, J. F., Lavín, A., and Retamales, J. B. (1999). Our southern strawberry heritage: *Fragaria chiloensis* of Chile. *HortScience*, 34(5):814–816.

[Hancock et al., 2008] Hancock, J. F., Sjulin, T. M., and Lobos, G. A. (2008). Strawberries. In *Temperate Fruit Crop Breeding: Germplasm to Genomics*, number January, pages 393–437.

[Haymes et al., 1997] Haymes, K. M., Henken, B., Davis, T. M., and Van De Weg, W. E. (1997). Identification of RAPD markers linked to a *Phytophthora fragariae* resistance gene (*Rpf1*) in the cultivated strawberry. *Theoretical and Applied Genetics*, 94(8):1097–1101.

[He et al., 2018] He, D., Saha, S., Finkers, R., and Parida, L. (2018). Efficient algorithms for polyploid haplotype phasing. *BMC Genomics*, 19.

[He et al., 2017] He, J. Q., Harrison, R. J., and Li, B. (2017). A novel 3D imaging system for strawberry phenotyping. *Plant Methods*, 13(1):93.

[He et al., 2019] He, L., Xiao, J., Rashid, K. Y., Jia, G., Li, P., Yao, Z., Wang, X., Cloutier, S., and You, F. M. (2019). Evaluation of genomic prediction for Pasmo resistance in flax. *International Journal of Molecular Sciences*, 20(2):359.

[He et al., 2016] He, S., Schulthess, A. W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., Ebmeyer, E., Reif, J. C., and Jiang, Y. (2016). Genomic selection in a commercial winter wheat population. *Theoretical and Applied Genetics*, 129(3):641–651.

[He et al., 2015] He, S., Zhao, Y., Mette, M. F., Bothe, R., Ebmeyer, E., Sharbel, T. F., Reif, J. C., and Jiang, Y. (2015). Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (*Triticum aestivum* L.). *BMC Genomics*, 16(1):1–12.

[Heffner et al., 2009] Heffner, E. L., Sorrells, M. E., and Jannink, J.-l. (2009). Genomic Selection for Crop Improvement. *Crop Science*, 49(February):1–12.

[Heide et al., 2013] Heide, O. M., Stavang, J. A., and Sønsteby, A. (2013). Physiology and genetics of flowering in cultivated and wild strawberries - A review. *Journal of Horticultural Science and Biotechnology*, 88(1):1–18.

[Henderson, 1984] Henderson, C. R. (1984). Applications of Linear Models in Animal Breeding Models. *University of Guelph*, page 384.

[Heslot et al., 2012] Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Science*, 52(1):146–160.

[Hoff et al., 2015] Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2015). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, 32(5):767–769.

[Holt and Yandell, 2011] Holt, C. and Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(1).

[Hooper et al., 2008] Hooper, L., Kroon, P. A., Rimm, E. B., Cohn, J. S., Harvey, I., Cornu, K. A. L., Ryder, J. J., Hall, W. L., and Cassidy, A. (2008). Flavonoids , flavonoid-rich foods , and cardiovascular risk : a meta-analysis of randomized controlled trials 1 , 2. *American Journal of Clinical Nutrition*, 88(1):38–50.

[Houde et al., 2004] Houde, M., Sylvain-Dallaire, N'Dong, D., and Sarhan, F. (2004). Overexpression of the acidic dehydrin WCOR410 improves freezing tolerance in transgenic strawberry leaves. *Plant Biotechnology Journal*, 2(5):381–387.

[hristian Rose et al., 2015] hristian Rose, J. C., Paulus, S., and Kuhlmann, H. (2015). Accuracy analysis of a multi-view stereo approach for phenotyping of tomato plants at the organ level. *Sensors (Basel, Switzerland)*, 15(5):9651–9665.

[Huang et al., 2020] Huang, Z., Sklar, E., and Parsons, S. (2020). Design of automatic strawberry harvest robot suitable in complex environments. *ACM/IEEE International Conference on Human-Robot Interaction*, pages 567–569.

[Huang et al., 2017] Huang, Z., Wane, S., and Parsons, S. (2017). Towards automated strawberry harvesting: Identifying the picking point. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10454 LNAI:222–236.

[Hytönen, 2009] Hytönen, T. (2009). *Regulation of strawberry growth and development.* PhD thesis.

[Ichijima, 1926] Ichijima, K. (1926). Cytological and genetic studies on Fragaria. *Genetics*, 11(6):590–604.

[Illa et al., 2011] Illa, E., Sargent, D. J., Lopez Girona, E., Bushakra, J., Cestaro, A., Crowhurst, R., Pindo, M., Cabrera, A., van der Knaap, E., Iezzoni, A., Gardiner, S., Velasco, R., Arús, P., Chagné, D., and Troggio, M. (2011). Comparative analysis of rosaceous genomes and the reconstruction of a putative ancestral genome for the family. *BMC Evolutionary Biology*, 11(1):9.

[Jensen et al., 2012] Jensen, B., Knudsen, I. M., Andersen, B., Nielsen, K. F., Thrane, U., Jensen, D. F., and Larsen, J. (2012). Characterization of microbial communities and fungal metabolites on field grown strawberries from organic and conventional production. *International Journal of Food Microbiology*, 160(3):313–322.

[Jensen et al., 2020] Jensen, S. E., Charles, J. R., Muleta, K., Bradbury, P. J., Casstevens, T., Deshpande, S. P., Gore, M. A., Gupta, R., Ilut, D. C., Johnson, L., Lozano, R., Miller, Z., Ramu, P., Rathore, A., Romay, M. C., Upadhyaya, H. D., Varshney, R. K., Morris, G. P., Pressoir, G., Buckler, E. S., and Ramstein, G. P. (2020). A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *Plant Genome*, 13(1):1–15.

[Johnson et al., 2014] Johnson, A. L., Govindarajulu, R., and Ashman, T.-l. (2014). Bioclimatic evaluation of geographical range in *Fragaria* (Rosaceae): consequences of variation in breeding system, ploidy and species age. *Botanical Journal of the Linnean Society*, 176(1):99–114.

[Juraske et al., 2009] Juraske, R., Mutel, C. L., Stoessel, F., and Hellweg, S. (2009). Life cycle human toxicity assessment of pesticides: Comparing fruit and vegetable diets in Switzerland and the United States. *Chemosphere*, 77(7):939–945.

[Kang et al., 2008] Kang, S. P., East, A. R., and Trujillo, F. J. (2008). Colour vision system evaluation of bicolour fruit: A case study with 'B74' mango. *Postharvest Biology and Technology*, 49(1):77–85.

[Kantartzi, 2013] Kantartzi, S. K. (2013). *Microsatellites. Methods and protocols.*

[Kärkkäinen and Sillanpää, 2012] Kärkkäinen, H. P. and Sillanpää, M. J. (2012). Back to basics for Bayesian model building in genomic selection. *Genetics*, 191(3):969–987.

[Karlsson et al., 2004] Karlsson, A. L., Aim, R., Ekstrand, B., Fjelkner-Modig, S., Schiött, A., Bengtsson, U., Björk, L., Hjerno, K., Roepstorff, P., and Emanuelsson, C. S. (2004). Bet v 1 homologues in strawberry identified as IgE-binding proteins and presumptive allergens. *Allergy: European Journal of Allergy and Clinical Immunology*, 59(12):1277–1284.

[Kazhdan et al., 2006] Kazhdan, M., Bolitho, M., and Hoppe, H. (2006). Poisson Surface Reconstruction. *Proceedings of the Symposium on Geometry Processing*, pages 61–70.

[Keilwagen et al., 2018] Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*, 19(1):189.

[Keilwagen et al., 2016] Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, 44(9):e89.

[Khammuang et al., 2005] Khammuang, S., Dheeranupattana, S., Hanmuangjai, P., and Wongroung, S. (2005). *Agrobacterium*-mediated transformation of modified antifreeze protein gene in strawberry. *Songklanakarin Journal of Science and Technology*, 27(4):693–703.

[Khlestkina and Salina, 2006] Khlestkina, E. K. and Salina, E. A. (2006). SNP markers: Methods of analysis, ways of development, and comparison on an example of common wheat. *Russian Journal of Genetics*, 42(6):585–594.

[Khoshnevisan et al., 2013] Khoshnevisan, B., Rafiee, S., and Mousazadeh, H. (2013). Environmental impact assessment of open field and greenhouse strawberry production. *European Journal of Agronomy*, 50:29–37.

[Kim et al., 2016] Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L. S., and Paterson, A. H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science*, 242:14–22.

[Kirsten, 2014] Kirsten, A. (2014). *The Mid-Atlantic Berry Guide.*

[Kjaer and Ottosen, 2015] Kjaer, K. H. and Ottosen, C.-O. (2015). 3D Laser Triangulation for Plant Phenotyping in Challenging Environments. *Sensors (Basel, Switzerland)*, 15(6):13533–47.

[Klee and Tieman, 2018] Klee, H. J. and Tieman, D. M. (2018). The genetics of fruit flavour preferences. *Nature Reviews Genetics*, 19(6):347–356.

[Klerks et al., 2004] Klerks, M. M., Lindner, J. L., Vaskova, D., Spak, J., Thompson, J. R., Jelkmann, W., and Schoen, C. D. (2004). Detection and tentative grouping of Strawberry crinkle virus isolates. *European Journal of Plant Pathology*, 110:45–52.

[Klodt et al., 2015] Klodt, M., Herzog, K., Töpfer, R., and Cremers, D. (2015). Field phenotyping of grapevine growth using dense stereo reconstruction. *BMC Bioinformatics*, 16(1):143.

[Klose et al., 2011] Klose, R., Penlington, J., and Ruckelshausen, A. (2011). Usability of 3D time-of-flight cameras for automatic plant phenotyping. *Bornimer Agrartechnische Berichte*, 69:93–105.

[Knight et al., 2005] Knight, V., Evans, K., Simpson, D., and Tobutt, K. R. (2005). Report on a desktop study to investigate the current world resources in Rosaceous fruit breeding programmes. (July).

[Kriaridou et al., 2020] Kriaridou, C., Tsairidou, S., Houston, R. D., and Robledo, D. (2020). Genomic Prediction Using Low Density Marker Panels in Aquaculture: Performance Across Species, Traits, and Genotyping Platforms. *Frontiers in Genetics*, 11(February):1–8.

[Kruijer et al., 2014] Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., Kooke, R., Keurentjes, J. J., and Van Eeuwijk, F. A. (2014). Marker-based estimation of heritability in immortal populations. *Genetics*, 199(2):379–398.

[Kurotobi et al., 2010] Kurotobi, T., Fukuhara, K., Inage, H., and Kimura, S. (2010). Glycemic Index and Postprandial Blood Glucose Response to Japanese Strawberry Jam in Normal Adults. *Journal of Nutritional Science and Vitaminology*, 56(3):198–202.

[Labadie et al., 2020] Labadie, M., Vallin, G., Petit, A., Ring, L., Hoffmann, T., Gaston, A., Potier, A., Schwab, W., Rothan, C., and Denoyes, B. (2020). Metabolite Quantitative Trait Loci for Flavonoids Provide New Insights into the Genetic Architecture of Strawberry (*Fragaria ananassa*) Fruit Quality. *Journal of Agricultural and Food Chemistry*, 68(25):6927–6939.

[Lafarga et al., 2018] Lafarga, T., Colás-Medà, P., Abadías, M., Aguiló-Aguayo, I., Bobo, G., and Viñas, I. (2018). Strategies to reduce microbial risk and improve quality of fresh and processed strawberries: A review. *Innovative Food Science and Emerging Technologies*, 52:197–212.

[Lande and Thompson, 1990] Lande, R. and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–756.

[Le Rouzic, 2014] Le Rouzic, A. (2014). Estimating directional epistasis. *Frontiers in Genetics*, 5(JUL):1–14.

[Li et al., 2020] Li, Y., Liu, T., Luo, H., and Liu, S. (2020). The transcriptional landscape of cultivated strawberry (*Fragaria x ananassa*) and its diploid ancestor (*Fragaria vesca*) during fruit development. *bioRxiv*.

[Li et al., 2019] Li, Y., Pi, M., Gao, Q., Liu, Z., and Kang, C. (2019). Updated annotation of the wild strawberry *Fragaria vesca* V4 genome. *Horticulture Research*, 6(1):61.

[Lin et al., 2002] Lin, L., Hedayat, a. S., Sinha, B., and Yang, M. (2002). Statistical Methods in Assessing Agreement: Models, Issues, and Tools. *Journal of the American Statistical Association*, 97(457):257–270.

[Lin, 1989] Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *International Biometric Society*, 45(1):255 – 268.

[Lin et al., 2014] Lin, Z., Hayes, B. J., and Daetwyler, H. D. (2014). Genomic selection in crops, trees and forages: A review. *Crop and Pasture Science*, 65(11):1177–1191.

[Lindeberg, 2012] Lindeberg, T. (2012). Scale Invariant Feature Transform. *Scholarpedia*, 7(5):10491.

[Liston et al., 2014] Liston, A., Cronn, R., and Ashman, T.-L. (2014). Fragaria: A genus with deep historical roots and ripe for evolutionary and ecological insights. *American Journal of Botany*, 101(10):1686–1699.

[Liston et al., 2020] Liston, A., Wei, N., Tennessen, J. A., Li, J., Dong, M., and Ashman, T. L. (2020). Revisiting the origin of octoploid strawberry. *Nature Genetics*, 52(1):2–4.

[Liu et al., 2019] Liu, H., Tessema, B. B., Jensen, J., Cericola, F., Andersen, J. R., and Sørensen, A. C. (2019). ADAM-Plant: A software for stochastic simulations of plant breeding from molecular to phenotypic level and from simple selection to complex speed breeding programs. *Frontiers in Plant Science*, 9(January):1–15.

[Liu et al., 2021] Liu, T., Li, M., Liu, Z., Ai, X., and Li, Y. (2021). Reannotation of the cultivated strawberry genome and establishment of a strawberry genome database. *Horticulture Research*, 8(1):41.

[Longhi et al., 2014] Longhi, S., Giongo, L., Buti, M., Surbanovski, N., Viola, R., Velasco, R., Ward, J. A., and Sargent, D. J. (2014). Molecular genetics and genomics of the Rosoideae: state of the art and future perspectives. *Horticulture Research*, 1(November 2013):1.

[Luby and Shaw, 2000] Luby, J. J. and Shaw, D. V. (2000). Does marker-assisted selection make dollars and sense in a fruit breeding program? : Molecular Genetics: Applications in Small and Tree Fruits: Where Is it Working. *HortScience*, 36(5):872–879.

[Lundberg et al., 2009] Lundberg, M., Töpel, M., Eriksen, B., Nylander, J. A., and Eriksson, T. (2009). Allopolyploidy in Fragariinae (Rosaceae): Comparing four DNA sequence regions, with comments on classification. *Molecular Phylogenetics and Evolution*, 51(2):269–280.

[Määttä-Riihinen et al., 2004] Määttä-Riihinen, K. R., Kamal-Eldin, A., and Törrönen, A. R. (2004). Identification and quantification of phenolic compounds in berries of Fragaria and Rubus species (family rosaceae). *Journal of Agricultural and Food Chemistry*, 52(20):6178–6187.

[Mahoney et al., 2010] Mahoney, L., Quimby, M., Shields, M., and Davis, T. (2010). Mitochondrial DNA transmission, ancestry, and sequences in *fragaria*. In *Acta Horticulturae*, pages 301–308.

[Mangandi et al., 2017] Mangandi, J., Verma, S., Osorio, L., Peres, N. A., van de Weg, E., and Whitaker, V. M. (2017). Pedigree-based analysis in a multiparental population of octoploid strawberry reveals QTL alleles conferring resistance to phytophthora cactorum. *G3: Genes, Genomes, Genetics*, 7(6):1707–1719.

[Marçais and Kingsford, 2011] Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

[Martín-Pizarro and Posé, 2018] Martín-Pizarro, C. and Posé, D. (2018). Genome editing as a tool for fruit ripening manipulation. *Frontiers in Plant Science*, 9(September):1–8.

[Mason and Tang, 2016] Mason, J. B. and Tang, S. Y. (2016). Folate status and colorectal cancer risk: A 2016 update. *Molecular Aspects of Medicine*, 53:73–79.

[Mathey et al., 2013] Mathey, M. M., Mookerjee, S., Gündüz, K., Hancock, J. F., Iezzoni, A. F., Mahoney, L. L., Davis, T. M., Bassil, N. V., Hummer, K. E., Stewart, P. J., Whitaker, V. M., Sargent, D. J., Denoyes, B., Amaya, I., Weg, E. V. D., and Finn, C. E. (2013). Large-scale standardized phenotyping of strawberry in RosBREED. *Journal of the AAmerican Pomological Society*, 67(4):205–216.

[Matias et al., 2017] Matias, F. I., Galli, G., Granato, I. S. C., and Fritsche-Neto, R. (2017). Genomic prediction of autogamous and allogamous plants by SNPs and haplotypes. *Crop Science*, 57(6):2951–2958.

[Mee Kin and Guan Huat, 2010] Mee Kin, C. and Guan Huat, T. (2010). Headspace solid-phase microextraction for the evaluation of pesticide residue contents in cucumber and strawberry after washing treatment. *Food Chemistry*, 123(3):760–764.

[Mehraj and Jamal Uddin, 2014] Mehraj, H. and Jamal Uddin, A. F. M. (2014). Correlation pathway for phenotypic variability study in strawberry. *International Journal of Sustainable Agricultural Technology*, 10(109):1815–1272.

[Meller Haral et al., 2012] Meller Haral, Y., Elad, Y., Rav-David, D., Borenstein, M., Shulchani, R., Lew, B., and Graber, E. R. (2012). Biochar mediates systemic response of strawberry to foliar fungal pathogens. *Plant and Soil*, 357(1):245–257.

[Mendoza and Aguilera, 2004] Mendoza, F. and Aguilera, J. (2004). Application of image analysis for classification of ripening bananas. *Food Engineering and Physical Properties*, 69(9):471–477.

[Meuwissen et al., 2001] Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-Wide dense marker maps. *Genetics*, 157(4):1819–1829.

[Meyer and Seaman, 2008] Meyer, J. P. and Seaman, M. A. (2008). A comparison of the exact Kruskal-Wallis distribution to asymptotic approximations for N. In *American Educational Research Association*, number March.

[Mezzetti et al., 2018] Mezzetti, B., Giampieri, F., Zhang, Y. T., and Zhong, C. F. (2018). Status of strawberry breeding programs and cultivation systems in Europe and the rest of the world. *Journal of Berry Research*, 8(3):205–221.

[Molenaar et al., 2018] Molenaar, H., Boehm, R., and Piepho, H. P. (2018). Phenotypic selection in ornamental breeding: It's better to have the BLUPs than to have the BLUEs. *Frontiers in Plant Science*, 871(November):1–14.

[Moll and Davis, 2017] Moll, R. and Davis, B. (2017). Iron, vitamin B12and folate. *Medicine (United Kingdom)*, 45(4):198–203.

[Mordini et al., 2009] Mordini, M., Nemecek, T., and Gaillard, G. (2009). Carbon & Water Footprint of Oranges and Strawberries: A Literature Review. page 76.

[Nagata et al., 2000] Nagata, M., Bato, P. M., Mitarai, M., Cao, Q., and Kitahara, T. (2000). Study on sorting system for strawberry using machine vision (Part 1). Development of software for determining the direction of strawberry (Akihime variety). *Journal of the Japanese Society of Agricultural Machinery*, 62(1):100–110.

[Nakaya and Isobe, 2012] Nakaya, A. and Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Annals of Botany*, 110(6):1303–1316.

[Negi et al., 2020] Negi, S., Sharma, G., and Sharma, R. (2020). Introgression and confirmation of everbearing trait in strawberry (*Fragaria x ananassa Duch.*). *Physiology and Molecular Biology of Plants*, 26(12):2407–2416.

[Nellist, 2018] Nellist, C. F. (2018). Disease Resistance in Polyploid Strawberry. In *The genomes of rosaceous berries and their wild relatives*, pages 79–94.

[Nellist et al., 2019] Nellist, C. F., Vickerstaff, R. J., Sobczyk, M. K., Marina-Montes, C., Wilson, F. M., Simpson, D. W., Whitehouse, A. B., and Harrison, R. J. (2019). Quantitative trait loci controlling *Phytophthora cactorum* resistance in the cultivated octoploid strawberry (*Fragaria ananassa*). *Horticulture Research*, 6(1):60.

[Njuguna, 2010] Njuguna, W. (2010). *Development and Use of Molecular Tools in Fragaria*. PhD thesis.

[Njuguna et al., 2013] Njuguna, W., Liston, A., Cronn, R., Ashman, T. L., and Bassil, N. (2013). Insights into phylogeny, sex function and age of Fragaria based on whole chloroplast genome sequencing. *Molecular Phylogenetics and Evolution*, 66(1):17–29.

[Noh et al., 2018] Noh, Y. H., Oh, Y., Mangandi, J., Verma, S., Zurn, J. D., Lu, Y. T., Fan, Z., Bassil, N., Peres, N., Cole, G., Acharya, C., Famula, R., Knapp, S., Whitaker, V. M., and Lee, S. (2018). High-throughput marker assays for FaRPc2-mediated resistance to Phytophthora crown rot in octoploid strawberry. *Molecular Breeding*, 38(8):1–11.

[Norman et al., 2018] Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. *G3: Genes, Genomes, Genetics*, 8(9):2889–2899.

[Nyine et al., 2018] Nyine, M., Uwimana, B., Blavet, N., Hibová, E., Vanrespaille, H., Batte, M., Akech, V., Brown, A., Lorenzen, J., Swennen, R., and Doležel, J. (2018). Genomic Prediction in a Multiploid Crop: Genotype by Environment Interaction and Allele Dosage Effects on Predictive Ability in Banana. *The Plant Genome*, 11(2):170090.

[Ogutu et al., 2012] Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6(Suppl 2):S10.

[Oh et al., 2020] Oh, Y., Chandra, S., and Lee, S. (2020). Development of Subgenome-Specific Markers for FaRXf1 Conferring Resistance to Bacterial Angular Leaf Spot in Allo-Octoploid Strawberry. *International Journal of Fruit Science*, 20(S2):S198–S210.

[Olatoye et al., 2019] Olatoye, M. O., Clark, L. V., Wang, J., Yang, X., Yamada, T., Sacks, E. J., and Lipka, A. E. (2019). Evaluation of genomic selection and marker-assisted selection in Miscanthus and energycane. *Molecular Breeding*, 39(12):1–16.

[Osorio et al., 2021] Osorio, L. F., Gezan, S. A., Verma, S., and Whitaker, V. M. (2021). Independent Validation of Genomic Prediction in Strawberry Over Multiple Cycles. *Frontiers in Genetics*, 11(January):1–13.

[Pajk et al., 2006] Pajk, T., Rezar, V., Levart, A., and Salobir, J. (2006). Efficiency of apples, strawberries, and tomatoes for reduction of oxidative stress in pigs as a model for humans. *Nutrition*, 22(4):376–384.

[Passey et al., 2003] Passey, A. J., Barrett, K. J., and James, D. J. (2003). Adventitious shoot regeneration from seven commercial strawberry cultivars (*Fragaria x ananassa Duch.*) using a range of explant types. *Plant Cell Reports*, 21(5):397–401.

[Paten et al., 2017] Paten, B., Novak, A. M., Eizenga, J. M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research*, 27(5):665–676.

[Paulus et al., 2013] Paulus, S., Dupuis, J., Mahlein, A.-k., and Kuhlmann, H. (2013). Surface feature based classification of plant organs from 3D laserscanned point clouds for plant phenotyping. *BMC Bioinformatics*, 14:1–12.

[Paulus et al., 2014] Paulus, S., Schumann, H., Kuhlmann, H., and Léon, J. (2014). High-precision laser scanning system for capturing 3D plant architecture and analysing growth of cereal plants. *Biosystems Engineering*, 121:1–11.

[Pedruzzi and Rouzine, 2019] Pedruzzi, G. and Rouzine, I. M. (2019). Epistasis detectably alters correlations between genomic sites in a narrow parameter window. *PLoS ONE*, 14(5):1–16.

[Pelham, 2017] Pelham, J. (2017). The Impact of Brexit on the UK Soft Fruit Industry. Technical report.

[Pereira, 2004] Pereira, V. (2004). Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biology*, 5(10):R79.

[Pet'ka et al., 2012] Pet'ka, J., Leitner, E., and Parameswaran, B. (2012). Musk strawberries: The flavour of a formerly famous fruit reassessed. *Flavour and Fragrance Journal*, 27(4):273–279.

[Piepho and Möhring, 2007] Piepho, H. P. and Möhring, J. (2007). Computing heritability and selection response from unbalanced plant breeding trials. *Genetics*, 177(3):1881–1888.

[Piepho et al., 2008] Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2):209–228.

[Pincot et al., 2020] Pincot, D. D., Hardigan, M. A., Cole, G. S., Famula, R. A., Henry, P. M., Gordon, T. R., and Knapp, S. J. (2020). Accuracy of genomic selection and long-term genetic gain for resistance to Verticillium wilt in strawberry. *Plant Genome*, 13(3):1–19.

[Poland et al., 2012] Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S. Y., Manes, Y., Dreisigacker, S., Crossa, J., Sanchez-Villeda, H., Sorrells, M., and Jannink, J. L. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *Plant Genome*, 5(3):103–113.

[Poling, 2012] Poling, E. B. (2012). Strawberry Plant Structure and Growth Habit. Technical report.

[Pott et al., 2020] Pott, D. M., Vallarino, J. G., Cruz-Rus, E., Willmitzer, L., Sánchez-Sevilla, J. F., Amaya, I., and Osorio, S. (2020). Genetic analysis of phenylpropanoids and antioxidant capacity in strawberry fruit reveals mQTL hotspots and candidate genes. *Scientific Reports*, 10(1):1–15.

[Potter et al., 2007] Potter, D., Eriksson, T., Evans, R. C., Oh, S., Smedmark, J. E. E., Morgan, D. R., Kerr, M., Robertson, K. R., Arsenault, M., Dickinson, T. A., and Campbell, C. S. (2007). Phylogeny and classification of Rosaceae. *Plant Systematics and Evolution*, 266(1-2):5–43.

[Potter et al., 2000] Potter, D., Luby, J. J., and Harrison, R. E. (2000). Phylogenetic Relationships among Species of *Fragaria* (Rosaceae) Inferred from Non-Coding Nuclear and Chloroplast DNA Sequences. *Systemic Botany*, 25(2):337–348.

[Powers, 2003] Powers, H. J. (2003). Riboflavin (vitamin B-2) and health. *The American journal of clinical nutrition*, 77(6):1352–60.

[Qin et al., 2008] Qin, Y., Teixeira da Silva, J. A., Zhang, L., and Zhang, S. (2008). Transgenic strawberry: State of the art for improved traits. *Biotechnology Advances*, 26(3):219–232.

[R Core Team, 2017] R Core Team (2017). *R: A language and environment for statistical computing.*

[Rader et al., 2016] Rader, R., Bartomeus, I., Garibaldi, L. A., Garratt, M. P., Howlett, B. G., Winfree, R., Cunningham, S. A., Mayfield, M. M., Arthur, A. D., Andersson, G. K., Bommarco, R., Brittain, C., Carvalheiro, L. G., Chacoff, N. P., Entling, M. H., Foully, B., Freitas, B. M., Gemmill-Herren, B., Ghazoul, J., Griffin, S. R., Gross, C. L., Herbertsson, L., Herzog, F., Hipólito, J., Jaggar, S., Jauker, F., Klein, A. M., Kleijn, D., Krishnan, S., Lemos, C. Q., Lindström, S. A., Mandelik, Y., Monteiro, V. M., Nelson, W., Nilsson, L., Pattemore, D. E., Pereira, N. D. O., Pisanty, G., Potts, S. G., Reemer, M., Rundlöf, M., Sheffield, C. S., Scheper, J., Schüepp, C., Smith, H. G., Stanley, D. A., Stout, J. C., Szentgyörgyi, H., Taki, H., Vergara, C. H., Viana, B. F., and Woyciechowski, M. (2016). Non-bee insects are important contributors to global crop pollination. *Proceedings of the National Academy of Sciences of the United States of America*, 113(1):146–151.

[Robinson, 1991] Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32.

[Romero-Gámez and Suárez-Rey, 2020] Romero-Gámez, M. and Suárez-Rey, E. M. (2020). Environmental footprint of cultivating strawberry in Spain. *International Journal of Life Cycle Assessment*, 25(4):719–732.

[Rowan et al., 2017] Rowan, B. A., Seymour, D. K., Chae, E., Lundberg, D. S., and Weigel, D. (2017). Methods for genotyping-by-sequencing. *Methods in Molecular Biology*, 1492:221–242.

[Rubinstein, 2015] Rubinstein, J. (2015). *Fragaria x ananassa:* Past, Present and Future Production of the Modern Strawberry. pages 1–30.

[Saiki et al., 1988] Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. a. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (New York, N.Y.)*, 239(4839):487–491.

[Salentijn et al., 2003] Salentijn, E. M., Aharoni, A., Schaart, J. G., Boone, M. J., and Krens, F. A. (2003). Differential gene expression analysis of strawberry cultivars that differ in fruit-firmness. *Physiologia Plantarum*, 118(4):571–578.

[Salzberg, 2019] Salzberg, S. L. (2019). Next-generation genome annotation: We still struggle to get it right. *Genome Biology*, 20(1):19–21.

[Savini et al., 2008] Savini, G., Giorgi, V., Scarano, E., and Neri, D. (2008). Strawberry plant relationship through the stolon. *Physiologia Plantarum*, 134(3):421–429.

[Schaart et al., 2011] Schaart, J. G., Kjellsen, T. D., Mehli, L., and Heggem, R. (2011). Towards the production of genetically modified strawberries which are acceptable to consumers. *Genes, Genomes and Genomics*, (Special Issue 1):102–107.

[Schirmer et al., 2015] Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6):e37.

[Schmidt et al., 2019] Schmidt, D. A., Campbell, N. R., Govindarajulu, P., Larsen, K. W., and Russello, M. A. (2019). Genotyping-in-Thousands by sequencing (GT-seq) panel development and application to minimally invasive DNA samples to support studies in molecular ecology. *Molecular Ecology Resources*, 20(1):114–124.

[Schwieterman et al., 2014] Schwieterman, M. L., Colquhoun, T. A., Jaworski, E. A., Bartoshuk, L. M., Gilbert, J. L., Tieman, D. M., Odabasi, A. Z., Moskowitz, H. R., Folta, K. M., Klee, H. J., Sims, C. A., Whitaker, V. M., and Clark, D. G. (2014). Strawberry flavor: Diverse chemical compositions, a seasonal influence, and effects on sensory perception. *PLoS ONE*, 9(2):e88446.

[Shamshad and Sharma, 2018] Shamshad, M. and Sharma, A. (2018). The Usage of Genomic Selection Strategy in Plant Breeding. *Next Generation Plant Breeding*.

[Shen et al., 2010] Shen, Z., Qu, W., Wang, W., Lu, Y., Wu, Y., Li, Z., Hang, X., Wang, X., Zhao, D., and Zhang, C. (2010). MPprimer: a program for reliable multiplex PCR primer design. *BMC bioinformatics*, 11(1):143.

[Shi and Lai, 2015] Shi, J. and Lai, J. (2015). Patterns of genomic changes with crop domestication and breeding. *Current Opinion in Plant Biology*, 24:47–53.

[Shulaev et al., 2011] Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J.-M., Rees, D. J. G., Williams, K. P., Holt, S. H., Ruiz Rojas, J. J., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. a., Troggio, M., Viola, R., Ashman, T.-L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant, D. W., Fox, S. E., Givan, S. a., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Lopez Girona, E., Zdepski, A., Wang, W., Kerstetter, R. a., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E., and Folta, K. M. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nature genetics*, 43(2):109–16.

[Signorell et al., 2021] Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Arppe, A., Baddeley, A., Barton, K., Bolker, B., Borchers, H. W., Caeiro, F., Champely, S., Chessel, D., Chhay, L., Cooper, N., Cummins, C., Dewey, M., Doran, H. C., Dray, S., Dupont, C., Eddelbuettel, D., Ekstrom, C., Elff, M., Enos, J., Farebrother, R. W., Fox, J., Francois, R., Friendly, M., Galili, T., Gamer, M., Gastwirth, J. L., Gegzna, V., Gel, Y. R., Graber, S., Gross, J., Grothendieck, G., Harrell, F. E., Heiberger, R., Hoehle, M., Hoffmann, C. W., Hojsgaard, S., Hothorn, T., Huerzeler, M., Hui, W. W., Hurd, P., Hyndman, R. J., Jackson, C., Kohl, M., Korpela, M., Kuhn, M., Labes, D., Leisch, F., Lemon, J., Li, D., Maechler, M., Magnusson, A., Mainwaring, B., Malter, D., Marsaglia, G., Marsaglia, J., Matei, A., Meyer, D., Miao, W., Millo, G., Min, Y., Mitchell, D., Mueller, F., Naepflin, M., Navarro, D., Nilsson, H., Nordhausen, K., Ogle, D., Ooi, H., Parsons, N., Pavoine, S., Plate, T., Prendergast, L., Rapold, R., Revelle, W., Rinker, T., Ripley, B. D., Rodriguez, C., Russell, N., Sabbe, N., Scherer, R., Seshan, V. E., Smithson, M., Snow, G., Soetaert, K., Stahel, W. A., Stephenson, A., Stevenson, M., Stubner, R., Templ, M., Lang, D. T., Therneau, T., Tille, Y., Torgo, L., Trapletti, A., Ulrich, J., Ushey, K., VanDerWal, J., Venables, B., Verzani, J., Iglesias, P. J. V., Warnes, G. R., Wellek, S., Wickham, H., Wilcox, R. R., Wolf, P., Wollschlaeger, D., Wood, J., Wu, Y., Yee, T., and Zeileis, A. (2021). Package DescTools'.

[Siles et al., 2013] Siles, J. A., Serrano, A., Martín, A., and Martín, M. A. (2013). Biomethanization of waste derived from strawberry processing: Advantages of pretreatment. *Journal of Cleaner Production*, 42:190–197.

[Silva et al., 2015] Silva, K. J. P., Brunings, A., Peres, N. A., Mou, Z., and Folta, K. M. (2015). The Arabidopsis NPR1 gene confers broad-spectrum disease resistance in strawberry. *Transgenic Research*, 24(4):693–704.

[Somasagara et al., 2012] Somasagara, R. R., Hegde, M., Chiruvella, K. K., Musini, A., Choudhary, B., and Raghavan, S. C. (2012). Extracts of Strawberry Fruits Induce Intrinsic Pathway of Apoptosis in Breast Cancer Cells and Inhibits Tumor Progression in Mice. *PLoS ONE*, 7(10):1–11.

[Sood et al., 2020] Sood, S., Bhardwaj, V., Kaushik, S., and Sharma, S. (2020). Prediction based on estimated breeding values using genealogy for tuber yield and late blight resistance in auto-tetraploid potato (*Solanum tuberosum* L.). *Heliyon*, 6(11):e05624.

[Soode et al., 2015] Soode, E., Lampert, P., Weber-Blaschke, G., and Richter, K. (2015). Carbon footprints of the horticultural products strawberries, asparagus, roses and orchids in Germany. *Journal of Cleaner Production*, 87(1):168–179.

[Soode-Schimonsky et al., 2017] Soode-Schimonsky, E., Richter, K., and Weber-Blaschke, G. (2017). Product environmental footprint of strawberries: Case studies in Estonia and Germany. *Journal of Environmental Management*, 203:564–577.

[Stanke and Morgenstern, 2005] Stanke, M. and Morgenstern, B. (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, 33(SUPPL. 2):465–467.

[Stevens, 2005] Stevens, M. D. (2005). *Sustainability of Cold-Climate Strawberry Production Systems*. PhD thesis.

[Sverrisdóttir et al., 2017] Sverrisdóttir, E., Byrne, S., Sundmark, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L., and Nielsen, K. L. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theoretical and Applied Genetics*, 130(10):2091–2108.

[Symons et al., 2012] Symons, G. M., Chua, Y. J., Ross, J. J., Quittenden, L. J., Davies, N. W., and Reid, J. B. (2012). Hormonal changes during non-climateric ripening in strawberry. *Journal of Experimental Botany*, 63(13):4741–4750.

[Tabatabaie and Murthy, 2016] Tabatabaie, S. M. H. and Murthy, G. S. (2016). Cradle to farm gate life cycle assessment of strawberry production in the United States. *Journal of Cleaner Production*, 127(April):548–554.

[Tayeh et al., 2015] Tayeh, N., Klein, A., Le Paslier, M. C., Jacquin, F., Houtin, H., Rond, C., Chabert-Martinello, M., Magnin-Robert, J. B., Marget, P., Aubert, G., and Burstin, J. (2015). Genomic prediction in pea: Effect of marker density and training population size and composition on prediction accuracy. *Frontiers in Plant Science*, 6(NOVEMBER):1–11.

[Tennessen et al., 2014] Tennessen, J. A., Govindarajulu, R., Ashman, T. L., and Liston, A. (2014). Evolutionary origins and dynamics of octoploid strawberry subgenomes revealed by dense targeted capture linkage maps. *Genome Biology and Evolution*, 6(12):3295–3313.

[Thompson and Jelkmann, 2004] Thompson, J. R. and Jelkmann, W. (2004). Strain diversity and conserved genome elements in Strawberry mild yellow edge virus. *Archives of Virology*, 149(10):1897–1909.

[Throop et al., 2005] Throop, J. A., Aneshansley, D. J., Anger, W. C., and Peterson, D. L. (2005). Quality evaluation of apples based on surface defects: Development of an automated inspection system. *Postharvest Biology and Technology*, 36(3):281–290.

[Tomczyk and Latté, 2009] Tomczyk, M. and Latté, K. P. (2009). Potentilla-A review of its phytochemical and pharmacological profile. *Journal of Ethnopharmacology*, 122(2):184–204.

[Trejo-téllez and Gómez-merino, 2014] Trejo-téllez, L. I. and Gómez-merino, F. C. (2014). Nutrient management in strawberry: effects on yield, quality and plant health. In *Strawberries*, pages 239–267.

[Ulrich and Olbricht, 2016] Ulrich, D. and Olbricht, K. (2016). A search for the ideal flavor of strawberry - Comparison of consumer acceptance and metabolite patterns in *Fragaria x ananassa Duch. Journal of Applied Botany and Food Quality*, 89:223–234.

[van Ooijen, 2009] van Ooijen, J. W. (2009). MapQTL 6. *Genome*, (April).

[Vázquez-Arellano et al., 2016] Vázquez-Arellano, M., Griepentrog, H. W., Reiser, D., and Paraforos, D. S. (2016). 3-D imaging systems for agricultural applicationsa review. *Sensors (Switzerland)*, 16(5).

[Verma et al., 2017] Verma, S., Bassil, N. V., Van De Weg, E., Harrison, R. J., Monfort, A., Hidalgo, J. M., Amaya, I., Denoyes, B., Mahoney, L. L., Davis, T. M., Fan, Z., Knapp, S., and Whitaker, V. M. (2017). Development and evaluation of the Axiom® IStraw35 384HT array for the allo-octoploid cultivated strawberry *Fragaria ananassa*. *Acta Horticulturae*, 1156(April):75–81.

[Viana et al., 2011] Viana, J. M. S., Faria, V. R., Fonseca e Silva, F., and de Resende, M. D. V. (2011). Best linear unbiased prediction and family selection in crop species. *Crop Science*, 51(6):2371–2381.

[Villarreal et al., 2016] Villarreal, N. M., Marina, M., Nardi, C. F., Civello, P. M., and Martínez, G. A. (2016). Novel insights of ethylene role in strawberry cell wall metabolism. *Plant Science*, 252:1–11.

[Vining et al., 2017] Vining, K. J., Salinas, N., Tennessen, J. A., Zurn, J. D., Sargent, D. J., Hancock, J., and Bassil, N. V. (2017). Genotyping-by-sequencing enables linkage mapping in three octoploid cultivated strawberry families. *PeerJ*, 5:e3731.

[Visscher et al., 1996] Visscher, P. M., Thompson, R., and Haley, C. S. (1996). Confidence intervals in QTL mapping by bootstrapping. *Genetics*, 143(2):1013–1020.

[Voss-Fels et al., 2019] Voss-Fels, K. P., Cooper, M., and Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, 132(3):669–686.

[Wang et al., 2019] Wang, K., Li, H., Xu, Y., Shao, Q., Yi, J., Wang, R., Cai, W., Hang, X., Zhang, C., Cai, H., and Qu, W. (2019). MFEprimer-3.0: Quality control for PCR primers. *Nucleic Acids Research*, 47(W1):W610–W613.

[Wang et al., 2015] Wang, X. L., Zhong, Y., Cheng, Z. M., and Xiong, J. S. (2015). Divergence of the bZIP Gene Family in Strawberry, Peach, and Apple Suggests Multiple Modes of Gene Evolution after Duplication. *International Journal of Genomics*, 2015.

[Wang et al., 2004] Wang, Z., Chen, Y., and Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, proteomics & bioinformatics / Beijing Genomics Institute*, 2(4):216–221.

[Wannemuehler, 2018] Wannemuehler, S. D. (2018). *Cost-Benefit Analysis of Marker-Assisted Selection in Rosaceous Fruit Crop Breeding Programs*. PhD thesis.

[Wannemuehler et al., 2020] Wannemuehler, S. D., Yue, C., Hoashi-Erhardt, W. K., Karina Gallardo, R., McCracken, V., and Gallardo, R. K. (2020). Cost-effectiveness analysis of a strawberry breeding program incorporating DNA-informed technology. *HortTechnology*, 30(3):365–371.

[Warner et al., 2010] Warner, D. J., Davies, M., Hipps, N., Osborne, N., Tzilivakis, J., and Lewis, K. A. (2010). Greenhouse gas emissions and energy use in UK-grown short-day strawberry (*Fragaria xananassa Duch.*) crops. *Journal of Agricultural Science*, 148(6):667–681.

[Watson et al., 2019] Watson, A., Hickey, L. T., Christopher, J., Rutkoski, J., Poland, J., and Hayes, B. J. (2019). Multivariate genomic selection and potential of rapid indirect selection with speed breeding in spring wheat. *Crop Science*, 59(5):1945–1959.

[Westoby et al., 2012] Westoby, M. J., Brasington, J., Glasser, N. F., Hambrey, M. J., and Reynolds, J. M. (2012). 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314.

[Whitaker, 2011] Whitaker, V. M. (2011). Applications of molecular markers in strawberry. *Journal of Berry Research*, 1(3):115–127.

[Whitaker et al., 2020] Whitaker, V. M., Knapp, S. J., Hardigan, M. A., Edger, P. P., Slovin, J. P., Bassil, N. V., Hytönen, T., Mackenzie, K. K., Lee, S., Jung, S., Main, D., Barbey, C. R., and Verma, S. (2020). A roadmap for research in octoploid strawberry. *Horticulture Research*, 7(1):1–17.

[Wietzke et al., 2018] Wietzke, A., Westphal, C., Gras, P., Kraft, M., Pfohl, K., Karlovsky, P., Pawelzik, E., Tscharntke, T., and Smit, I. (2018). Insect pollination as a key factor for strawberry physiology and marketable fruit quality. *Agriculture, Ecosystems and Environment*, 258(March):197–204.

[Williams et al., 2005] Williams, A., Ryan, D., Olarte Guasca, A., Marriott, P., and Pang, E. (2005). Analysis of strawberry volatiles using comprehensive two-dimensional gas chromatography with headspace solid-phase microextraction. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 817(1):97–107.

[Wilson et al., 2019] Wilson, F. M., Harrison, K., Armitage, A. D., Simkin, A. J., and Harrison, R. J. (2019). CRISPR/Cas9-mediated mutagenesis of phytoene desaturase in diploid and octoploid strawberry. *Plant Methods*, 15(1):1–13.

[Wingo et al., 2017] Wingo, T. S., Kotlar, A., and Cutler, D. J. (2017). MPD: multiplex primer design for next-generation targeted sequencing. *BMC Bioinformatics*, 18(1):14.

[Wu and Sun, 2013] Wu, D. and Sun, D.-W. (2013). Colour measurements by computer vision for food quality control - A review. *Trends in Food Science and Technology*, 29(1):5–20.

[Xiong et al., 2018] Xiong, Y., From, P. J., and Isler, V. (2018). Design and Evaluation of a Novel Cable-Driven Gripper with Perception Capabilities for Strawberry Picking Robots. *Proceedings - IEEE International Conference on Robotics and Automation*, (2):7384–7391.

[Xiong et al., 2019] Xiong, Y., Peng, C., Grimstad, L., From, P. J., and Isler, V. (2019). Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. *Computers and Electronics in Agriculture*, 157(December 2018):392–402.

[Xu and Zhao, 2010] Xu, L. and Zhao, Y. (2010). Automated strawberry grading system based on image processing. *Computers and Electronics in Agriculture*, 71(71S):32–39.

[Xu, 2003] Xu, S. (2003). Theoretical Basis of the Beavis Effect. *Genetics*, 165(4):2259–2268.

[Xu et al., 2014] Xu, S., Zhu, D., and Zhang, Q. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34):12456–12461.

[Xu and Crouch, 2008] Xu, Y. and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: From publications to practice. *Crop Science*, 48(2):391–407.

[Xue et al., 2001] Xue, H., Aziz, R. M., Sun, N., Cassady, J. M., Kamendulis, L. M., Xu, Y., Stoner, G. D., and Klaunig, J. E. (2001). Inhibition of cellular transformation by berry extracts. *Carcinogenesis*, 22(2):351–6.

[Yangilar, 2013] Yangilar, F. (2013). The application of dietary fibre in food industry : Structural features , effects on health and definition , obtaining and analysis of .... The Application of Dietary Fibre in Food Industry : Structural Features , Effects on Health and Definition , Obtain. *Journal of Food and Nutrition Research*, 1(3):13–23.

[Yu and Lawrence, 2012] Yu, Y. and Lawrence, L. (2012). *Agreement: Statistical Tools for Measuring Agreement. R package version 0.8-1.*

[Zhang and Chen, 2001] Zhang, C. and Chen, T. (2001). Efficient Feature Extraction for 2D/3D Objects in Mesh Representation. *Virtual Reality*, pages 1–4.

[Zhang et al., 2019] Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Frontiers in Genetics*, 10(MAR):1–10.

[Zhang et al., 2017] Zhang, S. D., Jin, J. J., Chen, S. Y., Chase, M. W., Soltis, D. E., Li, H. T., Yang, J. B., Li, D. Z., and Yi, T. S. (2017). Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytologist*, 214(3):1355–1367.

[Zhang et al., 2011] Zhang, S. D., Soltis, D. E., Yang, Y., Li, D. Z., and Yi, T. S. (2011). Multi-gene analysis provides a well-supported phylogeny of Rosales. *Molecular Phylogenetics and Evolution*, 60(1):21–28.

[Zhang et al., 2016] Zhang, Y., Teng, P., Shimizu, Y., Hosoi, F., and Omasa, K. (2016). Estimating 3D leaf and stem shape of nurserypaprika plants by a novel multi-camera photography system. *Sensors (Switzerland)*, 16(6).

[Zingaretti et al., 2019] Zingaretti, M. L., Monfort, A., and Pérez-Enciso, M. (2019). PSBVB: A versatile simulation tool to evaluate genomic selection in polyploid species. *G3: Genes, Genomes, Genetics*, 9(2):327–334.

[Zorrilla-Fontanesi et al., 2012] Zorrilla-Fontanesi, Y., Rambla, J.-L., Cabeza, A., Medina, J. J., Sanchez-Sevilla, J. F., Valpuesta, V., Botella, M. A., Granell, A., and Amaya, I. (2012). Genetic Analysis of Strawberry Fruit Aroma and Identification of *O-Methyltransferase FaOMT* as the Locus Controlling Natural Variation in Mesifurane Content. *Plant Physiology*, 159(2):851–870.

[Zuidmeer et al., 2006] Zuidmeer, L., Salentijn, E., Rivas, M. F., Mancebo, E. G., Asero, R., Matos, C. I., Pelgrom, K. T., Gilissen, L. J., and Van Ree, R. (2006). The role of profilin and lipid transfer protein in strawberry allergy in the Mediterranean area. *Clinical and Experimental Allergy*, 36(5):666–675.

[Zunino et al., 2011] Zunino, S. J., Parelman, M. A., Freytag, T. L., Stephensen, C. B., Kelley, D. S., Mackey, B. E., Woodhouse, L. R., and Bonnel, E. L. (2011). Effects of dietary strawberry powder on blood lipids and inflammatory markers in obese human subjects. *British Journal of Nutrition*, 108(05):900–909.